

The Control Group Is Out Of Control

Posted on April 28, 2014 by Scott Alexander



I

Allan Crossman calls parapsychology [the control group for science](#).

That is, in let's say a drug testing experiment, you give some people the drug and they recover. That doesn't tell you much until you give some other people a placebo drug you *know* doesn't work – but which they themselves believe in – and see how many of *them* recover. That number tells you how many people will recover whether the drug works or not. Unless people on your real drug do significantly better than people on the placebo drug, you haven't found anything.

On the meta-level, you're studying some phenomenon and you get some positive findings. That doesn't tell you much until you take some other researchers who are studying a phenomenon you *know* doesn't exist – but which they themselves believe in – and see how many of *them* get positive findings. That number tells you how many studies will discover positive results whether the phenomenon is real or not. Unless studies of the real phenomenon do

significantly better than studies of the placebo phenomenon, you haven't found anything.

Trying to set up placebo science would be a logistical nightmare. You'd have to find a phenomenon that definitely doesn't exist, somehow convince a whole community of scientists across the world that it does, and fund them to study it for a couple of decades without them figuring it out.

Luckily we have a natural experiment in terms of parapsychology – the study of psychic phenomena – which most reasonable people believe don't exist, but which a community of practicing scientists believes in and publishes papers on all the time.

The results are pretty dismal. Parapsychologists are able to produce experimental evidence for psychic phenomena about as easily as normal scientists are able to produce such evidence for normal, non-psychic phenomena. This suggests the existence of a very large “placebo effect” in science – ie with enough energy focused on a subject, you can *always* produce “experimental evidence” for it that meets the usual scientific standards. As Eliezer Yudkowsky puts it:

Parapsychologists are constantly protesting that they are playing by all the standard scientific rules, and yet their results are being ignored – that they are unfairly being held to higher standards than everyone else. I'm willing to believe that. It just means that the standard statistical methods of science are so weak and flawed as to permit a field of study

to sustain itself in the complete absence of any subject matter.

These sorts of thoughts have become more common lately in different fields. Psychologists admit to a [crisis of replication](#) as some of their most interesting findings turn out to be spurious. And in medicine, John Ioannides and others have been criticizing the research for a decade now and telling everyone they need to up their standards.

“Up your standards” has been a complicated demand that cashes out in a lot of technical ways. But there is broad agreement among the most intelligent voices I read ([1](#), [2](#), [3](#), [4](#), [5](#)) about a couple of promising directions we could go:

1. Demand very large sample size.
2. Demand replication, preferably exact replication, most preferably multiple exact replications.
3. Trust systematic reviews and meta-analyses rather than individual studies. Meta-analyses must prove homogeneity of the studies they analyze.
4. Use Bayesian rather than frequentist analysis, or even combine both techniques.
5. Stricter p-value criteria. It is far too easy to massage p-values to get less than 0.05. Also, make meta-analyses look

for “p-hacking” by examining the distribution of p-values in the included studies.

6. Require pre-registration of trials.
7. Address publication bias by searching for unpublished trials, displaying funnel plots, and using statistics like “fail-safe N” to investigate the possibility of suppressed research.
8. Do heterogeneity analyses or at least observe and account for differences in the studies you analyze.
9. Demand randomized controlled trials. None of this “correlated even after we adjust for confounders” BS.
10. Stricter effect size criteria. It’s easy to get small effect sizes in *anything*.

If we follow these ten commandments, then we avoid the problems that allowed parapsychology and probably a whole host of other problems we don’t know about to sneak past the scientific gatekeepers.

Well, [what now, motherfuckers?](#)



Bem, Tressoldi, Rabeyron, and Duggan (2014), full text available for download at the top bar of the link above, is parapsychology’s

way of saying “thanks but no thanks” to the idea of a more rigorous scientific paradigm making them quietly wither away.

You might remember Bem as the prestigious establishment psychologist who decided to try his hand at parapsychology and to his and everyone else’s surprise got positive results. Everyone had a lot of criticisms, some of which were [very very good](#), and the study [failed replication several times](#). Case closed, right?

Earlier this month Bem came back with a meta-analysis of ninety replications from tens of thousands of participants in thirty three laboratories in fourteen countries confirming his original finding, $p < 1.2 \times 10^{-10}$, Bayes factor 7.4×10^9 , funnel plot beautifully symmetrical, p-hacking curve nice and right-skewed, Orwin fail-safe n of 559, et cetera, et cetera, et cetera.

By my count, Bem follows all of the commandments except [6] and [10]. He apologizes for not using pre-registration, but says it’s okay because the studies were exact replications of a previous study that makes it impossible for an unsavory researcher to change the parameters halfway through and does pretty much the same thing. And he apologizes for the small effect size but points out that some effect sizes are legitimately very small, this is no smaller than a lot of other commonly-accepted results, and that a high enough p-value ought to make up for a low effect size.

This is *far* better than the average meta-analysis. Bem has always been pretty careful and this is no exception. Yet its conclusion is that psychic powers exist.

So – once again – what now, motherfuckers?



In retrospect, that list of ways to fix science above was a little optimistic.

The first nine items (large sample sizes, replications, low p-values, Bayesian statistics, meta-analysis, pre-registration, publication bias, heterogeneity) all try to solve the same problem: accidentally mistaking noise in the data for a signal.

We've placed so much emphasis on not mistaking noise for signal that when someone like Bem hands us a beautiful, perfectly clear signal on a silver platter, it briefly stuns us. "Wow, of the three hundred different terrible ways to mistake noise for signal, Bem has proven beyond a shadow of a doubt he hasn't done any of them." And we get so stunned we're likely to forget that this is only part of the battle.

Bem definitely picked up a signal. The only question is whether it's a signal of psi, or a signal of poor experimental technique. *None* of these commandments even *touch* poor experimental technique – or confounding, or whatever you want to call it. If an experiment is confounded, if it produces a strong signal even when its experimental hypothesis is true, then using a larger sample size will just make that signal even stronger.

Replicating it will just reproduce the confounded results again.

Low p-values will be easy to get if you perform the confounded experiment on a large enough scale.

Meta-analyses of confounded studies will obey the immortal law of “garbage in, garbage out”.

Pre-registration only assures that your study will not get any worse than it was the first time you thought of it, which may be very bad indeed.

Searching for publication bias only means you will get *all* of the confounded studies, instead of just some of them.

Heterogeneity just tells you whether all of the studies were confounded about the same amount.

Bayesian statistics, alone among these first eight, ought to be able to help with this problem. After all, a good Bayesian should be able to say “Well, I got some impressive results, but my prior for ψ is very low, so this raises my belief in ψ slightly, but raises my belief that the experiments were confounded *a lot*.”

Unfortunately, good Bayesians are hard to come by, and the researchers here seem to be making some serious mistakes. Here's Bem:

An opportunity to calculate an approximate answer to this question emerges from a Bayesian critique of Bem's (2011) experiments by Wagenmakers, Wetzels, Borsboom, & van der Maas (2011). Although Wagenmakers et al. did not explicitly claim psi to be impossible, they came very close by setting their prior odds at 10^{20} against the psi hypothesis. The Bayes Factor for our full database is approximately 10^9 in favor of the psi hypothesis (Table 1), which implies that our meta-analysis should lower their posterior odds against the psi hypothesis to 10^{11} .

Let me shame both participants in this debate.

Bem, you are abusing Bayes factor. If Wagenmakers uses your 10^9 Bayes factor to adjust from his prior of 10^{-20} to 10^{-11} , then what happens the next time you come up with another database of studies supporting your hypothesis? We all know you will, because you've amply proven these results weren't due to chance, so whatever factor produced these results – whether real psi or poor experimental technique – will no doubt keep producing them for the next hundred replication attempts. When those come in, does Wagenmakers have to adjust his probability from 10^{-11} to 10^{-2} ? When you get another hundred studies, does he have to go from 10^{-2} to 10^7 ? If so, then by [conservation of expected evidence](#) he should just update to 10^7 right now – or really to infinity, since you can keep coming up with more studies till the cows come home. But in fact he shouldn't do that, because at some point his thought process becomes “Okay, I already know that studies of this quality can consistently produce positive findings, so either psi is real or

studies of this quality aren't good enough to disprove it". This point should probably happen well before he increases his probability by a factor of 10^9 . See [Confidence Levels Inside And Outside An Argument](#) for this argument made in greater detail.

Wagenmakers, you are overconfident. Suppose God came down from Heaven and said in a booming voice "EVERY SINGLE STUDY IN THIS META-ANALYSIS WAS CONDUCTED PERFECTLY WITHOUT FLAWS OR BIAS, AS WAS THE META-ANALYSIS ITSELF." You would see a p-value of less than 1.2×10^{-10} and think "I bet that was just coincidence"? And then they could do another study of the same size, also God-certified, returning exactly the same results, and you would say "I bet that was just coincidence too"? YOU ARE NOT THAT CERTAIN OF ANYTHING. Seriously, *read the @#!\$ing Sequences*.

Bayesian statistics, at least the way they are done here, aren't gong to be of much use to anybody.

That leaves randomized controlled trials and effect sizes.

Randomized controlled trials are great. They eliminate most possible confounders in one fell swoop, and are excellent at keeping experimenters honest. Unfortunately, most of the studies in the Bem meta-analysis were already randomized controlled trials.

High effect sizes are really the only thing the Bem study lacks. And it is very hard to experimental technique so bad that it consistently produces a result with a high effect size.

But as Bem points out, demanding high effect size limits our ability to detect real but low-effect phenomena. Just to give an example, many physics experiments – like the ones that detected the Higgs boson or neutrinos – rely on detecting extremely small perturbations in the natural order, over millions of different trials. Less esoterically, Bem mentions the example of aspirin decreasing heart attack risk, which it definitely does and which is very important, but which has an effect size lower than that of his psi results. If humans have some kind of *very weak* psionic faculty that under regular conditions operates poorly and inconsistently, but does indeed exist, then excluding it by definition from the realm of things science can discover would be a bad idea.

All of these techniques are about reducing the chance of confusing noise for signal. But when we think of them as the be-all and end-all of scientific legitimacy, we end up in awkward situations where they come out super-confident in a study's accuracy simply because the issue was one they weren't geared up to detect. Because a lot of the time the problem is something more than just noise.

IV

Wiseman & Schlitz's [Experimenter Effects And The Remote Detection Of Staring](#) is my favorite parapsychology paper ever and sends me into fits of nervous laughter every time I read it.

The backstory: there is a classic parapsychological experiment where a subject is placed in a room alone, hooked up to a video link. At random times, an experimenter stares at them menacingly through the video link. The hypothesis is that this causes their galvanic skin response (a physiological measure of subconscious anxiety) to increase, even though there is no non-psychic way the subject could know whether the experimenter was staring or not.

Schultz is a psi believer whose staring experiments had consistently supported the presence of a psychic phenomenon. Wiseman, in accordance with [nominative determinism](#) is a psi skeptic whose staring experiments keep showing nothing and disproving psi. Since they were apparently the only two people in all of parapsychology with a smidgen of curiosity or rationalist virtue, they decided to team up and figure out why they kept getting such different results.

The idea was to plan an experiment together, with both of them agreeing on every single tiny detail. They would then go to a laboratory and set it up, again both keeping close eyes on one another. Finally, they would conduct the experiment in a series of different batches. Half the batches (randomly assigned) would be conducted by Dr. Schlitz, the other half by Dr. Wiseman. Because the two authors had very carefully standardized the setting, apparatus and procedure beforehand, “conducted by” pretty much just meant greeting the participants, giving the experimental instructions, and doing the staring.

The results? Schlitz's trials found strong evidence of psychic powers, Wiseman's trials found no evidence whatsoever.

Take a second to reflect on how this *makes no sense*. Two experimenters in the same laboratory, using the same apparatus, having no contact with the subjects except to introduce themselves and flip a few switches – and whether one or the other was there that day completely altered the result. For a good time, watch the gymnastics they have to do in the paper to make this sound sufficiently sensical to even get published. This is the only journal article I've ever read where, in the part of the Discussion section where you're supposed to propose possible reasons for your findings, both authors suggest maybe their co-author hacked into the computer and altered the results.

While it's nice to see people exploring Bem's findings further, *this* is the experiment people should be replicating ninety times. I expect *something* would turn up.

As it is, Kennedy and Taddonio [list ten similar studies](#) with similar results. One cannot help wondering about publication bias (if the skeptic and the believer got similar results, who cares?). But the phenomenon is sufficiently well known in parapsychology that it has led to its own host of theories about how skeptics emit negative auras, or the enthusiasm of a proponent is a necessary kindling for psychic powers.

Other fields don't have this excuse. In psychotherapy, for example, practically the only consistent finding is that whatever kind of psy-

chotherapy the person running the study likes is most effective. Thirty different meta-analyses on the subject have confirmed this with strong effect size ($d = 0.54$) and good significance ($p = .001$).

Then there's [Munder \(2013\)](#), which is a meta-meta-analysis on whether meta-analyses of confounding by researcher allegiance effect were themselves meta-confounded by meta-researcher allegiance effect. He found that indeed, meta-researchers who believed in researcher allegiance effect were more likely to turn up positive results in their studies of researcher allegiance effect ($p < .002$).

It gets worse. There's [a famous story](#) about an experiment where a scientist told teachers that his advanced psychometric methods had predicted a couple of kids in their class were about to become geniuses (the students were actually chosen at random). He followed the students for the year and found that their intelligence actually increased. This was supposed to be a Cautionary Tale About How Teachers' Preconceptions Can Affect Children.

Less famous is that the same guy did the same thing with rats. He sent one laboratory a box of rats saying they were specially bred to be ultra-intelligent, and another lab a box of (identical) rats saying they were specially bred to be slow and dumb. Then he had them do standard rat learning tasks, and sure enough the first lab found very impressive results, the second lab very disappointing ones.

This scientist – let's give his name, Robert Rosenthal – [then investigated three hundred forty five different studies](#) for evidence of the

same phenomenon. He found effect sizes of anywhere from 0.15 to 1.7, depending on the type of experiment involved. Note that this could also be phrased as “between twice as strong and twenty times as strong as Bem’s psi effect”. Mysteriously, animal learning experiments displayed the highest effect size, supporting the folk belief that animals are hypersensitive to subtle emotional cues.

Okay, fine. Subtle emotional cues. That’s way more scientific than saying “negative auras”. But the question remains – what went wrong for Schlitz and Wiseman? Even if Schlitz had done everything short of saying “The hypothesis of this experiment is for your skin response to increase when you are being stared at, please increase your skin response at that time,” and subjects had tried to comply, the whole point was that they didn’t *know* when they were being stared at, because to find that out you’d have to be psychic. And how are these rats figuring out what the experimenters’ subtle emotional cues mean anyway? I can’t figure out people’s subtle emotional cues half the time!

I know that standard practice here is to tell [the story of Clever Hans](#) and then say That Is Why We Do Double-Blind Studies. But first of all, I’m pretty sure no one does double-blind studies with rats. Second of all, I think most social psych studies aren’t double blind – I just checked the first one I thought of, Aronson and Steele on stereotype threat, and it certainly wasn’t. Third of all, this effect seems to be just as common in cases where it’s hard to imagine how the researchers’ subtle emotional cues could make a difference. Like Schlitz and Wiseman. Or like the psychotherapy experiments, where most of the subjects were doing therapy with individ-

ual psychologists and never even saw whatever prestigious professor was running the study behind the scenes.

I think it's a combination of subconscious emotional cues, subconscious statistical trickery, perfectly conscious fraud which for all we know happens much more often than detected, and things we haven't discovered yet which are at least as weird as subconscious emotional cues. But rather than speculate, I prefer to take it as a brute fact. Studies are going to be confounded by the allegiance of the researcher. When researchers who don't believe something discover it, that's when it's worth looking into.

V

So what exactly happened to Bem?

Although Bem looked hard to find unpublished material, I don't know if he succeeded. Unpublished material, in this context, has to mean "material published enough for Bem to find it", which in this case was mostly things presented at conferences. What about results so boring that they were never even mentioned?

And I predict people who believe in parapsychology are more likely to conduct parapsychology experiments than skeptics. Suppose this is true. And further suppose that for some reason, experimenter effect is real and powerful. That means most of the experiments conducted will support Bem's result. But this is still a weird

form of “publication bias” insofar as it ignores the contrary results of hypothetically experiments that were never conducted.

And worst of all, maybe Bem really did do an excellent job of finding every little two-bit experiment that no journal would take. How much can we trust these non-peer-reviewed procedures?

I looked through his list of ninety studies for all the ones that were both exact replications and had been peer-reviewed (with one caveat to be mentioned later). I found only seven:

| | |
|------------------------------|--------|
| Batthyany, Kranz, and Erber: | 0.268 |
| Ritchie 1: | 0.015 |
| Ritchie 2: | −0.219 |
| Richie 3: | −0.040 |
| Subbotsky 1: | 0.279 |
| Subbotsky 2: | 0.292 |
| Subbotsky 3: | −0.399 |

Three find large positive effects, two find approximate zero effects, and two find large negative effects. Without doing any calculatin’, this seems pretty darned close to chance for me.

Okay, back to that caveat about replications. One of Bem’s strongest points was how many of the studies included were exact replications of his work. This is important because if you do your own

novel experiment, it leaves a lot of wiggle room to keep changing the parameters and statistics a bunch of times until you get the effect you want. This is why lots of people want experiments to be preregistered with specific commitments about what you're going to test and how you're going to do it. These experiments weren't preregistered, but conforming to a previously done experiment is a pretty good alternative.

Except that I think the criteria for "replication" here were exceptionally loose. For example, Savva et al was listed as an "exact replication" of Bem, but it was performed in 2004 – seven years before Bem's original study took place. I know Bem believes in precognition, but that's going *too far*. As far as I can tell "exact replication" here means "kinda similar psionic-y thing". Also, Bem classily lists his own experiments as exact replications of themselves, which gives a big boost to the "exact replications return the same results as Bem's original studies" line. I would want to see much stricter criteria for replication before I relax the "preregister your trials" requirement.

(Richard Wiseman – the same guy who provided the negative aura for the Wiseman and Schiltz experiment – has started [a pre-register site for Bem replications](#). He says he has received five of them. This is very promising. There is also [a separate pre-register for parapsychology trials in general](#). I am both extremely pleased at this victory for good science, and ashamed that my own field is apparently behind parapsychology in the "scientific rigor" department)

That is my best guess at what happened here – a bunch of poor-quality, peer-unreviewed studies that weren't as exact replications as we would like to believe, all subject to mysterious experimenter effects.

This is not a criticism of Bem or a criticism of parapsychology. It's something that is inherent to the practice of meta-analysis, and even more, inherent to the practice of science. Other than a few very exceptional large medical trials, there is not a study in the world that would survive the level of criticism I am throwing at Bem right now.

I think Bem is wrong. The level of criticism it would take to prove a wrong study wrong is higher than that almost any existing study can withstand. That is not encouraging for existing studies.

VI

The motto of the Royal Society – Hooke, Boyle, Newton, some of the people who arguably invented modern science – was *nullus in verba*, “take no one's word”.

This was a proper battle cry for seventeenth century scientists. Think about the (admittedly kind of mythologized) history of Science. The scholastics saying that matter was this, or that, and justifying themselves by long treatises about how based on A, B, C, the word of the Bible, Aristotle, self-evident first principles, and the Great Chain of Being all clearly proved their point. Then other

scholastics would write different long treatises on how D, E, and F, Plato, St. Augustine, and the proper ordering of angels all indicated that clearly matter was something different. Both groups were pretty sure that the other had made a subtle error of reasoning somewhere, and both groups were perfectly happy to spend centuries debating exactly which one of them it was.

And then Galileo said “Wait a second, instead of debating exactly how objects fall, let’s just drop objects off of something really tall and see what happens”, and after that, Science.

Yes, it’s kind of mythologized. But like all myths, it contains a core of truth. People are terrible. If you let people debate things, they will do it forever, come up with horrible ideas, get them entrenched, play politics with them, and finally reach the point where they’re coming up with theories why people who disagree with them are probably secretly in the pay of the Devil.

Imagine having to conduct the global warming debate, except that you couldn’t appeal to scientific consensus and statistics because scientific consensus and statistics hadn’t been invented yet. In a world without science, *everything* would be like that.

Heck, just look at *philosophy*.

This is the principle behind the Pyramid of Scientific Evidence. The lowest level is your personal opinions, no matter how ironclad you think the logic behind them is. Just above that is expert opinion, because no matter how expert someone is they’re still only human.

Above that is anecdotal evidence and case studies, because even though you're finally getting out of people's heads, it's still possible for the content of people's heads to influence which cases they pay attention to. At each level, we distill away more and more of the human element, until presumably at the top the dross of humanity has been purged away entirely and we end up with pure unadulterated reality.



The Pyramid of Scientific Evidence

And for a while this went well. People would drop things off towers, or see how quickly gases expanded, or observe chimpanzees, or whatever.

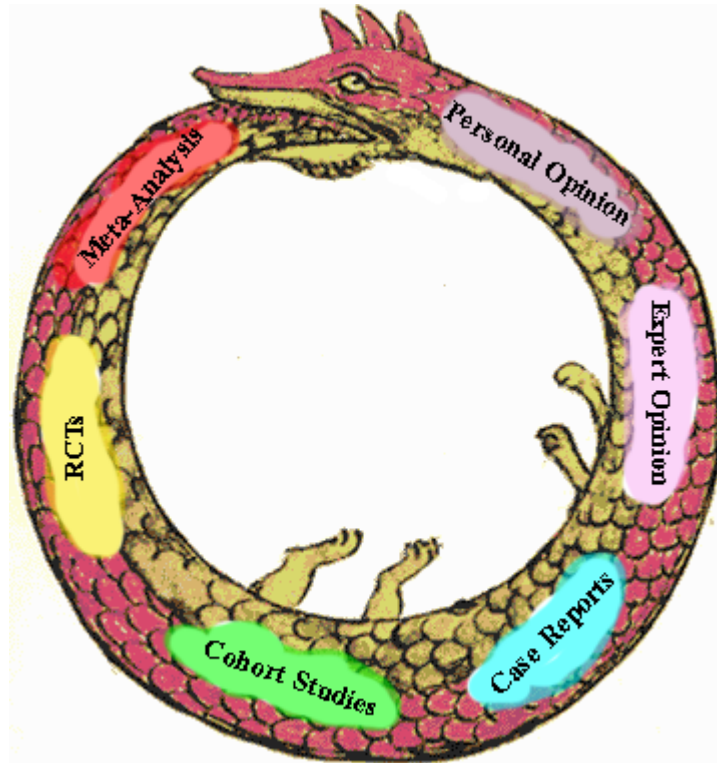
Then things started getting more complicated. People started investigating more subtle effects, or effects that shifted with the observer. The scientific community became bigger, everyone didn't know everyone anymore, you needed more journals to find out what other people had done. Statistics became more complicated,

allowing the study of noisier data but also bringing more peril. And a lot of science done by smart and honest people ended up being wrong, and we needed to figure out exactly which science that was.

And the result is a lot of essays like this one, where people who think they're smart take one side of a scientific "controversy" and say which studies you should believe. And then other people take the other side and tell you why you should believe different studies than the first person thought you should believe. And there is much argument and many insults and citing of authorities and interminable debate for, if not centuries, at least a pretty long time.

The highest level of the Pyramid of Scientific Evidence is meta-analysis. But a lot of meta-analyses are crap. This meta-analysis got $p < 1.2 \times 10^{-10}$ for a conclusion I'm pretty sure is false, and *it isn't even one of the crap ones*. Crap meta-analyses look [more like this](#), or even worse.

How do I know it's crap? Well, I use my personal judgment. How do I know my personal judgment is right? Well, a smart well-credentialed person like James Coyne agrees with me. How do I know James Coyne is smart? I can think of lots of cases where he's been right before. How do I know those count? Well, John Ioannides has published a lot of studies analyzing the problems with science, and confirmed that cases like the ones Coyne talks about are pretty common. Why can I believe Ioannides' studies? Well, there have been good meta-analyses of them. But how do I know if those meta-analyses are crap or not? Well...



The Ouroboros of Scientific Evidence

Science! YOU WERE THE CHOSEN ONE! It was said that you would destroy reliance on biased experts, not join them! Bring balance to epistemology, not leave it in darkness!



I LOVED YOU!!!!

Edit: [Conspiracy theory](#) by Andrew Gelman