Feeling the Future (Precognition Experiments)

ESP experiments that appeared to show a precognitive effect were reported in 2011 by American psychologist Daryl Bem under the heading 'Feeling the Future'. The studies broke new ground, being based on robust protocols widely used in psychology. Claims by critics that the findings could not be replicated and might be attributed to statistical bias and methodological flaws have been contested. This article by Chris Roe, professor of psychology at the University of Northampton, describes the experiments and assesses the sceptical critiques.

Introduction

Daryl J Bem is a social psychologist and emeritus professor at Cornell University. He received his PhD in 1964 from the University of Michigan and taught there before going on to join the faculty at Stanford, Carnegie-Mellon, Harvard, and Cornell. He had a distinguished career in psychology, then turned his attention to parapsychology; the self-perception theory of attitude formation and change has been named after him, and he was invited to co-author one of the core international psychology textbooks, known by generations of students as 'Hilgard and Atkinson'.[1]

Bem has also practised magic since childhood and was an early member of the Psychic Entertainers Association. It was because of this interest in methods of deception that he was invited in 1983 by Charles Honorton to review the protocols he had developed to test for ESP using the ganzfeld method, to see if the security precautions could be overcome by an expert magician. Bem was sufficiently impressed that he agreed to co-author a paper with Honorton if the protocol delivered above-chance results. The findings were highly significant, and so Bem published it in *Psychological Bulletin*,[2] which is ranked number 1 of 1,314 Psychology journals listed by Scimago Journal & Country Rank.[3]

Experimental Precursors

Bem's first independent research in parapsychology was reported in the conference paper *Precognitive Habituation: Replicable Evidence for a Process of Anomalous Cognition*.[4] It represented the culmination of a 'search for a straightforward laboratory demonstration of psi that could: (a) be observed using participants from the general population; (b) be conducted with no instrumentation beyond a desktop computer; (c) be evaluated by simple statistical tests; and (d) be replicated by any competent experimenter – including a skeptical one'.[5] He termed this the 'holy grail' for many psi researchers.[6]

As with later protocols used in the *Feeling the Future* paradigm, Bem's approach relied on taking a robust and well-known psychological phenomenon, then

changing the sequence of elements so that success at the task would be indicative of precognition. One supposed advantage of such a strategy is that empirical claims for psi effects should be protected from ad hoc methodological criticism, on the grounds that the methods have been broadly accepted in other areas of psychology, so that any criticisms here would necessarily undermine the claims for these other, mainstream phenomena.

The first phenomenon to be tested, the 'mere exposure' effect, was originally proposed by Zajonc[7] and describes the tendency for people to develop a preference for stimuli to which they have been exposed previously ('familiarity breeds liking', which may be a foundation assumption for the advertising industry). In the mere exposure protocol, the participant would be repeatedly exposed to a stimulus and, when later presented with two stimuli and asked to make a judgement as to which is more likeable, they would tend to select the previously exposed stimulus over the novel one. The effect occurs even where initial exposure is degraded or below the level of conscious awareness (for example, with very low illumination levels, or very short exposure times); indeed, effects may be even stronger where the participant has no conscious awareness of which stimulus they have been exposed to.[8] The effect has been described in more than 200 research articles[9] and occurs in animals as well as humans.

The mere exposure effect is explained in terms of a natural disposition to be wary of novel stimuli, which is experienced as an arousal response that is labelled as negative (for example, on encountering a spider that has never before been seen). However, repeated exposures to the stimulus generate a diminishing arousal response, to reflect the fact that previous encounters had no negative consequences (we have seen that same spider on previous journeys and nothing untoward has happened to us, so it does not make sense to waste resources on maintaining a high degree of alertness). When presented alongside a novel stimulus which elicits original levels of aversive arousal, the seen-before stimulus will seem less dislikeable. Where arousal is likely to be labelled positively rather than negatively (for instance with erotic images as stimuli), then a reduced arousal following repeated exposure would result in lower liking for that image in comparison with a new one.

Bem converted this into a precognition task by having participants make their judgement as to which of two images they preferred *before* they were repeatedly exposed to one of the two images (see Figure 1). Since the target image is selected randomly by the computer shortly after the participant has registered their preference, there is no obvious conventional mechanism to account for any effect. Nevertheless, Bem reported that more than 400 trials conducted by a number of researchers had yielded strong support both for the increased liking for negative stimuli (52.6% selection rate rather than chance expectation of 50%), and decreased liking for positive stimuli (48.0% selection rate).

 Conversion of the mere exposure protocol into a precognition task

Figure 1: Conversion of the mere exposure protocol into a precognition task

A mixture of high and low arousal images was taken from the International Affective Picture System (IAPS),[10] which consists of a set of digitized photographs that have been rated for valence and arousal by both male and female raters (male participants tend to rate the images less extremely than female participants). Images that were rated in IAPS as positive and high arousal (typically erotic) were regarded as quite mild, so these were supplemented with images from other internet sources. The 'precognitive habituation' effect that Bem was testing for did not occur with low-affect pictures, even though the mere exposure effect can be found with neutral, non-arousing stimuli.[11] He explains this in terms of the finding that mere exposure effects occur when there is a time interval of at least several minutes between the exposures and the preference judgments, whereas in the current procedure, the two events occur together within the same trial and are separated by only a few seconds. There were strong sex differences, with above-chance performance being entirely attributable to the female participants. This was explained in terms of stronger arousal responses by women to the stimuli (despite the stronger material for males).

Bem developed a two-item 'emotional reactivity scale' (with statements like 'In general, how intense are your emotional reactions to movies, videos, or photographs that are violent, scary, or gruesome?') and found that while a greater proportion of women than men scored as emotionally reactive, those men that did so produced the expected mere exposure effect, which offers a reasonable circumstantial case for Bem's explanation of the observed sex differences.

Bem also noticed that increasing the number of subliminal exposures of low arousal images from 4-8 up to 10-12 created a 'precognitive boredom effect'[12] in which an initial indifference to the stimulus (it was preferred 50% of the time, in line with chance expectation) became an aversion to it (selected 46.8% of the time). A further experiment with 200 participants confirmed the boredom effect, although only in participants who were low in arousability and in boredom tolerance. The effect could not be accounted for in terms of bias in the method of randomization. He concluded, 'the study reported here demonstrates that participants low in Arousability or Boredom Tolerance show precognitive aversion on low-arousal pictures. Perhaps everyone can display psi when the task matches his or her personality'.[13]

In a 2008 presentation at the Parapsychological Association Convention, Bem introduced three additional time-reversed psychological effects: Precognitive Approach/Avoidance, Retroactive Priming, and Precognitive Memory. For the first of these, the participant is shown a picture and its mirror image side by side and asked to indicate which they prefer. The stimuli were low-arousal, emotionally neutral pictures from the IAPS set, and mirror images were used to ensure that alternatives were exactly matched for likability. After a selection had been made, the computer would randomly designate one of the alternatives as the target; if the participant had selected the target they would be 'rewarded' by the subliminal presentation of a positively-valenced picture, but if they had selected the non-target they would be 'punished' by being subliminally presented with a negatively-valenced picture. Analysis of data from 150 participants showed a significant

preference (51.6%) for the target images that led to subliminal rewards. This represents a time reversed version of a standard reinforcement paradigm.

Experiment 2 is derived from a standard priming paradigm that has been used extensively in psychology.[14] The participant is presented supraliminally with a target image and is asked to respond as quickly as possible (but without errors) to indicate whether the image is positive or negative, for example by pressing respectively a left key or a right key. Immediately before the image is presented, a positive or negative word (known as the 'prime') is presented, usually so briefly that the participant reports only seeing a flash of light rather than recognizing what word has been presented. Nevertheless, the meaning of the prime affects the participant's reaction times in responding to the image; where the prime is congruent with the overt image (for instance, the word 'beautiful' followed by an image of some flowers) the participant's reactions are typically quicker than they would be without a prime, but where the prime is incongruent (for instance the word 'disgusting' followed by an image of some flowers), the correct response is slowed down relative to reaction times without a prime. Such 'semantic' priming effects are well established and considered to be robust.[15]

Bem included a 'classic' priming task, but also a condition in which the elements were reversed so that participants were presented with the image first, and only after they had reacted to it were they presented with a subliminal prime — by which point, of course, it would be too late for the prime to affect reaction times by any conventional means. Bem reports that data from 100 participants showed that they were on average 21 milliseconds faster on congruent trials than on incongruent trials with the 'classic' priming task, but that they were also 15 milliseconds faster on the time-reversed (precognitive) version.

The third experiment looks at the effects of practice on word recall. A staple of psychology undergraduate research methods classes, the basic effect is that participants recall more items on a list of presented words if they have had an opportunity to 'practise' them, particularly if they process them more deeply, such as by finding ways in which the words might be linked semantically, a technique known as *clustering*.[16] This can be demonstrated by only allowing participants to practise some of the presented words and then showing that they recall more of these than the words that are presented but not practised. In Bem's time-reversed version, participants are given a chance to practise with 24 of 48 presented words, but only *after* they have completed the recall task. This seems akin to sitting an exam and then revising for it afterwards. Nevertheless, Bem again reports evidence of a precognition effect, with his 100 participants recalling more of the to-be-practised words than the control words.

'Feeling the Future' Article

The experiments described above was reported in conference presentations and published as small-circulation conference proceedings, so had not been subject to the intensive peer review process that is intended to maintain scientific standards or methodological quality.[17] They were included with other experimental data in a summary paper titled 'Feeling the Future',[18] published in 2011 in the *Journal of*

Personality and Social Psychology, which encourages papers that report on experimental series rather than individual studies. Here, Bem reported on nine formal experiments in the sequence given in Table 1. The listing starts with simpler designs for ease of exposition rather than reflecting chronological order, and so does not readily map onto the experiments described in Bem's conference papers (he only cites the 2003 publication). I have indicated provenance where that was possible.

The first experiment, involving the detection of erotic stimuli, is not really a time-reversed standard protocol but is based on traditional forced choice ESP testing methods. Participants are presented with two curtains and are asked to choose the one they believe conceals a picture. The curtains of their selection are opened to reveal either a picture or a blank space. The pictures could be erotic or non-erotic images (some trials used positive, neutral, or negative images, and different 'strength' erotic images were used for men and women). Whether or not their curtain revealed a picture, and if so what type, were determined randomly by the computer after they had made their selection, so this was a precognitive design. By chance alone, participants should select the curtain that concealed a picture 50% of the time; for trials with erotic images (which were presumed to be desirable), they selected the correct curtain 53.1% of the time, which was statistically significant.

Experiment 2 looks to be the avoidance experiment described earlier,[19] although the stimuli are referred to as 'closely matched pictures' rather than mirror images of the same picture. In a laudable attempt at transparency, Bem describes changes to the protocol after the first 100 of 150 trials were complete, intended to capitalize on observed preferences. However, this proved to be a source of concern for researchers who regard all elements of a protocol as fixed for the duration of an experiment.

Experiment 3 is the priming study described earlier.[20] The later report includes more analysis permutations to reflect different ways of dealing with the non-normal distribution of reaction time data, but the outcome remains substantially the same.

Experiment 4 is a previously unreported replication of the priming effect, but with the primes being semantically related to the accompanying picture (so that, for example, the positive and negative primes for a picture of a basket of fruit are 'luscious' and 'bitter') rather than a random positive or negative word. The precognitive effect was confirmed, but the adjustment did not increase the effect size as hoped.

Experiments 5 and 6 describe the reversed mere exposure effect (named here 'retroactive habituation') that we began with.[21] However, it is difficult to reconcile the two descriptions: in Bem's 2003 study, series 100 comprises 104 women and 49 men, series 200 had 52 women and 48 men, and series 300 included 62 participants with gender unspecified for a total of 315. Here, in contrast, study I has 63 women and 37 men, and study II has 87 women and 63 men, giving a total of 250. Percentage selection rates for to-be-exposed negative and erotic images are dissimilar across the two reports. However, these differences do not affect the conclusions that can be drawn.

Experiment 7 tests Retroactive Induction of Boredom, and is a recasting of the precognitive boredom effect discussed earlier.[22] This is the only experiment in this suite that did not produce a significant outcome in support of the primary hypothesis, but it still found the predicted effect for participants who were high on 'stimulus seeking' (which I take to be equivalent to emotionally reactive, as described earlier). A second iteration of this experiment was not conducted.

Experiments 8 and 9 concern Retroactive Facilitation of Recall. The first of these was previously reported in Bem's 2008 report and has been described earlier. To recap, participants did recall significantly more of the to-be-practised words than the control words.

Experiment 9 was a replication but with an additional practice exercise, which re-presented the target words organized according to their suggested categories (foods, animals, etc.). The sample was smaller (50 participants in total rather than the usual 100) but the effect was stronger and independently significant, confirming the precognitive effect.

In summary, then, the article presents nine discrete experiments, of which eight were independently significant. The mean effect size, d, was .22, which is very similar to effect sizes reported for other psi phenomena.[23] Bem ends with advice to those who may be interested in replicating his findings, particularly drawing attention to the relationship between study power and likelihood of finding a significant effect.

Phenomenon tested and experiment	First reported	Sample size	d full sample	p full sample
Precognitive approach/avoidance				
1. Detection of Erotic Stimuli		50M 50F	.25	.01
1. Avoidance of Negative Stimuli	Bem (2008) Exp 1	43M 107F	.20	.009
Retroactive priming				
1. Retroactive Priming I	Bem (2008) Exp 2	31M 69F	.26	.007
1. Retroactive Priming II		43M 57F	.23	.014

Retroactive habituation				
1. Retroactive Habituation I Negative trials	Bem (2003) series 200?	37M 63F	.22	.014
1. Retroactive Habituation II Negative trials	Bem (2003) series 100?	63M 87F	.15	.037
1. Retroactive Induction of Boredom	Bem (2005)?	60M 140F	.09	.096
Retroactive facilitation of recall				
1. Facilitation of Recall I	Bem (2008) Exp 3	36M 64F	.19	.029
1. Facilitation of Recall II		16M 34F	.42	.002
Mean effect size (d)			.22	

Table 1: Results of the 'feeling the future' experimental series (adapted from Table 7 of Bem, 2011) and their provenance

Critical Reaction

The scientific community's reaction to the article's publication was mainly negative. A *New York Times* article[24] noted that 'the decision may delight believers in so-called paranormal events, but it is already mortifying scientists', and quotes the emeritus professor of psychology at the University of Oregon, Ray Hyman, a high-profile sceptic: 'It's craziness, pure craziness. I can't believe a major journal is allowing this work in'. Jarrett[25] included Bem's study among the '10 Most Controversial Psychology Studies Ever Published', alongside notorious research such as Zimbardo's Stanford Prison Experiment and Milgram's 'Shock Experiments'. Chambers[26] identified the publication of Bem's article as the tipping point that triggered a methodological and statistical crisis in psychology (rather than the dramatic high profile revelations of fraud perpetrated by psychologists such as Diederik Stapel or Jan Smeesters, the dismal failure of the Open Science Replication project, or the discovery that many psychologists admitted to questionable research practices, which are all considered in detail later on in his book).

Engber[27] described the research as 'both methodologically sound and logically insane' and quotes Wagenmakers' experience of reading Bem's ESP paper, 'I had to

put it away several times … Reading it made me physically unwell.' In the same article, Uli Schimmack, a psychologist at the University of Toronto asserted 'I don't have to believe any of these results because they're clearly fudged.' A subsequent article[28] attributes the findings to hypothesizing after results are known, on the grounds that the suite of experiments was more successful and consistent than should be expected statistically.

In attempting to explain this vociferous rejection of Bem's findings, Lacsap[29] observed

> after speaking to quite a few of my colleagues about this [paper], I realize that the willingness to take these results seriously – as opposed to dismissing them out of hand – is a function … of the PRIOR probability that such effects exist … People were bugged by the result, not the methodology. As a matter of fact, the experimental approach (with several substudies) would have passed muster in most fields, including psychology, without a second thought if the results had been more in line with expectations. No one would have batted an eye, no one would have attempted a replication. This should give those with a concern for the state of the field pause for thought. How many results that are wrong do we believe because we expect them to be true?

James Alcock

A more detailed critique is offered by psychologist James Alcock, a sceptical critic of parapsychology.[30] Alcock's general thesis is that any observed methodological shortcomings are indicative of more general sloppiness, so that even if the former could not account in themselves for the reported findings, a sceptic is justified in dismissing them as likely due to these other, undetected flaws: 'when one finds that the chemist began with dirty test tubes, one can have no confidence in the chemist's findings, and one must wonder about other, as yet undetected, contamination.'[31]

For experiment 1, Alcock reasonably draws attention to variations in the specifics of the task (such as how many stimuli of each type) across the series of trials, which is unusual but not unprecedented in an exploratory experimental series; it might have been preferable to characterize the first 40 trials as experiment 1a and the remaining 80 trials as experiment 1b. Unfortunately, he quickly descends to hyperbole: 'What is going on here?! Setting aside the confusion about the stimulus set, no competent researcher dramatically modifies an experiment two-fifths of the way into it! To do so is to seriously compromise any subsequent analysis and interpretation' [32] He also is bemused by the adoption of different stimuli for males and females; despite the rationale and precedent for this from mainstream studies that I have already discussed, Alcock finds it to be 'the most baffling description of research materials and procedures that I have ever encountered'.

Bem[33] responds by explaining how tailoring stimulus material for different participant cohorts is accepted practice for some lines of psychology research. He also challenges Alcock's tendency to make vague allusions to problems instead of stipulating how his identified flaws might have materially affected the study outcome, commenting, 'If Alcock believes that having different sets of erotic

stimuli for men and women or for gay and heterosexual participants is a flawed procedure, then he should spell out how and why he thinks this could possibly lead to false positive results.'

Alcock's main concern, however, is with how data are analysed, in particular with Bem's use of one-tailed instead of two-tailed predictions, which he argues hedges things in Bem's favour. This is a rather technical point, but bluntly, if the effect is in the expected direction then some outcomes will be classed as significant by 1-tailed test that would be regarded as non-significant by 2-tailed test.[34] These days, there is a tendency in psychology to prefer a more conservative two-tailed approach even if there is a rationale for predicting the direction of effect (partly on the grounds that if effects are so meagre they may be of little practical value). In this respect Alcock has a point in criticizing Bem for adopting one-tailed tests, because even though there is a directional expectation, there is not a strong empirical or theoretical basis to justify it for these exploratory experiments. Reverting to two-tailed tests would render three of the experiments marginally nonsignificant (Retroactive Habituation II Negative trials and Erotic trials, Facilitation of Recall I).

Alcock also raises concerns that Bem conducted multiple analyses without making a statistical correction, including post hoc testing for differential effects and recalculations that use nonparametric rather than parametric statistics to check that possible violations of parametric assumptions do not lead to spurious outcomes. Correcting for these would render the outcome nonsignificant, in his view. Alcock's solution is to use the Bonferroni method, but this has fallen out of favour, being regarded as an overly stringent response that adversely affects studies that are already statistically underpowered.[35]

In any case, Bem[36] argues that the objection did not apply to the actual analyses he conducted, commenting,

> It is illegitimate and misleading to perform multiple tests on a set of data without adjusting the resulting significance levels to take into account the number of separate analyses conducted. This is well known to experimental psychologists, but, in fact, it does not apply to any of the analyses in my article. Alcock has memorized the right words about multiple tests, but does not appear to understand the logic behind those words. … Perhaps Alcock wants me to change my conclusion that there were no significant effects on non-erotic pictures to the conclusion that there were really really no significant effects on non-erotic pictures.

Alcock also refers to a concern (initially communicated to him by sceptic colleague Ray Hyman) that across the series of experiments there was a very large negative correlation (-.91) between effect size and sample size, which is considered dubious since ordinarily we might expect to observe a positive correlation. However, this seems confused. Firstly, by definition an effect size is independent of sample size; a reported effect size is simply an estimate of the actual effect size in the population, with larger-sample studies more closely approximating it, while smaller studies are more susceptible to sampling effects that can produce more variation in that estimate (this is the basic assumption of funnel plots as used in meta-analyses).

Alcock may be mistaking effect size for *significance*: for a given effect size, we should expect to see increasing significance as the sample size increases.

Secondly, the calculation of a correlation is highly suspect when there is a lack of variance in the values being correlated. While Bem has one study with 50 trials and one with 200 trials, the others are all either 150 trials (two cases) or 100 trials (five cases). In my view, this is not sufficiently varied to give a meaningful outcome from a correlation analysis.

Thirdly, changes to the sample sizes are not the only differences between the studies – the 200-trial study is the unsuccessful boredom study while the last study in the series, a 50-trial recall facilitation experiment gave an effect that was nearly twice as large as for any other experiment. Correlating these seems rather like comparing apples with oranges.

There are no other concerns raised by Alcock that are substantial enough to warrant consideration here. It is surprising, then, that Ritchie, Wiseman and French[37] felt justified in claiming that 'Alcock (2011) … has outlined numerous experimental flaws in [Bem's] design'; tellingly, they do not elaborate on that claim.

Splitting Trials

The unequal numbers of trials across different experiments does raise a more substantial concern, however. Yarkoni[38] explains:

> There's some reason to think that the 9 experiments Bem reports weren't necessarily designed as such. Meaning that they appear to have been 'lumped' or 'splitted' post hoc based on the results. For instance, Experiment 2 had 150 subjects, but the experimental design for the first 100 differed from the final 50 in several respects. They were minor respects, to be sure (e.g., pictures were presented randomly in one study, but in a fixed sequence in the other), but were still comparable in scope to those that differentiated Experiment 8 from Experiment 9 (which had the same sample size splits of 100 and 50, but were presented as two separate experiments). There's no obvious reason why a researcher would plan to run 150 subjects up front, then decide to change the design after 100 subjects, and still call it the same study. A more plausible explanation is that Experiment 2 was actually supposed to be two separate experiments (a successful first experiment with N = 100 followed by an intended replication with N = 50) that was collapsed into one large study when the second experiment failed–preserving the statistically significant result in the full sample. Needless to say, this kind of lumping and splitting is liable to additionally inflate the false positive rate.

The fact that each experiment is a multiple of 50 suggests that blocks were pre-planned and do not involve optional stopping mid-series. However, following Yarkoni, it would have been preferable to have a uniform block of trials (i.e., always 50 or 100 for a discrete experiment) for each formal experiment so as to field against concerns about lumping or splitting trials.

Wagenmakers et al

Wagenmakers, Wetzels, Borsboom, and van der Maas[39] criticized Bem's work on the grounds that it treated what were essentially a series of exploratory studies as if they were confirmatory ones. The former designs have licence to explore the data quite liberally in order to identify potentially interesting patterns and would be free to consider serendipitous effects that had not been predicted in advance. However, those observations would not in themselves count as evidence of the effects and would need to be confirmed in follow-up studies as formal pre-specified predictions about new data. They claim that Bem's experiments 'highlight the relative ease with which an inventive researcher can produce significant results even when the null hypothesis is true'.[40] Indeed, the ways in which Bem divides the data (for instance, by stimulus type, by gender, by sensation seeking) does give the impression of an attempt to find a multivariate 'recipe for success'. This is a reasonable strategy if it is subsequently confirmed in an independent replication, but in itself is highly susceptible to inflating the likelihood of finding a significant effect somewhere.[41]

Wagenmakers et al also assert that the statistical approach adopted by Bem (and common to most psychological research) overstates the evidence against the null hypothesis, particularly where sample sizes are relatively large. They prefer a Bayesian analysis which gives an estimate of the prior probability of a given effect and calculates how that probability shifts as a result of the observed data. Of ten critical tests they conducted, three yielded 'substantial' evidence in favour of the null hypothesis, six provided evidence in favour of an effect that was only 'anecdotal', and only one (Facilitation of Recall II) gave 'substantial' evidence for an effect, leading them to conclude that 'Bayesian reanalysis of Bem's experiments … demonstrated that the statistical evidence was, if anything, slightly in favor of the null hypothesis'.[42]

Bem, Utts and Johnson[43] responded, arguing that the authors incorrectly selected an unrealistic prior distribution for their analysis and that a Bayesian analysis using a more reasonable distribution yields strong evidence in favour of the psi hypothesis. The arguments are technical, but essentially psi studies tend to give an average effect size in the range .15-.25, which is broadly comparable to effect sizes for psychology as a whole, whereas Wagenmakers et al assumed that if the null hypothesis were false (there was a real effect size) there was more than a 50% likelihood that the effect size would be greater than 0.8. When a more realistic 'knowledge-based' prior is used, five of the nine experiments gave either 'strong' or 'substantial' evidence in favour of an effect, and the combined Bayes factor greatly exceeds Wagenmakers et al's criterion for 'extreme' evidence in favour of an effect.

Rouder and Morey[44] were also critical of Wagenmakers et al's analysis, arguing that it was inappropriate for making assessments across multiple experiments. Their corrected analysis, while more sympathetic to Bem's claim, was not sufficient to sway an appropriately skeptical reader (in part because of concerns about unreported failures to replicate). Wagenmakers et al[45] offered a rejoinder to an early draft of Bem, Utts and Johnson's response, but this has not been subject to peer review and offers little of substance.

Replications

Many of the concerns raised about the possible exploratory nature of Bem's experiments can be resolved by independent replication. A high-profile failure to replicate was reported by Ritchie, Wiseman and French.[46] They focused on retroactive facilitation of recall (Bem's experiments 8 and 9), with each author overseeing an independent study involving 50 participants. All trials were conducted in-person, either by the author or a research assistant / student (as was the case for Bem's original experiments). The experimental design is said to have been pre-registered, but no details are available (the published link does not work and the project is not included in the Koestler Parapsychology Unity Study Registry).

All three experiments are reported to be nonsignificant; in two cases this is because the mean difference in recall for practice words and control words is very small, but replication 2 gives a 1-sample t-test value of 1.57, which is a suggestive effect. The authors regard this as nonsignificant because the effect is in the opposite direction to prediction (participants recalled more of the control words than practice words) and so would be rejected by a 1-tailed test. However, it seems an odd decision to adopt 1-tailed tests given that they echo criticisms of Bem for using them, especially when experimenter effects linked to their scepticism of psi (versus openness to it) have been observed for other psi experiments – see Roe (2016)[47] for a fuller consideration. The t-value for the uncorrected weighted mean recall score is $t = 3.09$, which for a sample of 50 participants would give a highly significant ($p < .005$) missing effect even if corrected for multiple analyses. Nevertheless, it is clear that none of these replication attempts confirmed Bem's original findings.

Ritchie, Wiseman & French submitted a report for publication in the *Journal of Personality and Social Psychology* and were surprised when it was rejected. They attributed this to journal editors having little appetite for publishing failures to replicate, though the journal also rejected submissions that claimed to support the Bem findings.[48] While antipathy for null results may be generally true in the social sciences, and is likely to have had an impact on the published record as a whole by skewing it to the positive,[49] it is a surprising attribution to make in this case. The article seems unlikely to meet the journal's criteria that submissions will be evaluated on the basis of the statistical power of the study that is carried out, and the number and power of previous replications of the same finding. In this case, three low powered experiments have little prospect of providing an adequate refutation of the original studies.

I conducted a power analysis[50] to estimate the likelihood that a study with sample size 50 would produce an outcome that was significant at $p = .05$ (1-tailed) given an effect size d of .19 (as reported in Bem's experiment 8). This produced a power estimate of .37, meaning there is only a 37% chance that an individual study would successfully replicate the original effect where the effect is real but small. A simple binomial analysis indicates that a collection of three such studies would all be nonsignificant about 25% of the time. However, if we use the much larger effect size $d = .42$ from Bem's experiment 9, then the power of each replication attempt increases markedly to 90% and the likelihood that none of the 3 is independently significant reduces massively to 0.1%.

Nevertheless, the authors attracted a lot of media attention that was sympathetic to the claim that their initial publication difficulties were due to the mistreatment of failed replications, featuring for example in articles in *New Scientist*[51] *The Guardian*[52] and even in the *Stanford Encyclopedia of Philosophy*'s entry on 'Reproducibility of Scientific Results'.[53] The British Psychological Society's professional member magazine *The Psychologist* devoted an issue to concerns about replication that was opened by a summary of the Ritchie, Wiseman and French replication failure.[54]

A more substantial replication attempt, which thus addresses problems of interpretation caused by a gross lack of statistical power, was reported by Galak, LeBoef, Nelson and Simmons.[55] This comprised seven experiments and more than 3,000 participants, and focused on Bem's facilitation of recall effect, on the reasonable grounds that 'the other findings reported in Bem (2011)[56] hinge on nuanced affective responses' that can be 'be sensitive to subtle variation in the intensity and character of the stimuli'.[57] Here in contrast, participants are simply shown a list of words in the knowledge that they will subsequently be asked to recall as many as they can.

This series of experiments adheres broadly to Bem's approach, but nevertheless incorporates changes; for example, experiments 1, 2, 6 and 7 were conducted online, experiment 2 used (unspecified) different words and different categories, and experiment 6 included a 'standard' recall task. Participation pathways to the online experiment 7 – by far the largest study – included hyperlinks from an online report that described the original Bem study. It is not clear whether participants recruited by this method might have thereby been introduced to the set of test words used in the original study and (presumably) re-used here. Sample sizes for the seven experiments are 112, 158, 124, 109, 211, 175 and 2,469. The odd values are explained by a strategy of stopping once threshold values have been passed but does not explain why different threshold values were used in the first place. Given the criticisms of inconsistency lodged against Bem, these variations may be important.

Only one of the seven experiments – experiment 4 – showed a significant effect suggesting precognition (using a one-tailed p value), and the combined effect was very close to zero. Interestingly, the three experiments conducted in-person gave t values of 1.28, 1.77, and -0.71 (Bem's original recall experiments gave t values of 1.92 and 2.96), whereas the online experiments gave t-values of -1.20, 0.00, -.33 and -.23. Considering just the in-person experiments gives a positive but non-significant effect size of 0.07 ($Z = 0.940$, $p = 0.347$).[58]

From this it seems that adopting an online protocol is not a valid variation. Online research clearly has a number of advantages, particularly with respect to generating large samples of participants and enabling people to participate at times that are convenient for them. However, marked disadvantages result from participation not being monitored at any level by an experimenter: there are no checks that participants are attending to the task to the exclusion of all distractions; it is not possible to verify that participants are not cheating by writing down the words as they appear; there is no facility to check whether participants are selectively withdrawing from the study (for example, if they discover that the words they have

recalled are not among the words they subsequently have a chance to practise). To their credit, Galak et al attempted to gauge participant attentiveness, but the approaches they incorporated (to ask people if they had been attentive, and to measure how long it took to complete the task) seem naïve and crude respectively. Until more effective methods have been built into their designs, data collected online is likely to remain of dubious validity.

Galak et al[59] additionally presented a meta-analysis of all replication attempts to date, including their own suite of experiments and the replication failures by Ritchie et al.[60] All studies in the database involved the facilitation of recall effect, and all were in-person tests apart from the four experiments by Galak et al described above. The overall average effect size of .04 is considerably smaller than Bem's (2011) average effect size (.29) and is not statistically different from zero. This effect size is weighted by sample size and may have been disproportionately affected by Galak et al's experiment 7, which had 2,469 participants (over 60% of the total; when this one study is excluded, the median sample size reduces to 85.5 and the mode is only 50.). I have already noted that there may be issues with online experiments; indeed, a separate analysis by Galak et al[61] that excludes the online experiments gives a significant effect of .09.

In 2016, Bem, Tressoldi, Rabeyron and Duggan[62] published a more comprehensive meta-analysis that encompassed all the 'feeling the future' protocols. They retrieved 69 attempted replications as well as 11 other experiments that 'tested for the anomalous anticipation of future events in alternative way'. If Bem's original studies are included, the total sample comprises 90 experiments from 33 different laboratories located in 14 different countries, and involved 12,406 participants. The replications should resolve some of the controversy surrounding Bem's original work, since they were designed from the outset as confirmatory studies that were constrained to test for the specific effects described by Bem – 31 are described as 'exact replications' and 38 as 'modified replications'. The overall effect size (Hedges' g) is 0.09, which is significant ($p = 1.2 \times 10^{-10}$) and is interpreted by the authors as 'decisive evidence for the experimental hypothesis'.[63] Even when Bem's original experiments are removed from the analysis, the result remains highly significant.

There were differences in outcome across experiment-types, with 'fast-thinking protocols' which require quick judgements that do not allow time for reflection (such as the priming task) producing larger effects than the 'slow-thinking protocols' (such as memorizing and recalling words). It is interesting to note that the flurry of failures to replicate Bem's findings had all focused on this latter task type.

Concerns about selective reporting are tested by comparing outcomes from peer reviewed publications with 'unpublished' studies (including conference proceedings); these did not differ in outcome, suggesting there was no overt publishing bias. It is possible to calculate the number of unpublished studies that average a null result which would be needed to cancel out the observed effect; in this case there would need to be more than 1,000 unpublished experiments, which is extremely unrealistic.

Conclusion

In summary, Bem's original proposal to adapt well-established psychology protocols so that they become a test for precognition is laudable. While it did not protect the work from methodological criticism, it has encouraged researchers who would not normally get involved in parapsychological research to conduct replication attempts. Some of the criticisms of Bem's stimulus paper are without merit, but others have legitimately drawn attention to inconsistencies and ambiguities in the way the studies were conducted and organized.

Particular concerns around differentiating between exploratory and confirmatory studies can be resolved by independent replication attempts. Arguably, sceptical critiques have attributed more weight to the three high-profile replication attempts than they deserve; in particular, the initial (negative) meta-analysis was compromised by the inclusion of online experiments with extremely large sample sizes. A more sophisticated understanding of the relationship between effect size and study power could have led to a more realistic understanding of the likelihood of achieving statistical significance where one is testing for a small but robust effect. The more recent meta-analysis claims that effects can be replicated statistically and provides useful indicators for the next wave of replication attempts, particularly to map and explain the apparent advantage of fast-thinking protocols.

Chris Roe

Literature

Atkinson, R.L., Atkinson, R.C., Smith, E.E., Bem, D.J., & Nolen-Hoeksema, S. (2000). *Hilgard's Introduction to Psychology* (13th ed.). Harcourt College Publishers.

Alcock, J. (2011). Back from the future: Parapsychology and the Bem Affair. *Skeptical Inquirer*, March/April.

Aldhous, P. (2011). Journal rejects studies contradicting precognition. *New Scientist*, 5 May 2011.

Baptista, J., Derakhshani, M., & Tressoldi, P. E. (2015). Explicit anomalous cognition: A review of the best evidence in ganzfeld, forced choice, remote viewing and dream studies. In E. Cardeña, J. Palmer, & D. Marcusson-Clavertz (Eds.), Parapsychology: A handbook for the 21st century (192-214). Jefferson, NC: McFarland.

Bem, D.J. (2003). Precognitive habituation: Replicable evidence for a process of anomalous cognition. *Proceedings of Presented Papers: The Parapsychological Association 46th Annual Convention*, 6-20.

Bem, D.J. (2004). Precognitive Avoidance and Precognitive Déjà vu. *Proceedings of Presented Papers: The Parapsychological Association 47th Annual Convention*, 431-32.

Bem, D.J. (2005). Precognitive Aversion. *Proceedings of Presented Papers: The Parapsychological Association 48th Annual Convention*, 31-35.

Bem D.J. (2008). Feeling the future III: Additional experimental evidence for apparent retroactive influences on cognition and affect. *Proceedings of Presented Papers: The Parapsychological Association 51st Annual Convention*, 24-32.

Bem, D.J. (2011a). Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology* 100, 407-25.

Bem, D.J. (2011b). Response to Alcock's 'Back from the Future: Comments on Bem'. *Skeptical Inquirer*, March/April.

Bem, D. J., & Honorton, C. (1994). Does psi exist? Evidence for an anomalous process of information transfer. *Psychological Bulletin* 115, 4-18.

Bem, D.J., Tressoldi, P.E., Rabeyron, T., & Duggan, M. (2016). Feeling the Future: A Meta-Analysis of 90 Experiments on the Anomalous Anticipation of Random Future Events. F1000Research, doi: 10.12688/f1000research.7177.2

Bem, D.J., Utts, J., & Johnson, W.O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology* 101, 716-19.

Carey, B. (2011). Journal's Paper on ESP Expected to Prompt Outrage. Benedict Carey. *The New York Times*, Jan. 5.

Chambers, C. (2017). *The 7 deadly sins of psychology: A manifesto for reforming the culture of scientific practice.* Princeton University Press.

Engber, D. (2017). Daryl Bem proved ESP is real. Which means science is broken. *Slate Magazine*, May 17.

Fidler, F., & Wilcox, J. (2018). Reproducibility of Scientific Results. *Stanford Encyclopedia of Philosophy*.

French, C. (2012). Precognition studies and the curse of the failed replications. *The Guardian*, 15 March.

Galak, J., LeBoef, R., Nelson, L., & Simmons, J. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology* 103, 933-48.

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. Technical Report A-8. University of Florida, Gainesville, FL.

Lascap, [no initial] (2010). Is reality retrocausal? *Pascal's Pensées* Posted on November 1, 2010. [web page]

Lockhart, R. S., & Craik, F. I. M. (1990). Levels of processing: A retrospective commentary on a framework for memory research. *Canadian Journal of Psychology/Revue canadienne de psychologie 44*/1, 87-112.

Meyer, D.E. (2014). Semantic priming well established. *Science* 345/6196, 523.

Monahan, J.L., Murphy, S.T., Zajonc, R.B. (2000). Subliminal mere exposure: Specific, general, and diffuse effects. *Psychological Science* 11, 462-66.

Mossbridge, J., Tressoldi, P., & Utts, J. (2012). Predictive physiological anticipation preceding seemingly unpredictable stimuli: A metaanalysis. *Frontiers in Psychology* 3, 390.

Perneger, T.V. (1998). What's wrong with Bonferroni adjustments. *BMJ* 316, 1236.

Ritchie, S.J., Wiseman, R., & French, C.C. (2012a). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLOS One*.

Ritchie, S.J., Wiseman, R., & French, C.C. (2012b). Replication, replication, replication. *The Psychologist* 25/5, May, 346-48.

Roe, C.A. (2016). Experimenter as subject: What can we learn from the experimenter effect? *Mindfield* 8/3, 89-97.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review* 18, 682-89. doi:10.3758/s13423-011-0088-7

Schlitz, M., & Delorme, A. (2021). Examining implicit beliefs in a replication attempt of a time-reversed priming task. F1000Research.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology* 13/2, 90-100.

Van den Bussche, E., Van den Noortgate, W., & Reynvoet, B. (2009). Mechanisms of masked priming: A meta-analysis. *Psychological Bulletin 135*/3, 452-77.

Wagenmakers, E-J., Wetzels, R., Borsboom, D., & van der Maas, H.L.J. (2011). Why psychologists must change the way they analyze their data: The case of psi — Comment on Bem (2011). *Journal of Personality and Social Psychology* 100, 426-32.

Wagenmakers, E-J., Wetzels, R., Borsboom, D., Kievit, R.A., & van der Maas, H.L.J. (2011). *Yes, psychologists must change the way they analyze their data: Clarifications for Bem, Utts, and Johnson (2011).* Unpublished manuscript.

Yarkoni, T. (2011, 10 January). The psychology of parapsychology, or why good researchers publishing good articles in good journals can still get it totally wrong [Web log post].

Zajonc, R.B. (1968). Attitudinal effects of mere exposures. *Journal of Personality and Social Psychology* 9, 1-27.

Zajonc, R.B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science* 10/6, 224-28. doi:10.1111/1467-8721.00154

Endnotes

Footnotes

- 1. e.g., Atkinson et al. (2000).
- 2. Bem & Honorton, (1994).
- 3. Scimago Journal & Country Rank https://www.scimagojr.com/journalrank.php?area=3200
- 4. Bem (2003) See https://dbem.org/pubs.html for a detailed list (does not include published conference papers)
- 5. Bem (2004), 431.
- 6. Bem (2003), 6.
- 7. Zajonc (1968).
- 8. Zajonc (2001).
- 9. Monaghan et al. (2000).
- 10. Lang et al. (2008).
- 11. Bornstein (1989).
- 12. Bem (2005).
- 13. Bem (2005), 35.
- 14. cf. Van den Bussche et al. (2009).
- 15. Meyer (2014).
- 16. Lockhart & Craik (1990).
- 17. PA conference submissions are peer reviewed, but time constraints mean that the review process may not be as thorough as it is for journal submissions.
- 18. Bem (2011).
- 19. Bem (2008), experiment 1
- 20. Bem (2008), experiment 2.
- 21. Bem (2003).
- 22. Bem (2005)
- 23. e.g., Baptista, Derakhshani & Tressoldi, 2015; Mossbridge, Tressoldi & Utts (2012).
- 24. Carey (2011).
- 25. Jarrett (2014).
- 26. Chambers (2017).
- 27. Engber (2017).
- 28. Schimmack (2012).
- 29. Lacsap (2010).
- 30. Alcock (2011).
- 31. Alcock (2011), 4.
- 32. Alcock (2011), 5.
- 33. Bem (2011b).
- 34. see, e.g., https://cxl.com/blog/one-tailed-vs-two-tailed-tests/.
- 35. see, e.g. Perneger (1998).
- 36. Bem (2011b).
- 37. Ritchie et al. (2012b).
- 38. Yarkoni (2011).
- 39. Wagenmakers et al (2012).
- 40. Wagenmakers et al (2012), 427.
- 41. see Steegen et al's 2016 multiverse analysis of data that illustrates how outcome and conclusions can be dramatically affected by quite subtle

changes to various analytic decisions.

- 42. Wagenmakers et al (2012), 431.
- 43. Bem et al. (2011).
- 44. Rouder and Morey (2011).
- 45. Wagenmakers et al. (2011).
- 46. Ritchie, Wiseman & French (2012a).
- 47. Roe (2016).
- 48. Aldhous (2011).
- 49. Schmidt (s2009).
- 50. at https://www.ai-therapy.com/psychology-statistics/sample-size-calculator
- 51. Aldhous (2011).
- 52. French (2012).
- 53. Fidler & Wilcox (2018).
- 54. Ritchie, Wiseman & French (2012b).
- 55. Galak et al. (2012).
- 56. Bem (2011).
- 57. Bem (2011), 934.
- 58. Thanks to Patrizio Tressoldi for calculating these statistics.
- 59. Galak et al. (2012).
- 60. Ritchie et al. (2012a).
- 61. Galak et al. (2012)
- 62. Bem et al. (2016)
- 63. Bem (2011), 7.

© Psi Encyclopedia