Note: The success rate of experiments in the Koestler Parapsychology Unit Study Registry

Chris Phillips
11 February 2023

This note is an attempt to evaluate the success rate of experimental studies of psi pre-registered with the Koestler Parapsychology Unit at Edinburgh University.

The Registry

The KPU Registry is an attempt to overcome the potential problems of selective publication and "researcher degrees of freedom" in experimental parapsychology. It offers an opportunity to pre-register publicly the hypotheses to be tested, the experimental protocols to be used and the planned statistical analyses of the data. The hypotheses are classified as either confirmatory or exploratory. According to the Registry's current guidance, confirmatory hypotheses may be either based on previous empirical evidence or purely theoretical. But it is suggested that well-designed confirmatory analyses should satisfy certain conditions: they should use data obtained by established methods, with a sample size offering sufficient statistical power, all decisions about analysis that may affect the results should be specified in advance, and there should be criteria capable of providing evidence against the hypothesis, not just in favour of it.

The Registry was established in 2012, and in its first ten years 73 studies were pre-registered (a list is available here).

Earlier this year I spent some time reading the registration documents and the published results of the registered studies, with the intention of performing a kind of meta-analysis to evaluate the strength of the evidence accumulated so far. For several reasons (including the fact that some authors did not bother to report numerical p-values when they were not significant) I decided this wasn't feasible.

But in September, David Marks published his evaluation of the studies in a series of three blog posts (1 , 2 and 3). This was not a meta-analysis, but an assessment based on a simple measure of overall success, namely the total number of confirmatory hypotheses that were supported in the registered studies.

As I had already reached my own conclusions about this, which in some respects differed from Marks's, I thought it would be useful to outline them here.

The reported success rate

The hypotheses considered in this assessment are those concerning psi, described as confirmatory at registration, for which results had been published by the end of October 2022 in the form of either a peer-reviewed paper, a preprint or a report to the Registry. Three studies containing such hypotheses (numbers 21, 38 and 43) are excluded because the sample size was much smaller than specified at registration (and, in the third of these, other data required to test the hypotheses were not collected).

The hypotheses satisfying these criteria for inclusion are contained in 25 of the 73 registered studies. Some of these studies include more than one confirmatory hypothesis, and in these cases the multiple hypotheses are often not independent of one another, and are sometimes not very

clearly differentiated. The relevant studies, with my assessment of the number of qualifying hypotheses indicated in parentheses, are those numbered:
1 (3), 2 (1), 3 (1), 4 (1), 6 (3), 7 (1), 8 (1), 9 (1), 10 (1),
11 (1), 12 (2), 13 (4), 15 (3), 16 (1), 18 (2), 19 (1),
25 (1), 26 (2), 27 (1), 36 (4), 37 (1), 46 (2), 49 (2), 51 (1) and 56 (3).

This gives a total of 44 hypotheses to be assessed. For these hypotheses, various statistical tests are specified in the registration documents. In every study a frequentist test is included, sometimes in conjunction with a Bayesian test. A significance level of 0.05 (or equivalently a 95% confidence interval) is specified in all but two studies (and in these no significance level at all is given). On the face of it, this means that on the basis of pure chance the expected number of positive tests would be about 5% of the total.[1]

The reported success rate is higher then this:
1 (2), 2 (1), 8 (1), 9 (1), 11 (1), 13 (2), 26 (1), 46 (1), 56 (2)
That is 12 positive tests in 9 studies, whereas chance expectation based on a significance level of 0.05, applied to the 44 hypotheses considered, would be only 2.2.

Evaluation of positive results

Unfortunately, a closer examination shows that most of the reported positive tests can't be taken at face value.

For a detailed discussion of the problems with these results, see the separate notes on studies. Essentially they fall into three categories:

(1) Problems with the experimental protocol (Studies 1 - for both the positive results - 2, 8, 9 and 46). In two cases the protocol was changed after registration in a way that invalidates the statistical analysis (Studies 1 and 2). In another, there appears to have been a serious problem with the output of a random number generator (Study 9).

(2) A change from the registered statistical analysis, which would have given a negative result, to a different one, which produces a positive result (Study 11).

(3) The use of invalid statistical analyses, which tend to overestimate the apparent statistical significance of the results (Studies 13 and 56 - for both the positive results in each study).

If these are excluded, only a single positive test is left - one of the two in Study 26.

This is the same conclusion that Marks reached, though in most cases his reasons for excluding the positive results were different from mine.[2]

1 Marks considered almost exactly the same registered studies, though he classed number 43 as having a negative result rather than insufficient data, and omitted number 56 (presumably he was unaware of the published results because the registry had not been notified of them). Rather than counting individual hypotheses, he simply attempted to classify each study according to whether all, some or none of the confirmatory tests were positive.
2 Marks's reasons differ from mine for Studies 2, 9, 11, 13 and 46. Note also that he described the result of Study 4 as unclear and omitted Study 6 from consideration altogether (because a link to the published paper on the KPU website was broken). The results for the confirmatory hypotheses were negative for both of these. For further details, see the separate notes on studies.

Digression: One- and two-tailed tests

One of the positive tests excluded because of a problem with the protocol, in Study 46, illustrates a recurring difficulty in interpreting the registration documents. In this case, the initial statement of the relevant hypothesis in words was directional, but in the statistical section, a two-tailed test was specified. In the event, the two-tailed test gave a positive result, but the apparent effect was in the direction opposite to that expected (see the separate notes on studies).[3]

The same apparent inconsistency, between the statement in words of a directional hypothesis and the specification of a two-tailed statistical test, is to be found in a number of the KPU registration documents. Indeed, there seems to be a general tendency in parapsychology to use two-tailed tests regardless of whether the hypothesis is directional. Certainly a two-tailed test may be justified if there is considered to be a possibility of "psi missing" - an effect in the direction opposite to that expected - even if that possibility is relatively small. But there is no doubt that for a wholly directional hypothesis the most powerful test for a given significance level is one-tailed. Particularly in the context of a pre-registered study, it is very difficult to see any advantage in using a two-tailed test when the hypothesis is wholly directional (see note on two-tailed tests).

However, an important purpose of pre-registration is to minimise the freedom of the researcher (or others) to make decisions about analysis after the results are known. So it is certainly not desirable for commentators to modify pre-registered statistical analyses without very pressing reasons. It is particularly undesirable for them to do that when they have knowledge of the results. Moreover, if one two-tailed test were to be replaced by a one-tailed test because the hypothesis was directional, for consistency the same would need to be done in all such cases. Nearly all the studies fix the significance level (at 0.05), so the replacement of two-tailed tests by one-tailed tests would potentially add positive results as well as removing them (for example, in one of the tests in Study 4).

Finally, it may be noted that in two studies with directional hypotheses where one-tailed tests were specified, they produced p-values very close to 1 ($p=0.991$ in Study 18 and $p=0.985$ and 0.972 in two alternative analyses in Study 51). Of course, particularly given the number of tests performed, it is not impossible that these values arose purely by chance. But alternatively they could reflect either an unexpected anomalous effect in the direction opposite to that expected, or an unidentified problem in the execution of the experiments or the analysis of the data.

Conclusions about the success rate

Returning to the question of the overall success rate of the studies, we have arrived at a total of only a single confirmatory hypothesis that produced a positive result.

This needs to be viewed in relation to the total number of confirmatory hypotheses tested. In evaluating this number, we need to consider whether the 11 positive results excluded above should instead be counted as negative results, or whether the problems identified mean that these hypotheses should be excluded from consideration altogether. On the basis of the published results, we can say that for Study 11 (in which the test applied was different from the pre-registered one) the result of the pre-registered test would have been negative, and also that for the result reported as positive in the first experiment of Study 56, a valid version of the pre-registered test would have been negative. For the other 9 problematical hypotheses it is not possible to say whether, if the problems could be resolved, the results would be positive or negative. So in assessing the success

3 Because the verbal description of the hypothesis was directional, Marks treated the result of this test as negative, and argued that claiming significance on the basis of the pre-registered test appeared "a bit bizarre".

rate, these 9 must be excluded from consideration altogether.

Equally, in order to avoid introducing any bias, we must consider whether similar problems affected any of the hypotheses for which negative results were reported. That is the case for the two hypotheses of Study 49, for which the published tests were different from the pre-registered ones, and for the hypothesis of Study 56 that gave a negative result, which must be excluded for the same reason as the one that gave a positive result (see separate notes on studies). The published results were negative, but it is not clear what the results of a valid test of the hypotheses would have been, so these 3 must also be excluded from consideration.

The end result of excluding these 12 hypotheses, is that just one confirmatory test out of a possible 32 produced a positive result.

Conclusions about psi

What conclusions about psi can be drawn from a success rate of 1 out of 32?

The most straightforward conclusion is that with 32 hypotheses and a significance level (in other words an expected false positive rate) of 0.05, we should expect to get on average 1.6 positive results by pure chance. So the success rate is no greater than would be expected by chance, and offers no evidence for the existence of psi.

Does such a low success rate offer evidence *against* the existence of psi? The answer to that question is not so straightforward.[4]

The answer depends partly on the statistical power of the tests applied. If the statistical power (that is, the probability of a positive result if the psi hypothesis being tested is true) were low, that would mean we should expect a high false negative rate, so that no conclusion could be drawn from a smallness of the number of positive results. But for most of these hypotheses, estimates of statistical power were included in the registration documents. Even excluding studies for which the statistical power was overestimated because the registered analyses were invalid, power was estimated at 80% or more for 23 of the 32 hypotheses considered. Taken at face value, that would imply an expectation of at least 18 positive results, not 1. Clearly something was seriously wrong with the expectations of the researchers.

Of course, estimates of the statistical power of experiments depend on estimates of the size of the effect being studied, and estimating the effect size is difficult when even the existence of the effect is controversial. It is notoriously difficult to prove a negative, and the failure to observe an effect may indicate only that it is weaker than expected, rather than that it is not there at all. And several factors may conspire the reduce the statistical power of the registered studies. One is the inclusion of "confirmatory" hypotheses that are purely theoretical, without previous empirical support. Another is the use of two-tailed tests for hypotheses that are wholly directional. But even so, the results of the registered studies reported so far are very difficult to reconcile with the picture of a well-behaved, reproducible psi effect present in the general population, whose existence remains controversial only because of prejudice and lack of resources. Evidently parapsychology is still a long way away from developing reliably replicable experimental protocols, despite the claims sometimes made (for example, in relation to presentiment studies in recent decades).

Evidently the success rate would be consistent with the conclusion that psi does not exist. If, on the

4 Though in the title of his third blog post David Marks offers a very definite answer: "The Psi Hypothesis Has Been Blown Sky High".

other hand, psi does exist, then there must be some reason - as yet not understood - why it is so difficult to demonstrate replicably in the laboratory. Perhaps the most plausible explanation would be that the capacity for psi is found only comparatively rarely in the population, rather than being universally present. Of course, it has even been suggested that positive results in parapsychology experiments may represent a psi effect associated with the experimenter rather than the participants. It is perhaps worth noting that, while the total number of hypotheses tested was quite large at 32, the total number of experimenters testing them was much smaller. The 32 hypotheses formed part of 20 studies, but they represented the work of only 7 different first authors.

Conclusions about pre-registration

Setting aside the questions about the success rate of the studies and the existence of psi, reading the registration documents and trying to evaluate the published results was an instructive experience.

It certainly brought home to me how difficult it must be to run a registry of experimental parapsychology studies. Especially given a degree of reluctance among research workers to pre-register studies at all, it is understandable that the KPU Registry currently operates primarily as a means of offering the opportunity of registration to the community, rather than seeking to act as an arbiter of the quality of the submitted projects or the validity of the published results.

The problem is that parapsychology, probably more than any other branch of science, needs to be - like Caesar's wife - above suspicion with regard to the quality of its experimental protocols and the validity of its analysis. The suggestion is often made that all the positive evidence for the existence of psi can be explained either by pure chance or by questionable research practices. That may seem unfair, in placing a greater burden of proof on parapsychology than on other branches of science. But being realistic, if parapsychology wants to convince the scientific community as a whole that psi effects really exist and are a legitimate area of research, it needs to go further in ensuring that when it does present positive evidence, the methods used to obtain it are beyond reproach.

If that is what parapsychologists want, they will need to take pre-registration more seriously, as a means of demonstrating that their work is being carried out to a high standard of scientific rigour, and also as a means of forestalling as many as possible of the common criticisms that are levelled against them.

To my mind, a more rigorous pre-registration system would have several aspects:

(1) **Specification.** A much more complete description of the work proposed, and particularly of the analysis techniques. As far as data processing and statistical analysis are concerned, the procedures should be specified in enough detail that two different people, following the published prescription independently, could be guaranteed of getting the same results. If that isn't the case, then "researcher degrees of freedom" remain, and the conclusions are open to question on the ground that the freedom may have been exploited to produce the best possible results.

(2) **Validity.** As well as sufficient detail, there needs to be an assurance that the analysis techniques are valid. There is no point in pre-registering in detail what is going to be done if, when it comes to publication, what has been proposed is discovered to be invalid. Exact methods of analysis are obviously preferable to those that rely on approximations. Though exact methods are not always available, they often are, particular with the help of appropriate experimental design. In that case, the use of an approximate method lays the conclusions open to criticism unnecessarily.

(3) **Statistical power.** Given that the very existence of psi phenomena is still at issue within the

scientific community as a whole, if studies are intended to be confirmatory, they should be designed to ensure adequate statistical power. It may be tempting to base estimates of power on over-optimistic assumptions about effect sizes, but the result of that will be under-powered studies. Equally, the use of two-tailed tests when the effect is expected to be wholly directional amounts to a pointless sacrifice of statistical power for no advantage. In the context of the KPU registry's distinction between different types of hypotheses, it seems preferable to limit confirmatory hypotheses to those for which significant empirical support is claimed, and to exclude purely theoretical or speculative hypotheses, even when they are precisely specified.

(4) **Fraud-proofing.** Clearly parapsychology research can expect to be robustly criticised, and that will include the suggestion that fraud may be involved, unfair though that may seem to parapsychologists. Although pre-registration cannot of itself exclude the possibility of fraud, nowadays that possibility can be minimised by technological means, and fraud-proofing should form an essential part of the registration process. For example, in experimental work random numbers (for example, for the selection of targets in a precognition study) are commonly obtained from a remote Internet server. There would be nothing particularly difficult in combining this with the transmission in real time of measured data (for example, the participant's guesses about the targets) to either the same or a different Internet server. That would greatly reduce the scope for fraud (and for some studies - for example those dealing with precognition, clairvoyance and psychokinesis) could effectively eliminate it, in the context of an appropriate protocol.

(5) **The role of the registry.** A more rigorous registration process incorporating these aspects would inevitable require a more active role for the registry, in terms of ensuring that the details of the proposed work were adequately described, the experimental and analytical methods were valid, the power was sufficient and appropriate fraud-proofing safeguards had been incorporated. Realistically, that would require a formal review process analogous to that used by peer-reviewed journals, which would represent a much more active intervention than in the KPU registry (where authors are sometimes asked to clarify their proposals, but where judgments about appropriateness and validity are not viewed as part of the registry's remit).

Certainly such a system of pre-registration would be very different from, and much more demanding of effort than, the facility currently offered by the KPU registry. And admittedly, even as things are now, only a relatively small proportion of the parapsychological work being funded, presented at meetings and published in journals is pre-registered. Nor do the organisers of meetings or the publishers of journals consider pre-registration a high enough priority to make it a condition of acceptance. It may well be that there is not sufficient appetite in the field for a more rigorous pre-registration system. It partly depends on how strongly parapsychologists want to convince the scientific world in general that their field of research is a worthwhile one.

But of course, to move in this direction would not require unanimity among parapsychologists, or anything like it, and even a small collection of rigorously pre-registered studies would have the potential to be the modern-day equivalent of William James's "single white crow".

[I am grateful to Patrizio Tressoldi and Julia Mossbridge for providing information about their registered studies, and to Jim Kennedy for comments on a draft of this note.]