Comment on Jakob, Deschamps & Maier: How to Read a Paper[1]

Peter A. Bancel

Institut Métapsychique International

Abstract: In this issue of JAEX, Jakob et al. (2024) report results from an experiment to test whether an implicit psychokinetic effect related to unconscious intention can alter the outcome of quantum events. The experiment is well conceived and executed. However, the claim of evidence to support the experimental hypothesis is undermined by errors of analysis and of omission. Some are not obvious and require a careful reading to discern. I explain the nature of the mistakes, where and how they arise, and the consequences for the inferences the authors' draw. I emphasize that a helpful way to detect such problems is by following the chain of inference in experimental reports. Editors, reviewers, and particularly authors can protect the integrity of the scientific literature by reading carefully in this way.

Keywords: Bayesian *t* test, optional stopping, multiple testing, psychokinesis, psi, mind-matter interaction, replication crisis.

Highlights

- An interesting paper by Jakob et al. (2024) reports supporting evidence for a micro-PK effect related to unintentional personality traits.
- The methods and procedures have errors and omissions that undermine the paper's conclusions.
- When reading scientific papers, it is helpful to examine carefully the chain of inferences that lead to research conclusions.

[1] Address correspondence to: Peter A. Bancel, Ph. D., contact@imiresearch.fr

https://doi.org/10.31156/jaex.25420

I thank the editor-in-chief of JAEX for the opportunity to comment on the preceding paper, *Investigating the Effect of Cluster C Personality Styles on Micro-Psychokinesis*, for which I was a reviewer. In their paper, Jakob et al. (henceforth, *Jakob*) claim strong evidence for an unconscious (i.e., unintentional) psychokinesis effect (PK) related to personality traits. The claim of statistical evidence, and subsequent inferences to PK, are based on a Bayesian analysis of experimental data that I believe is misleading. I explain why the statistical inferences the authors make are not well-supported by their analyses. Along with this critique, I discuss the authors' contribution in the context of the so-called "replication crisis" in science to emphasize the importance and challenge of getting things right in the primary literature. This broader point is relevant for researchers, journal editors, and reviewers. It is a problem that although largely recognized and acknowledged, requires ongoing effort to address. I feel it is appropriate to address the issue in the pages of JAEX since these concerns, and the evolving response of the reform movement in statistics, are regularly highlighted in specialty journals in other fields of research.

I start with due praise for *Jakob*. The authors identify an important question in psi research, argue carefully for its connection to theory, conceive and execute a well-planned experiment, and are attentive to current best practices. In addition, the work is part of an ongoing research program at Ludwig-Maximilians-Universität (LMU) and it is laudable to make a sustained effort in a line of psi research. The paper is worth a read.

The research question they address is whether psi effects, and PK in particular, are associated with unconscious intention. This is important for its connection to theoretical models, such as those derived from the Pauli-Jung *Unus Mundus* model of synchronicity and others (see *Jakob* and references therein), but also because it investigates psi effects that potentially circumvent the inherent variability in psi agent effort. Our understanding of the circumstances of such implicit psi is cloudy, to say the least. The prospect that these effects could be reliable brightened with the now famous experiments of Bem (2011) and a subsequent meta-analysis of replications (Bem et al., 2016). Those hopes have dimmed since recent high-powered, multi-lab confirmatory studies have failed to find effects (Kekecs et al., 2023; Muhmenthaler et al., 2022; Schlitz et al., 2021). *Jakob* innovate around these questions and, in doing so, provide a complementary approach to the attempts at straightforward replication.

The design and execution of *Jakob* benefits from applying up-to-date practices: empirical verification of their personality trait metric; implementation of an online design to access a diverse participant base; use of Bayesian statistics that allow stopping rules to conserve resources; and, preregistration, which is now standard practice in parapsychology as elsewhere. Despite these advantages, methodological and procedural mistakes undermine their conclusions. Among these are: failure to account for multiple analyses; stipulating a primary hypothesis with an uncontrolled type I error rate; and not following preregistered procedure.

The problems in *Jakob* arise in part from the challenges to apply Bayesian methods in data analysis. Statistical analysis is tricky. Statisticians have lamented the pitfalls of Null Hypothesis Significance Testing (NHST) and (mis)use of *p*-values for a long time. Not surprisingly, these issues play a central role in the current "replication crisis" (Ioannidis, 2005) and they remain challenging to remedy (see for example, Wasserstein et al., 2019, and articles in the same issue of *The American Statistician*). It is thus natural that there is a learning curve for less familiar Bayesian techniques as well (for an accessible discussion see Schmalz et al., 2023). I have learned things in preparing this comment and appreciate, not for the first time, the value in reviewing statistical subtleties, so the comment aims to be helpful and tutorial. The purpose is to identify some pitfalls that may not at first be obvious, and thus aid the reader in a better assessment of statistical inferences.

To summarize, *Jakob* tested 2,403 participants with three consecutive psi tasks and divided the participants in groups with high or low tendencies relative to each of three personality traits (PTs; as determined from responses to a questionnaire). The separate psi tasks are PT-specific and each consists of 30 trials with random stimuli (sequentially presented word phrases) that are either neutral or relevant for the PT. An equi-probable quantum-based random bit generator (qRNG) selected whether neutral or relevant stimuli were presented. Under the Null hypothesis of no psi effect (H_0) the expected average count of relevant stimuli was thus 15 for the 30-trial tasks. The theoretical proposal is that an unconscious affinity for PT-relevant stimuli will bias the qRNG outcomes, thereby associating reality emergence (qRNG outcomes) with hidden intentional states (the PTs) via a PK effect.

The data were divided into 6 sets: a high or low PT group for each of the 3 PT-specific tasks. The hypotheses stated that data from the high PT groups would yield strong evidence of excess relevant stimuli (H1), whereas data from low PT groups would not. The criterion for evidence is a Bayes Factor (BF) > 10. A BF threshold of 10 is conventionally considered strong evidence in Bayesian analysis (Schönbrodt & Wagenmakers, 2018). Groups with low PT scores are controls and are hypothesized to yield BFs < 10, relative to either H0 or H1. Specifically, the registered primary hypothesis was that *one or more* of the 3 high PT groups would surpass the evidence criterion BF > 10, in either direction for or against the hypotheses. *Jakob* found one high PT dataset (labeled the *dependent* or DE-group) with BF > 10 favoring H1, and that the 3 control datasets all had BF < 10.

Jakob conclude that the primary hypothesis was confirmed and that the hypothesis of no strong evidence for PK in the control groups (with low PT scores) was supported. The experiment is presented as confirmatory, which strengthens the statistical inference because all procedures are specified in a preregistration. An important feature is *Jakob*'s use of a Bayesian stopping rule to halt data accumulation. The preregistration states: "we will add participants until a) our desired level of evidence is reached at least for one of the three dependent variables (BF_{10} = 10, BF01 = 1/10) or b) or till a maximum number of participants (*N* = 1,000) without any trend towards 10 has been reached." To sum up, the paper concludes: "… this study provided confirmatory evidence for a pre-registered micro-PK effect for one out of three PTs."; "[for that] target group, strong evidence (BF_{10} > 10) was found. By contrast, no strong evidence for H1 was observed within the control groups."

To assess *Jakob*'s conclusions, it is useful to recall how experiments are done. Often, one begins with a chain of inference whereby a theoretical notion is translated into an experimental hypothesis that distinguishes between effects. An experiment is then designed and executed to produce data that can be tested according to some analytical procedure. The experiment only concludes when a reverse chain of inference leads from the data back to theory. The inferential return chain starts with a statistical inference about the data such as a claim that the data do not agree with H0. The inference may be hypothetico-deductive or Bayesian, or a mixture of these as is the case in *Jakob*. Further inferences in the chain tend to be qualitative, providing interpretative context for the statistical inference. They include experimental inferences to link the statistical claim to an effect, and theoretical inferences that link the effect to theory. In the case of *Jakob*, experimental inferences might involve reasoning that a statistical result is due to PK and not, say, instrumental error. Theoretical inferences might involve reasoning that the PK effect is implicit and associated with the participant PTs, but not due to an intention of the experimenter. Experiments can be confirmatory, like *Jakob*, or exploratory. The difference is the degree to which pre-stated procedures and inferences are followed, with confirmatory experiments adding force to statistical inferences and exploratory studies favoring discovery.

A number of problems undermine the chain of inference to *Jakob*'s conclusions. Most have to do with the statistical inference and it is instructive to see how these arise. They can be variously traced to statistical misunderstandings, procedural flaws, or definitional ambiguities. Some are clear-cut, others less so, but all weaken *Jakob*'s statistical claims and by extension the chain of inference to their conclusions. The point here is not to be overly critical, but rather to help researchers who do experiments (which is not easy!), and readers who follow the chain of inference in a paper. At issue in *Jakob* are: 1) Claims of evidence that are not adjusted for multiple testing; 2) lack of context for Bayesian *t* tests. 3) an undetermined type I error rate of the primary hypothesis; 4) failure to follow the pre-registered procedure; 5) inadequate treatment of the control group analysis.

In the interest of readability, I first give a succinct statement of the objections, and then follow with further comments.

Multiple Testing Invalidates the Inference to "Strong" Evidence.

Jakob's claim of strong evidence is overstated because the Bayesian *t* tests are not independent and therefore require adjustments for multiple testing. This is a compelling criticism of the paper because *Jakob* relies on the claim of strong evidence to support inference to a PK effect. Typically, test statistics are adjusted to control for the false positive error rate (FPE) when multiple tests are performed. Methods such as Bonferroni or False Discovery Rate corrections on *p*-values adjust for the fact that the probability of a type I error increases with the number of tests. However, opinions vary on how to apply corrections because frequentist theory does not determine when and how multiple testing should be treated. Within Bayesian statistics the situation is clear: no adjustment is needed if tests are independent (Sjölander & Vansteelandt, 2019).

To see this, note that the BF is proportional to the ratio of posterior probabilities of H0 and H1. From this, we can see how test independence affects the posterior probability for H1 when multiple tests are made. Imagine an experiment that tests two hypotheses, HA and HB, with data sets XA and XB. An example would be considering two of *Jakob*'s PT groups under their respective H1s. The posterior probability, P(HA, HB | XA, XB) is the (Bayesian) quantity of interest: it is the probability that the hypotheses are true given all the data that were collected. The posterior probability can be expanded using Bayes's rule:

$$P\left(H_A,H_B|X_A,X_B\right)=\frac{P\left(X_A,X_B|H_A,H_B\right)P\left(H_A,H_B\right)}{P\left(X_A,X_B\right)}$$

If the hypotheses and datasets are independent, the probabilities factorize:

$$P\left(H_A,H_B|X_A,X_B\right)=\frac{P\left(X_A|H_A\right)P\left(H_A\right)P\left(X_B|H_B\right)P\left(H_B\right)}{P\left(X_A\right)P\left(X_B\right)}$$

Summing over hypothesis parameters B (taking the marginal),

$$\int_B P\left(H_A,H_B|X_A,X_B\right)=\frac{P\left(X_A|H_A\right)P\left(H_A\right)}{P\left(X_A\right)}\int_B\frac{P\left(X_B|H_B\right)P\left(H_B\right)}{P\left(X_B\right)}$$

$$P\left(H_A|X_A,X_B\right)=\frac{P\left(X_A|H_A\right)P\left(H_A\right)}{P\left(X_A\right)}$$

$$P\left(H_A|X_A,X_B\right)=P\left(H_A|X_A\right)$$

where the last step is just from applying Bayes' rule. So, with independent tests, the probability of HA does not change when there is a second test. A consequence is that no adjustment for multiple testing is needed for the posterior probabilities or BFs (the proportionality factor between ratios of prior and posterior probabilities, given the data). This conclusion does not apply when tests are dependent, which holds if the hypotheses and/or the data sets are associated. In that case, the probabilities do not factorize and the BF of a test on A cannot be considered independently of data set B. In *Jakob*, it is easy to see that both the data sets and the hypotheses are associated. For hypotheses, *any plausible* theoretical association implies dependence, and *Jakob* is clear that the three personality traits they examine are associated. It is precisely because of the common characteristic of fear reactions that the PTs are chosen for study. The primary hypothesis that *one or more* of the high-PT groups should exhibit an effect, and the choice of identical prior probabilities, together express the assumed association implicitly.

The data sets for PT groups are also dependent. Each participant contributed to all three tests and the participant groupings into high and low PT are highly correlated (correlations of about .65). *Jakob* notes the group correlations, but overlooks that the data dependencies imply that BFs cannot be reported separately without taking the multiple tests into account. To see this better, consider the participant groups as treatments on the qRNG outcomes. Because the treatments are correlated, the data sets, in general, will contain dependencies. In Bayesian language one would say that learning something about one high-PT group (from its data) provides information on the others (and their data).

All this is to say that *Jakob* make an invalid inference of strong evidence for an effect by considering one selected BF as an independent statistic. Instead, the BFs should be adjusted for the multiple tests on dependent hypotheses and data. There are ways to do this with hierarchical Bayesian models that incorporate hypothesis associations into higher dimensional priors (Berry & Hochberg, 1999; Gelman et al., 2012). In hierarchical models, adjustments for multiple tests are built in and model outputs can be taken at face value from a Bayesian perspective. But the models are quite involved and still do not address the problem of the dependencies in *Jakob*'s data. An alternative is a hybrid approach of frequentist and Bayesian reasoning (Sjölander & Vansteelandt, 2019). What remains is that the assertion that BF > 10 is strong evidence cannot be used, as such, to support claims further up the chain of inference.

Lack of Context for Bayesian t-Tests

It is important for Bayesian analysis to include sensitivity checks when reporting BFs (Kruschke, 2021). This is particularly important when using informed priors and for small effects, both of which pertain to *Jakob*. For instance, *Jakob* use a Cauchy(0.05, 0.05) prior that yields a final BF=10.41 for the DE-high group. Tight, informed priors such as this can be sensitive to parameter changes or to slight relaxations of the constraint of a point estimate for H0. For example, with a modified prior of Cauchy(0.1, 0.1) the BF reduces by nearly half to BF=5.78. When calculated BFs are not robust, authors should either qualify their inferences accordingly, or explain a priori why a tight choice of prior is preferred over neighboring ones. *Jakob* choose their prior because it was used in some previous papers, but that is not a reason to ignore the BF's sensitivity to parameter choices. The issue has special relevance in psi research where experimenter psi has been shown to influence statistical outcomes in cases where parameter selection is an avenue for psi effects (Bancel, 2017). *Jakob* mention the possibility of experimenter PK as a possible caveat to their conclusions, but parameter choice is another way experimenter psi could influence the results. Such influences aside, the sensitivity of the BFs to parameter choices is a factor of concern for statistical inference and this is not considered in *Jakob*.

Large, Unreported Type I Error of the Primary Hypothesis

Jakob's primary hypothesis is a test for BF > 10 among the 3 high PT groups. With optional stopping, the false positive error (FPE) rate of the test is roughly .15 (15%) which is far higher than the typical standard of .05 for confirmatory studies. *Jakob*'s large type I error is not reported or discussed. *Jakob* mentions in the Discussion that the primary hypothesis (at least one BF > 10) is "a relatively weak postulate", but they neither explain the weakness, nor present a FPE estimate, nor do they consider further the impact of FPE on the inference to "confirmatory evidence for a preregistered micro-PK effect." Both Bayesian and frequentist methods are subject to type I errors and FPE must be considered when making inferences about a test. When *N* is fixed, the criterion BF > 10 does a good job of controlling type I error, with a FPE of about .006 when *N* is comparable to that of the data in *Jakob*. However, with fully optional stopping, the FPE increases to ~ .05 for a single dependent variable and to ~ .15 when considering one "success" in three. If the stopping checks are performed only once per day, as *Jakob* specifies in preregistration, the FPE decreases slightly to perhaps ~.14. But we don't know the real check frequency and so the true FPE remains high and undetermined.

Although it is not statistically *incorrect* to use a test with high FPE, it is not advised, and whatever the FPE may be, it is necessary to make inferences with FPE estimates in mind. Further, the goal of any paper should be to provide key information transparently to allow readers to assess the inferential chain. To omit a discussion of FPE when it is large is an instance of inadequate reporting. Last, a strong recommendation is to include FPE analysis into the experimental design at the start. There are ways to adequately control FPE adapted to different scientific goals and resource limitations. These should be carefully considered when planning analysis of Bayes factors and setting stopping rules (Schönbrodt & Wagenmakers, 2018). *Jakob* have needlessly increased the FPE rate by checking the BF too early and frequently. Examples of some alternatives and details of the FPE estimates stated above are given in the Appendix.

Failure to Clarify and Follow Pre-Registered Procedures

Jakob ignores the pre-registered stopping rule. Data collection should have stopped at *N* ≈ 820 when the high DE-PT group first yielded BF > 10 (in Figure 2 of their paper). This deviation from the preregistration adds an undetermined degree of experimenter freedom to the protocol. Part of the problem is due to a lack of clarity in the stopping rule. The rule states that data collection halts either at BF > 10, or when *N* =1,000 *unless a trend towards 10 has been reached*. The rule is ambiguous on several counts. Does the stop at BF > 10 apply only to high PT groups? If BF < 10, does the stop occur when any group attains N = 1,000, or all groups, or high PT groups only? What is meant by “a trend towards 10” for continuing past N = 1000, and how is the trend determined? What then is the stopping rule if collection continues after N = 1,000? These ambiguities allow experimenter degrees of freedom to enter and it is difficult to assess their impact on inferences. Removing such hidden freedoms is precisely the reason for preregistration. *Jakob* undermines the value of preregistration by failing to state the rule clearly.

Ambiguities aside, the rule to stop at the first instance of BF > 10 is clear enough, but it was ignored. Figure 2 in *Jakob* shows BF > 10 at $N \approx$ 820 for the DE-PT group, yet data collection continued. In an early version of the paper, collection in fact stopped when all 6 groups passed N = 1000. The result was BF=7.8 for the DE-PT group. Then, in the final version of the paper, data collection continued further until BF again reached 10 for the DE-PT group at N =1,400. The apparent justification for continuing was that a “trend towards 10 was observed” when BF was at 7.8 and N = 1,260. We are left to puzzle how this trend toward 10 was determined (there is in fact a trend away from 10 in Figure 2) and why the stop was not done at the earlier crossing of BF > 10 in the first place.

An argument can be made that these details are unimportant because the Bayes factor allows stopping for *any* reason as long as all the data are used to calculate BF. This is true when the objective is to report a BF for a single independent dataset. In that case, the BF supplies the relative update from prior to posterior probability. However, *Jakob* don’t use the BF only in this way. The 3 BFs were used as binary variables input to a frequentist test for their primary hypothesis which predicts that one or more of the BFs > 10. A consequence of the failure to follow registration procedures is that the assessment of type I error for the primary hypothesis is muddled further. It is not even clear how to model the decision process that actually occurred, should we want to estimate the impact of decisions on the rate of FPE.

It is essential to follow preregistered procedures when claiming results are confirmatory. Should lapses occur, reports should at least explain the full history of data collection and discuss implications for inference. When explanations are post-hoc, confirmatory claims need to be qualified accordingly.

Inadequate Control Group Analysis

Jakob’s claim that control groups show no strong evidence for an effect is misleading. This is because 1) no direct comparison between the target and control groups was made and, 2) different tests were used for the groups.

A common procedural error in psychology (Makin & Orban de Xivry, 2019) and elsewhere (Nieuwenhuis et al., 2011) is to draw inferences from two groups without performing a direct test of difference. *Jakob* make this mistake and aggravate it by reporting a one-tailed test for target groups and a two-tailed test for control groups. No direct mean difference test is given. Reporting the difference between two separate tests is not equivalent to a direct difference test between groups, and this latter statistic is the relevant one for inferring that the groups differ. A direct test is needed because it is conceivable to have two groups for which a difference test is inconclusive, and yet individual tests alternately support H1 and H0, and thus a seeming difference (Palfi & Dienes, 2020). The proper inference is based on the direct test. Furthermore, *Jakob*'s two-tailed test on control data penalizes these BFs relative to the target groups.

A Reassessment of Jakob's Statistical Inference

We can reassess *Jakob*'s claims in light of the comments above. A fair paraphrasing of *Jakob*'s statistical inferences is: The primary hypothesis was substantiated by confirmatory evidence for one group; that group yielded strong evidence in favor of H1. Other groups, including all controls, did not provide evidence at that level. The hypothesis confirmation and the strong statistical evidence for the DE group supports inference to a PK effect.

A reassessment of these conclusions might read: Although our primary hypothesis was satisfied *sensu stricto*, it does not provide adequate evidential support for an effect because of a high false positive error rate. The FPE exceeds ~15% but a precise estimate is hampered by procedures that either deviated from preregistration or were stated unclearly. The procedural uncertainties also weaken claims that the results are confirmatory in the conventional sense. Furthermore, evidence for the group with BF > 10 must be assessed downwards because of test dependencies. A clearer statement of the strength of evidence unfortunately eludes us due to dependencies in the hypotheses and data groups. With our statistical results weakened and uncertain, we cannot claim adequate evidential support for hypothesis confirmation or for an inference to a PK effect.

The critique arrives at a statistical inference that is contrary to *Jakob*. The reassessment is not about adjusting *Jakob*'s criteria: if strong Bayesian evidence for an effect were provided it would be fine to present an argument for PK, for that would be the next step in the chain of inference. Rather, we find that upon careful reading, *Jakob*'s own criteria are either not met (evidence is not strong because of multiple testing), or are inadequate (the primary 1-in-3 hypothesis has an unacceptable FPE rate and control comparisons are lacking). A consequence is that the chain of inference comes to a halt, and a precise statement of results cannot be made.

Discussion

How to Read a Paper

An experimental paper presents us with a loop of reasoning that goes from theory to data, and from data back to theory. This is mirrored in the typical structure of a journal article, with its sections of Introduction, Methods, Results, Discussion, and Conclusions. The loop of reasoning is notably Bayesian: we update our prior beliefs and knowledge of theory, given the data. However, the updating is not merely analytical as would be the case, say, in calculating a Bayes factor. For readers, it involves scrutinizing the strength – and our own comprehension – of each link in the chain of reasoning. Although perhaps obvious, it is useful to recall this approach when preparing to read a paper carefully, especially for reviewers and editors and of course for the authors themselves.

Applying a chain of reasoning examination to *Jakob* explains why the beginning of the paper is so satisfying: the link from theory to experiment to method is particularly well done and careful. We can also pinpoint where things go amiss. This happens at the stage of hypothesis formulation which mixes up confirmatory and exploratory intentions and, as emphasized above, fails to identify hypothesis dependencies. In hindsight. The confusion could have been avoided by simply comparing high and low PT data without the breakdown into separate PT groups, as this would obviate the dependencies of hypotheses. The data dependencies could be resolved as well by limiting the test to those participants with uniformly high or low scores on all three PT scales. Doing so would remove the correlation between test groups. Formulating the hypothesis this way allows a clean confirmatory test, and subsequent tests of individual PT groups then could be registered as exploratory. But, formulating a clean hypothesis this way does not give evidence for an effect. It yields a high-low difference *t* score of 0.84 with an estimated BF of about 1 under the *Jakob* prior. For all high PT data alone BF10 = 1.32 (*M* = 15.083; *SD* = 2.73) and *t*(2699) = 1.58; *p* = .06.

Aside from exposing logical weaknesses, the chain of reasoning approach also helps to identify errors of omission. Careful validation of the steps in statistical reasoning can expose gaps in information that is needed to move to the next link in the chain. The failure to give the FPE rate, or omitting a sensitivity check of the BF priors, are examples. Of course, it requires some statistical sophistication to spot these omissions, and it is the role of editors, reviewers, and authors to collaborate well and thoroughly so that such oversights are avoided.

In fact, considering peer review as a collaboration as opposed to, say, an adversarial process, is not a bad way to think about it. It highlights the mutual goal of all concerned, which is to assure the integrity of the primary literature. Therein lies the crux of the current crisis in science: the integrity of the primary literature has been compromised and once that happens, it is extremely complicated to remedy. This is true across the sciences, but perhaps even more so for smaller fields of research such as parapsychology with its limited readership. Mistaken conclusions that get published can persist as cited literature and thereby hinder progress. An example in *Jakob* is a citation (Jahn et al., 2000) to support evidence of decline effects. The citation's use is mistaken and has been refuted in popular, academic and peer-reviewed literature (Ehrlich, 2021; Varvoglis & Bancel, 2015, 2016). Nevertheless, it is cited erroneously in psi research to this day. *Jakob* risks this propagation of mistaken inferences with their present work. In their Conclusions, they state "Overall, our findings align with earlier studies documenting evidence for the influence of unconscious observers' intentions on a QRNG," citing previous papers by the group. But the apparent alignment is mistaken, as this critique has explained in some detail. So better to have an ounce of prevention than a pound of cure.

That being said, useful prescriptions such as preregistration, adopting new statistical methods, or commitments to publish null results are not sufficient. Papers need to transparently include full information about the chain of reasoning that leads to a conclusion (Wagenmakers et al., 2021). This allows readers to assess for themselves a paper's conclusions and also learn from the authors' experiences and missteps (for an example of a nicely done paper from the literature on meta-science, see Scheel et al., 2021). For those involved in the particular collaboration that is the process of publication, careful reading helps to maintain as reliable a scientific record as possible, and by extension, that the literature may provide us all with a good read.

References

Bancel, P. A. (2017). Searching for global consciousness: A 17-year exploration. *EXPLORE*, *13*(2), 94–101. https://doi.org/10.1016/j.explore.2016.12.003

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Bem, D., Tressoldi, P., Rabeyron, T., & Duggan, M. (2016). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events. *F1000Research*, *4*, 1188.

https://doi.org/10.12688/f1000research.7177.2

Berry, D. A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, *82*(1), 215–227. https://doi.org/10.1016/S0378-3758(99)00044-0

Ehrlich, R. (2021). *Eight preposterous propositions*. Princeton University Press. https://doi.org/10.1515/9780691228402

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jahn, R., Dunne, B., Bradish, G., Dobyns, Y., Lettieri, A., Nelson, R., ... & Walter, B. (2000). Mind/ Machine interaction consortium: portREG replication experiments. *Journal of Scientific Exploration*, *14*(4), 499–555.

Kekecs, Z., Palfi, B., Szaszi, B., Szecsi, P., Zrubka, M., Kovacs, M., … Aczel, B. (2023). Raising the value of research studies in psychological science by increasing the credibility of research reports: The transparent Psi project. *Royal Society Open Science*, *10*(2), 191375. https://doi.org/10.1098/rsos.191375

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour*, *5*(10), Article 10. https://doi.org/10.1038/s41562-021-01177-7

Makin, T. R., & Orban de Xivry, J.-J. (2019). Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife*, *8*, e48175. https://doi.org/10.7554/eLife.48175

Muhmenthaler, M. C., Dubravac, M., & Meier, B. (2022). The future failed: No evidence for precognition in a large scale replication attempt of Bem (2011). *Psychology of Consciousness: Theory, Research, and Practice*. https://doi.org/10.1037/cns0000342

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), Article 9. https://doi.org/10.1038/nn.2886

Palfi, B., & Dienes, Z. (2020). Why Bayesian "evidence for H1" in one condition and Bayesian "evidence for H0" in another condition does not mean good-enough Bayesian evidence for a difference between the conditions. *Advances in Methods and Practices in Psychological Science*, *3*(3), 300–308. https://doi.org/10.1177/2515245920913019

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007467. https://doi.org/10.1177/25152459211007467

Schlitz, M., Bem, D., Marcusson-Clavertz, D., Cardeña, E., Lyke, J., Grover, R., … & Delorme, A. (2021). Two replication studies of a time-reversed (psi) priming task and the role of expectancy in reaction times. *Journal of Scientific Exploration*, *35*(1), Article 1. https://doi.org/10.31275/20211903

Schmalz, X., Biurrun Manresa, J., & Zhang, L. (2023). What is a Bayes factor? *Psychological Meth-*

ods, *28*(3), 705–718. https://doi.org/10.1037/met0000421

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

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

Varvoglis, M., & Bancel, P. A. (2015). Micro-psychokinesis. In In E. Cardeña, J. Palmer, & D. Marcusson-Clavertz (Eds.), *Parapsychology: A handbook for the 21st century* (pp. 266–281). McFarland & Co.

Varvoglis, M., & Bancel, P. A. (2016). Micro-psychokinesis: Exceptional or universal? *Journal of Parapsychology*, *80*(1), 37–44.

Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š., … Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, *5*(11), Article 11. https://doi.org/10.1038/s41562-021-01211-8

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond “ *p* < 0.05.” *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Appendix

Calculation of False Positive Error (FPE) for Bayes factors in Jakob.

Monte Carlo calculations can be used to estimate the FPE rate. The BFs are marginal probabilities that are typically estimated via a Markov Chain Monte Carlo (MCMC) procedure and this can take considerable computer time. To speed the FPE estimates with stopping rules (which need millions of BF estimates), it is useful to have a closed expression to approximate BF from values *N* and simulation *t* scores, which have a monotonic relation to the BFs. The (somewhat tedious) details follow.

As a starting point, observe that for *p* values < .1 and a standard default prior of Cauchy(0, 0.71) (Wagenmakers, 2022):

$$BF\,(p,N) \sim \frac{1}{3p\sqrt{N}}$$

Using the closed form expression for the *p* value of a *t* score (in a normal approximation), we can try a general form for BF as:

$$BF(t,d) = \frac{9}{\sqrt{N}\left(1 - \mathrm{Erf}\left(\frac{t}{d}\right)\right)}$$

The scale factor *d* depends on both *N* and the prior, and the numerator value 9 is a convenience factor determined from fits. Next is to find an approximation for the scale factor *d*(*N*) under the Cauchy(0.05,0.05) prior used by *Jakob*. To do this, MCMC BFs are calculated for random data using a representative set of *N*s (using the statistical platform R, version 4.3.1). The *N* values used here are (200, 300, 500, 700, 900, 1000, 1200, 1400, 2000, 3000), which span the size of data sets in *Jakob*. By inspection, we find a fairly simple form for the scale factor:

$$d(N) = 1 + ae^{g/N} + be^{\frac{h}{\sqrt{N}}}$$

Values of *d*(*N*) that give a best fit to BF(*t*, *d*) in the vicinity of BF=10 are then determined for the *N* values above by iterating *d* and visually assessing the fit near BF=10 in a graphical plot. This takes a few minutes and is straightforward to do. The (*d*, *N*) value pairs are then fitted using *Mathematica's* NonlinearModelFit routine to the expression above. This yields parameter values for *d*(*N*) that are adapted to the Cauchy prior used in *Jakob*, as:

$$(a,b) = (-1.4831, +2.072) \qquad (g,h) = (-211.2, -4.4428)$$

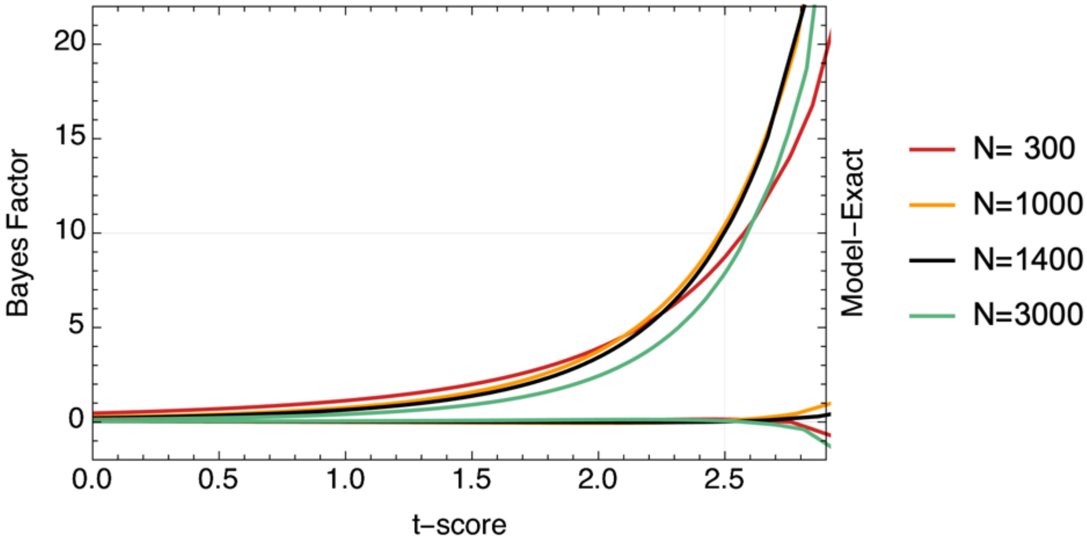
The BFs for the *Jakob* prior can thus be estimated quickly for a given *t* and *N* as:

$$BF(t,d) = \frac{9}{\sqrt{N}\left(1 - \mathrm{Erf}\left(\frac{t}{d(N)}\right)\right)}$$

Figure A.1 shows the accuracy of BF(*t*, *d*) for several sizes of data set. The *t* scores are values of Student *t* tests on simulated data with the 30-trial tasks, as in *Jakob*. The exact BFs (from full MCMC calculations done in R) versus *t* score are shown in Figure A.1 in the traces that curve upward. Note that *t* scores of around 2.5 correspond to BFs near 10, but that the BF value depends on N. The horizontal traces show the difference between exact BFs and the estimations using BF(*t*, *N*) outlined above. The estimates are quite good across the range of *N* and for *t* scores up to 2.6; beyond that, the deviations from exact BFs increase, but the fractional accuracy is still high for BFs well beyond 20. An interesting detail is that, for a fixed value of the *t* score, BFs are lowered at extremes of both small and large *N*. This is a consequence of Bayesian analysis when effects are small. For small effects, the Null is favored more for small samples, and BFs are more stringent with large samples (for a discussion, see Rouder et al., 2009).

Figure A.1

Analytical Estimates of BF as a Function of Sample N and T-Test Value



Note: *The relation between BF and the one-sided t test for fixed sample N. The full MCMC calculations of BFs, for a selection of sample sizes (Ns) are shown as the four rising curves. The BFs rise sharply as t-test values rise above ~2. The horizontal traces show, for the same selection of Ns, the difference (Model-Exact) between calculated BFs and their values given by the approximate formula for BF(t, N) above. The negligible differences indicate that the formula is highly accurate over the selected range of Ns and t-scores.*

Monte Carlo estimations of the FPE rate are shown in Table A.1. The FPE rate depends on the stopping rule and a variety of scenarios are simulated. Details are as follows. Null data arrays are generated using *Mathematica's* (version 12.3.1) RandomInteger function to obtain 10,000 surrogate data sets of *N*Final trials each. The sequential *t* scores for each array are calculated (that is, the score at entry *n* is the *t* value for trials 1 through *n*). The *t* scores are then converted to BFs using BF(*t*, *d*). The FPE rate is estimated by counting the fraction of arrays with a BF > 10 over a range of trials that begins at *N*Start and ends at *N*Final. Table A.1 shows the FPE for *N*Start set to 100 and 500, and for values of *N*Final of 300, 1000, 1400 and 3000. Depending on the stopping rule, data can be checked for the stop criterion continuously or more sparsely, at either fixed or opportunistic intervals. *Jakob* specified daily checking which, if followed, would impose a sampling rate of approximately 20%. The tests in Table A.1 comprise sampling rates of 100, 40, 20, 2 and 0 per cent. Errors for the FPE estimate are estimated in turn by a bootstrap resampling of the 10,000 MC arrays of each scenario. This gives an estimation of FPE for a single Bayesian *t* test BF as a function of total array length, the start check position and the sampling rate.

The relevant FPE estimates for *Jakob* are for at least one instance of BF > 10 in a set of 3 arrays. This roughly triples the FPE estimates for single tests and is shown in the first half of Table A.1. Note that the FPE is fairly insensitive to the sampling rate. Significant control of the FPE rate only occurs as the sampling becomes sparse. Setting the position of the first check, *N*Start, to larger values also provides a way to control the FPE. In the Table, the errors of estimates are tighter for the case of no stopping, and these are indicated separately in the last column.

Table A.1

False Positive Error for Bayes Factors > 10

3 Tests						
		FPE (± 0.018)				FPE (± 0.008)
N_{Start}	N_{Final}	Continuous	40%	20%	2%	No stopping
100	1000	0.141	0.133	0.130	0.092	0.017
500	1000	0.077	0.073	0.072	0.058	0.020
100	1400	0.157	0.152	0.147	0.109	0.018
500	1400	0.117	0.114	0.110	0.093	0.023
1 Test						
		FPE (±0.002)				FPE (± 0.001)
N_{Start}	N_{Final}	Continuous	40%	20%	2%	No stopping
100	1000	0.05	0.047	0.046	0.032	0.006
500	1000	0.026	0.025	0.024	0.02	0.007
100	1400	0.056	0.053	0.052	0.038	0.006
500	1400	0.04	0.039	0.038	0.032	0.008

Commentaire sur Jakob et al.: Comment Lire un Article

Peter A. Bancel

Résumé:Dans ce numéro de JAEX, Jakob et al (2024) rapportent les résultats d'une expérience visant à tester si un effet psychokinétique implicite lié à l'intention inconsciente peut modifier l'issue d'événements quantiques. L'expérience est bien conçue et exécutée. Cependant, l'allégation de preuves à l'appui de l'hypothèse expérimentale est compromise par des erreurs d'analyse et d'omission. Certaines ne sont pas évidentes et nécessitent une lecture attentive pour être discernées. J'explique la nature des erreurs, où et comment elles se produisent, et les conséquences sur les conclusions faites par les auteurs. L'accent est missur le fait qu'un moyen utile de détecter ces problèmes consiste à suivre la chaîne d'inférence dans les rapports expérimentaux. Les rédacteurs, les reviewers, et surtout les auteurs peuvent protéger l'intégrité de la littérature scientifique en ayant une lecture attentive incluant les éléments ici développés.

Translation into French by Antoine Bioy, Ph. D.

Kommentar zu Jakob et al.: Wie man einen Artikel liest

Peter A. Bancel

Zusammenfassung: In dieser Ausgabe von JAEX berichten Jakob et al. (2024) über die Ergebnisse eines Experiments, in dem getestet wurde, ob ein impliziter psychokinetischer Effekt, der mit einer unbewussten Intention zusammenhängt, das Ergebnis von Quantenereignissen verändern kann. Das Experiment ist gut geplant und durchgeführt worden. Die Behauptung jedoch, es läge ein Beweis zugunsten der experimentelle Hyothese vor, wird jedoch durch fehlerhafte Analyse und durch Auslassungen unterminiert. Einige sind nicht offensichtlich und erfordern eine sorgfältige Lektüre, um sie zu erkennen. Ich erkläre die Art der Fehler, wo und wie sie auftreten, und die Konsequenzen für die Schlussfolgerungen, die die Autoren ziehen. Es wird betont, dass eine hilfreiche Methode, solche Probleme zu erkennen, darin besteht, auf die Abfolge der Schlussfolgerungen in den experimentellen Protokollen zu achten. Herausgeber, Gutachter und insbesondere Autoren können zum Schutz der Integrität der wissenschaftlichen Literatur beitragen, indem sie auf diese Weise sorgfältig lessen.

Translation into German by Eberhard Bauer, Ph. D.

Comentário sobre Jakob et al.: Como Ler um Artigo

Peter A. Bancel

Resumo: Neste número da JAEX, Jakob et al (2024) relatam resultados de um experimento que visava testar se um efeito psicocinético implícito relacionado à intenção inconsciente pode alterar o resultado de eventos quânticos. O experimento é bem concebido e executado. No entanto, a alegação de evidência para apoiar a hipótese experimental fica comprometida por erros de análise e omissão. Alguns não são óbvios e exigem uma leitura cuidadosa para serem reconhecidos. Explico a natureza dos erros, onde e como surgem, e as consequências para as inferências que os autores fazem. Enfatiza-se que uma maneira útil de detectar tais problemas é seguindo a cadeia de inferências em relatórios experimentais. Editores, revisores e, especialmente, autores podem proteger a integridade da literatura científica lendo cuidadosamente dessa forma.

Translation into Portuguese by Antônio Lima, Ph. D.

Comentario Sobre Jakob et al.: Cómo Leer un Artículo

Peter A. Bancel

Resumen: En este número de *JAEX*, Jakob et al. (2024) reportan los resultados de un experimento para evaluar si un efecto psicokinético implícito relacionado con una intención inconsciente puede alterar el resultado de eventos cuánticos. El experimento está bien concebido y ejecutado. Sin embargo, la pretensión de apoyo a la hipótesis experimental se ve socavada por errores de análisis y de omisión. Algunos no son obvios y requieren una lectura cuidadosa para discernirlos. Explico la naturaleza de los errores, dónde y cómo se producen, y las consecuencias para las inferencias que extraen los autores. Subrayo que una forma útil de detectar estos problemas es seguir la cadena de inferencias en los informes experimentales. Editores, revisores y, sobre todo, autores pueden proteger la integridad de la literatura científica leyendo de este modo.

Translation into Spanish by Etzel Cardeña, Ph. D.