the WINNOWER

SOCIAL SCIENCES

Why a meta-analysis of 90 precognition studies does not provide convincing evidence of a true effect

DANIEL LAKENS [1]

1. Eindhoven University of Technology

CORRESPONDENCE:
lakens@gmail.com

DATE RECEIVED:
June 12, 2015

KEYWORDS:
esp, pre-cognition, pcurve, publication bias, meta-analysis

A meta-analysis of 90 studies on precognition by Bem, Tressoldi, Rabeyron, & Duggan has been circulating recently. I have looked at this meta-analysis of precognition experiments for an earlier blog post. I had a very collaborative exchange with the authors, which was cordial and professional, and led the authors to correct the mistakes I pointed out and answer some questions I had. I thought it was interesting to write a pre-publication peer review of an article that had been posted in a public depository, and since I had invested time in commenting on this meta-analysis anyway, I was more than happy to accept the invitation to peer-review it. This blog is a short summary of my actual review - since a pre-print of the paper is already online, and it is already cited 11 times, perhaps people are interested in my criticism on the meta-analysis. I expect that many of my comments below apply to other meta-analyses by the same authors (e.g., this one), and a preliminary look at the data confirms this. Until I sit down and actually do a meta-meta-analysis, here's why I don't think there is evidence for pre-cognition in the Bem et al meta-analysis.

Only 18 statistically significant precognition effects have been observed in the last 14 years, by just 7 different labs, as the meta-analysis by Bem, Tressoldi, Rabeyron, and Duggan reveals. 72 studies reveal no effect. If research on pre-cognition has demonstrated anything, it is that when you lack a theoretical model, scientific insights are gained at a painstakingly slow pace, if they are gained at all.

The questions the authors attempt to answer in their meta-analysis is whether there is a true signal in this noisy set of 90 studies. If this is the case, it obviously does not mean we have proof that precognition exists. In science, we distinguish between statistical inferences and theoretical inferences (e.g., Meehl, 1990). Even if a meta-analysis would lead to the statistical inference that there is a signal in the noise, there is as of yet no compelling reason to draw the theoretical inference that precognition exists, due to the lack of a theoretical framework as acknowledged by the authors. Nevertheless, it is worthwhile to see if after 14 years and 90 studies something is going on.

In the abstract, the authors conclude: there is "an overall effect greater than 6 sigma, $z = 6.40$, $p = 1.2 \times 10\text{-}10$ with an effect size (Hedges' g) of 0.09. A Bayesian analysis yielded a Bayes Factor of 1.4×109, greatly exceeding the criterion value of 100 for "decisive evidence" in support of the experimental hypothesis." Let's check the validity of this claim.

Dealing with publication bias.

Every meta-analysis needs to deal with publication bias to prevent the meta-analytic effect size estimate being anything else than the inflation from 0 that emerges because people are more likely to

share positive results. Bem and colleagues use Begg and Mazumdar's rank correlation test to examine publication bias, stating that: "The preferred method for calculating this is the Begg and Mazumdar's rank correlation test, which calculates the rank correlation (Kendall's tau) between the variances or standard errors of the studies and their standardized effect sizes (Rothstein, Sutton & Borenstein, 2005)."

I could not find this recommendation in Rothstein et al., 2005. From the same book, Chapter 11, p. 196, about the rank correlation test: Similarly, from *"the test has low power unless there is severe bias, and so a non-significant tau should not be taken as proof that bias is absent (see also Sterne et al., 2000, 2001b, c)"* . the Cochrane handbook of meta-analyses: *"The test proposed by Begg and Mazumdar (Begg 1994) has the same statistical problems but lower power than the test of Egger et al., and is therefore not recommended."*

When the observed effect size is tiny (as in the case of the current meta-analysis), just a small amount of bias can yield a small meta-analytic effect size estimate that is statistically different from 0. In other words, whereas a significant test result is reason to worry, a non-significant test result is not reason not to worry.

The authors also report the trim-and-fill method to correct for publication bias. It is known that when publication bias is induced by a p-value boundary, rather than an effect size boundary, and there is considerable heterogeneity in the effects included in the meta-analysis, the trim-and-fill method might not perform well enough to yield a corrected meta-analytic effect size estimate that is close to the true effect size (Peters, Sutton, Jones, Abrams, & Rushton, 2007; Terrin, Schmid, Lau, & Olkin, 2003, see also the Cochrane handbook). I'm not sure what upsets me more: The fact that people continue to use this method, or the fact that the people who use this method still report the uncorrected effect size estimate in their abstract.

Better tests for publication bias

PET-PEESE meta-regression seems to be the best test to correct effect size estimates for publication bias we currently have. This approach is based on first using the precision-effect test (PET, Stanley, 2008) to examine whether there is a true effect beyond publication bias, and then following up on this test (if the confidence intervals for the estimate exclude 0) by a PEESE (precision-effect estimate with standard error, Stanley and Doucouliagos, 2007) to estimate the true effect size.

In the R code where I have reproduced the meta-analysis (see below), I have included the PET-PEESE meta-regression. The results are clear: the estimated effect size when correcting for publication bias is 0.008, and the confidence intervals around this effect size estimate do not exclude 0. In other words, there is no good reason to assume that anything more than publication bias is going on in this meta-analysis.

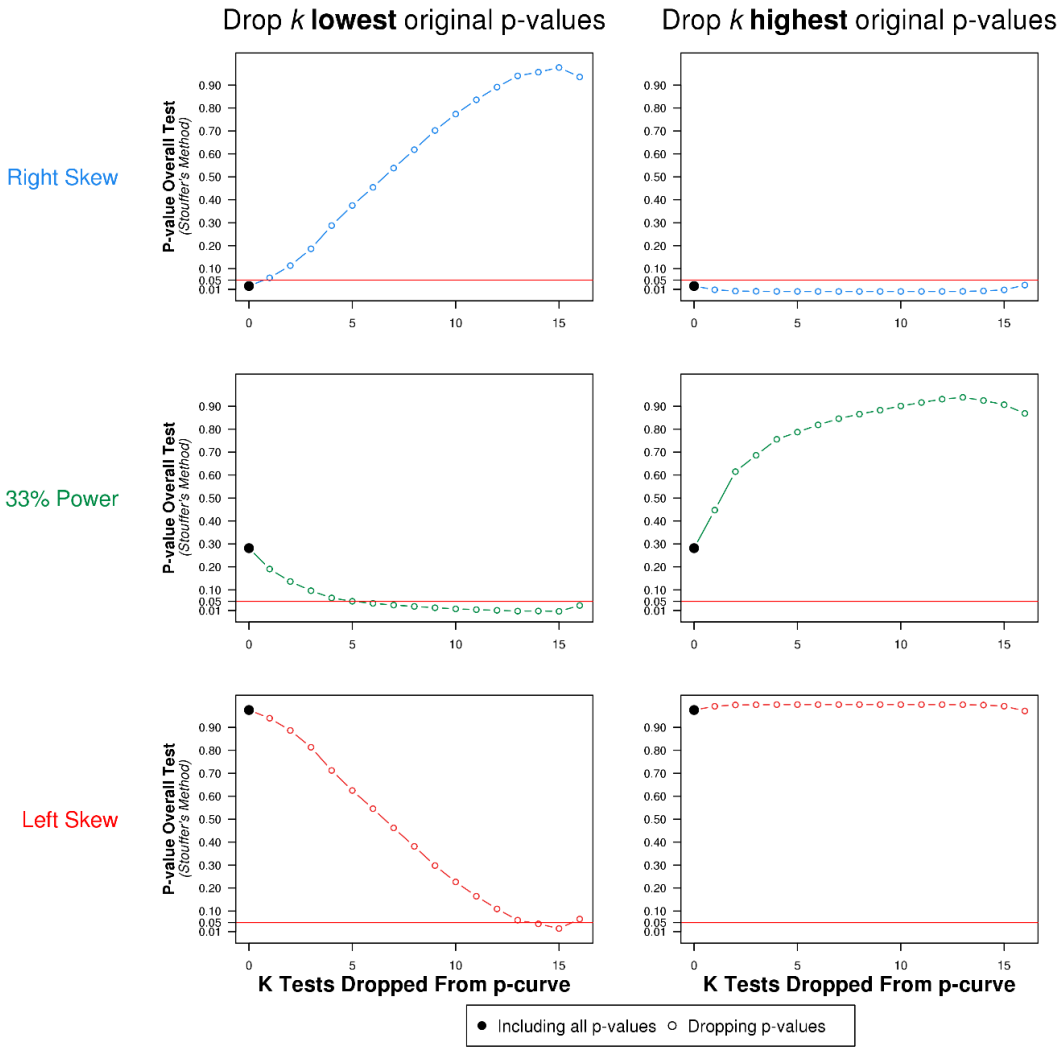
Perhaps it will help to realize that if precognition had an effect size of Cohen's dz = 0.09, to have 90% power to examine an effect with an effect size estimate of 0.09, an alpha level of 0.05, and performing a two-sided t-test, you'd need 1300 participants. Only 1 experiment has been performed with a sufficiently large sample size (Galak, exp 7), and this experiment did not show an effect. Meier (study 3) has 1222 participants, and finds an effect at a significance level of 0.05. However, using a significance level of 0.05 is rather silly when sample sizes are so large (see http://daniellakens.blogspot.nl/2014/05/the-probability-of-p-values-as-function.html) and when we calculate a Bayes Factor using the t-value and the sample size, we see this results in a JZS Bayes Factor of 1.90 - nothing that should convince us.

Estimating the evidential value with *p*-curve and *p*-uniform.

The authors report two analyses to examine the effect size based on the distribution of p-values. These techniques are new, and although it is great the authors embrace these techniques, they should be used with caution. *(I'm skipping a quite substantial discussion of the p-uniform test that was part of*

the review. The short summary is that the authors didn't know what they were doing).

The new test of the p-curve app returns a statistically significant effect when testing for right skew, or evidential value, when we use the test values the authors use (the test has recently been updated - in the version the authors used, the p-curve analysis was not significant). However, the p-curve analysis now also include an exploration of how much this test result depends on a single p-value, by plotting the significance levels of the test if the k most extreme p-values are removed. As we see in the graph below (blue, top-left), the test for evidential value returns a p-value above 0.05 after excluding only 1 p-value, which means we cannot put a lot of confidence in these results.



I think it is important to note that I have already uncovered many coding errors ina previous blog post, even though the authors note that 2 authors independently coded the effect sizes. I feel I could keep pointing out more and more errors in the meta-analysis (instead, I will just recommend to include a real statistician as a co-author), but let's add one to illustrate how easily the conclusion in the current p-curve analysis changes.

The authors include Bierman and Bijl (2013) in their spreadsheet. The raw data of this experiment is shared by Bierman and Bijl (and available at: https://www.dropbox.com/s/j44lvj0c561o5in/Main%20datafile.sav - another excellent example of open science), and I can see that although Bierman and Bijl exclude one participant for missing data, the reaction times that are the basis for the effect size estimate in the meta-analysis are not missing. Indeed, in the master thesis itself (Bijl & Bierman, 2013), all reaction time data is included. If I

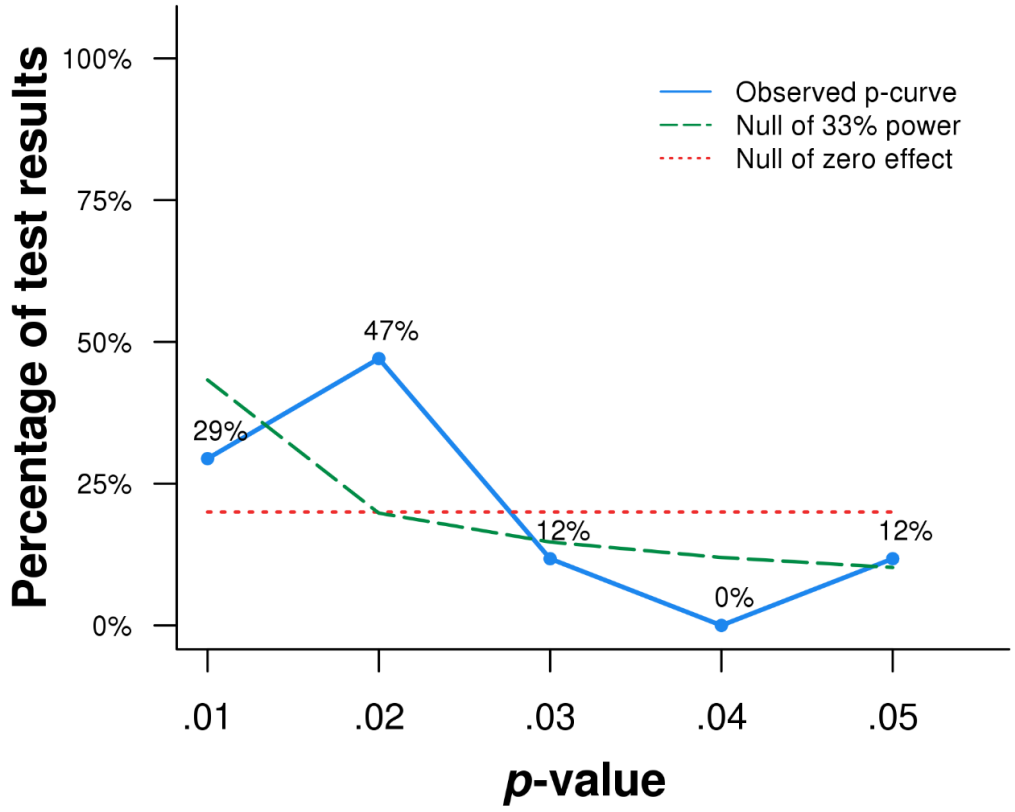
reanalyze the data, I find the same result as in the master thesis:

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	rtctrl_norm - rttrgt_norm	,03469	,13530	,01641	,00194	,06744	2,114	67	,038

I don't think there can be much debate about whether all reaction time data should have been included (and Dick Bierman agrees with me in personal communication), and I think that the choice to report reaction time data from 67 instead of 68 participants in one of those tiny sources of bias that creep into the decisions researchers almost unconsciously make (after all, the results were statistically significant from zero regardless of the final choice). However, for the p-curve analysis (which assumes authors stop their analysis when p-values are smaller than 0.05) this small difference matters. If we include $t(67)=2.11$ in the p-curve analysis instead of $t(67)=2.59$, the new p-curve test no longer indicates the studies have evidential value.

Even if the p-curve test based on the correct data would have shown there is evidential value (although it is comforting it doesn't) we should not be mindlessly interpreting the p-values we get from the analyses. Let's just look at the plot of our data. We see a very weird p-value distribution with many more p-values between 0.01-0.02 then between 0.00-0.01 (whereas the reverse pattern should be observed, see for example Lakens, 2014).



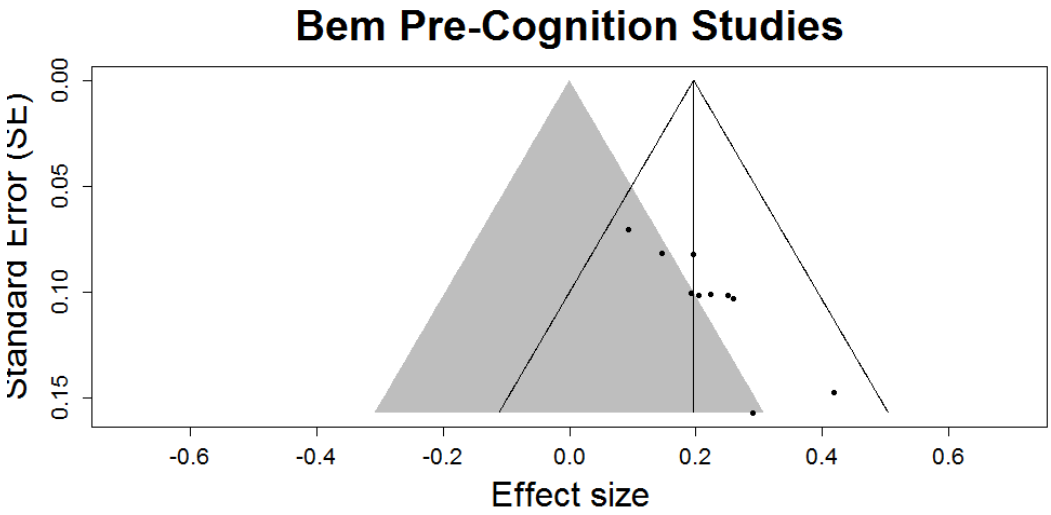
Remember that p-curve is a relatively new technique. For many tests we use (e.g., the t-test) we first perform assumption checks. In the case of the t-test, we check the normality assumption. If data isn't normally distributed, we cannot trust the conclusions from a t-test. I would severely doubt whether we can trust the conclusion from a p-curve if there is such a clear deviation from the expected distribution. Regardless of whether the p-curve tells us there is evidential value or not, the p-curve doesn't look like a 'normal p-value distribution'. Consider the p-curve analysis as an overall F-test for an ANOVA. The p-

curve tells us there is an effect, but if we then perform the simple effects (looking at p-values between 0.00-0.01, and between 0.01-0.02) our predictions about what these effects look like is not confirmed. This is just my own interpretation of how we could improve the p-curve test, and it will useful to see how this test develops. For now, I just want to conclude it is debatable whether the conclusion there is an effect has passed the p-curve test for evidential value (I would say it has not), and passing the test is not immediately a guarantee there is evidential value.

In the literature, a lot has been said about the fact that the low-powered studies reported in Bem (2011) strongly suggest there are an additional number of unreported experiments, or that the effect size estimates were artificially inflated by p-hacking (see Francis, 2012). The authors mention the following when discussing the possibility that there is a file-drawer (page 9):

"In his own discussion of potential file-drawer issues, Bem (2011) reported that they arose most acutely in his two earliest experiments (on retroactive habituation) because they required extensive preexperiment pilot testing to select and match pairs of photographs and to adjust the number and timing of the repeated subliminal stimulus exposures. Once these were determined, however, the protocol was "frozen" and the formal experiments begun. Results from the first experiment were used to rematch several of the photographs used for its subsequent replication. In turn, these two initial experiments provided data relevant for setting the experimental procedures and parameters used in all the subsequent experiments. As Bem explicitly stated in his article, he omitted one exploratory experiment conducted after he had completed the original habituation experiment and its successful replication."

This is not sufficient. The power for his studies is too low to have observed the number of low p-values reported in Bem (2011) without having a much more substantial file-drawer, or p-hacking. It simply is not possible, and we should not accept vague statements about what has been reported. Where I would normally give researchers the benefit of the doubt (our science is built on this, to a certain extent) I cannot do this when there is a clear statistical indication that something is wrong. To illustrate this, let's take a look at the funnel plot for just the studies by Dr. Bem.



Data outside of the grey triangle is statistically significant (in a two-sided test). The smaller the sample size (and the larger the standard error), the larger the effect size needs to be to show a statistically significant effect. If you would report everything you find, effect sizes should be randomly distributed around the true effect size. If they all fall on the edge of the grey triangle, there is a clear indication the studies were selected based on their (one-sided) p-value. It's also interesting to note that the effect size estimates provided by Dr Bem are twice as large as the overall meta-analytic effect size estimate. The fact that there are no unpublished studies by Dr Bem in his own meta-analysis, even when the statistical signs are very clear that such studies should exists, is for me a clear sign of bias.

Now you can publish a set of studies like this ina top journal in psychology as evidence for precognition, but I just use these studies to explain to my students what publication bias looks like in a funnel plot.

For this research area to be taken seriously be scientists, it should make every attempt to be free from bias. I know many researchers in this field, among others Dr Tressoldi, one of the co-authors, are making every attempt to meet the highest possible standards, for example by publishing pre-registered studies (e.g., https://koestlerunit.wordpress.com/study-registry/registered-studies/). I think this is the true way forward. I also think it is telling us something that if replications are performed, these consistently fail to replicate the original results (such as a recent replication by one of the co-authors, Rabeyron, 2014, which did not replicate his own original results - note his original results are included in the meta-analysis, but his replication is not). Publishing a biased meta-analysis stating in the abstract there is "decisive evidence" in support of the experimental hypothesis' while upon closer scrutiny, the meta-analysis fails to provide any conclusive evidence of the presence of an effect (let alone support for the hypothesis that psi exists) would be a step back, rather than a step forward.

No researcher should be convinced by this meta-analysis that psi effects exist. I think it is comforting that PET meta-regression indicates the effect is not reliably different from 0 after controlling for publication bias, and that p-curve analyses do not indicate the studies have evidential value. However, even when statistical techniques would all conclude there is no bias, we should not be fooled into thinking there is no bias. There most likely will be bias, but statistical techniques are simply limited in the bias they can reliably indicate.

I think that based on my reading of the manuscript, the abstract of the manuscript in a future revision should read as follows:

In 2011, the Journal of Personality and Social Psychology published a report of nine experiments purporting to demonstrate that an individual's cognitive and affective responses can be influenced by randomly selected stimulus events that do not occur until after his or her responses have already been made and recorded, a generalized variant of the phenomenon traditionally denoted by the term precognition (Bem, 2011). To encourage replications, all materials needed to conduct them were made available on request. We here report a meta-analysis of 90 experiments from 33 laboratories in 14 countries which yielded an overall effect size (Hedges' g) of 0.09, which after controlling for publication bias using a PET-meta-regression is reduced to 0.008, which is not reliably different from 0, 95% CI [-0.03; 0.05]. These results suggest positive findings in the literature are an indication of the ubiquitous presence of publication bias, but cannot be interpreted as support for psi-phenomena. In line with these conclusions, a p-curve analysis on the 18 significant studies did not provide evidential value for a true effect. We discuss the controversial status of precognition and other anomalous effects collectively known as psi, and stress that even if future statistical inferences from meta-analyses would result in an effect size estimate that is statistically different from zero, the results would not allow for any theoretical inferences about the existence of psi as long as there are no theoretical explanations for psi-phenomena.