# Back from the Future: Parapsychology and the Bem Affair

Psychologist Daryl Bem has reported data suggesting that individuals' future experiences can influence their responses in the present. Careful scrutiny of his report reveals serious flaws in procedure and analysis, rendering this interpretation untenable.

JAMES E. ALCOCK

A flurry of media attention is being directed toward the prepublication distribution of Daryl Bem's forthcoming research paper "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect."[1] Bem claims to have found evidence of marvelous psychic abilities that transcend time and allow the future to reach backward to change the past. Both the academic stature of its author, a respected emeritus professor of psychology at Cornell University, and the fact that it was to be published in the American Psychological Association's (APA) *Journal of Personality and Social Psychology* have made this report particularly newsworthy.

Parapsychology has long struggled, unsuccessfully, for acceptance in the halls of science. Could this article be its breakthrough? After all, the article apparently provides evidence compelling enough to persuade the editors of the world's preeminent social-psychology journal of its worthiness. However, this is hardly the first time that there has been media excitement about new "scientific" evidence of the paranormal. Over the past eighty-odd years, this drama has played out a number of times, and each time parapsychologists ultimately failed to persuade the scientific world that parapsychological phenomena (psi) actually exist. Recalling George Santayana's now-clichéd dictum, "Those who cannot remember the past are condemned to repeat it," we should approach Bem's work using a historical framework to guide us. Consider the following:

1. In 1934, Joseph Banks Rhine published *Extra-Sensory Perception* (Rhine & McDougall, 1934/2003), summarizing his careful efforts to bring parapsychology into the laboratory through application of modern psychological methodology and statistical analysis. Based on a long series of card-guessing experiments, Rhine wrote: "It is independently established on the basis of this work alone that Extra-Sensory Perception is an actual and demonstrable occurrence" (p. 210). Elsewhere, he wrote: "We have, then, for physical science, a challenging need for the discovery of the energy mode involved. Some type of energy is inferable and none is known to be acceptable . . ." (166).

Despite Rhine's confidence that he had established the reality of extrasensory perception, he had not done so. Methodological problems with his experiments eventually came to light, and as a result parapsychologists no longer run card-guessing studies and rarely even refer to Rhine's work.

2. Physicist Helmut Schmidt conducted numerous studies throughout the 1970s and '80s that putatively demonstrated that humans (and animals) could paranormally influence and/or predict the output of random event generators. Some of his claims were truly extraordinary: for example, that a cat in a garden shed, which was heated only by a lamp controlled by a random event generator, was able—through psychokinetic manipulation—to turn the lamp on more often than would be expected by chance. Schmidt's claim to have put psi on a solid scientific footing garnered considerable attention, and his published research reported very impressive p values.[2] In my own extensive review of his work, I concluded that Schmidt had indeed accumulated impressive evidence that something other than chance was involved (Alcock 1988). However, I found serious methodological errors throughout his

**Excitement about Helmut Schmidt's research gradually dwindled to the point that his work became virtually irrelevant, even within the field of parapsychology itself.**

work that rendered his conclusions untenable, and the "something other than chance" was attributable to methodological flaws.

As with Rhine, excitement about Schmidt's research gradually dwindled to the point that his work became virtually irrelevant, even within the field of parapsychology itself.

3. The 1970s gave rise to "remote viewing," a procedure through which an individual seated in a laboratory can supposedly receive psychic impressions of a remote location that is being visited by someone else. Physicists Russell Targ and Harold Puthoff claimed that their series of remote-viewing studies demonstrated the reality of psi. This attracted huge media attention, and their dramatic findings (Targ and Puthoff 1974) were published in *Nature*, one of the world's top scientific journals. At first, their methodology seemed unassailable; years later, when more detailed information became available, it became obvious that there were fundamental flaws in procedure that could readily account for their sensational findings. When other researchers repeated Targ and Puthoff's procedure with the flaws intact, significant results were obtained; with the flaws removed, outcomes were not significant (Marks and Kamman 1978, 1980).

Add Targ and Puthoff to the list of "breakthrough" psi researchers whose work is now all but forgotten.

4. In 1979, Robert Jahn, then dean of the School of Engineering and Applied Science at Princeton University, established the Princeton Engineering Anomalies Research (PEAR) unit to study putative paranormal phenomena such as psychokinesis. Like Schmidt, Jahn was particularly interested in the possibility that people can predict and/or influence purely random subatomic processes. Given his superb academic and scientific credentials, his claims of success drew particular attention within the scientific community. When his laboratory closed in 2007, Jahn concluded that "over the laboratory's 28-year history, thousands of such experiments, involving many millions of trials, were performed by several hundred operators. The observed effects were usually quite small, of the order of a few parts in ten thousand on average, but they compounded to highly significant statistical deviations from chance expectations" (PEAR, n.d.).

However, parapsychologists themselves were among the most severe critics of his work, and their criticisms were in line with my own (Alcock 1988). More importantly, several replication attempts have been unsuccessful (Jeffers 2003), including a large-scale international effort led by Jahn himself (Jahn et al. 2000).

5. In the1970s, the ganzfeld, a concept borrowed from contemporaneous psychological research into the effects of sensory deprivation, was brought into parapsychological research. Parapsychologists reasoned that psi influences may be so subtle that they are normally drowned out by information carried through normal sensory channels. Perhaps if a participant were in a situation relatively free of normal stimulation, then extrasensory information would have a better opportunity to be recognized. The late Charles Honorton carried out a large number of ganzfeld studies and claimed that his meta-analysis[3] of this work substantiated the reality of psi. Hyman (1985) carried out a parallel meta-analysis that contradicted that conclusion. Hyman and Honorton (1986) subsequently published a "joint communiqué" in which they agreed that the ganzfeld results were not likely to be due to chance, but they thought that replication involving more rigorous standards was essential before final conclusions could be drawn.

Daryl Bem subsequently published an overview of ganzfeld research in the prestigious *Psychological Bulletin* (Bem and Honorton 1994), claiming that the accumulated data were clear evidence of the reality of paranormal phenomena. That effort failed to be convincing, in part because a number of meta-analyses have been carried out since then with contradictory results (e.g., Bem et al. 2001; Milton and Wiseman 1999). Recently, the issue was raised again in the pages of *Psychological Bulletin*, with papers from Storm et al. (2010a, 2010b) and Hyman (2010). While Storm and coworkers argued that their meta-analyses demonstrate paranormal influences, Hyman pointed to serious shortcomings in their analysis and reminded us that the ganzfeld procedure has failed to yield data that are capable of being replicated by neutral scientists.

Because of the lack of clear and replicable evidence, the ganzfeld procedure has not lived up to the promise of providing the long-sought breakthrough that would lead to the acceptance of psi by mainstream science.

Add Honorton (and Bem the first time around) to the list.

The lesson in this history is that new claims of impressive evidence for psi

should give one pause. Early excitement is often misleading, and as Ray Hyman has pointed out, it often takes up to ten years before the shortcomings of a new approach in parapsychological research become evident.

One must also keep in mind that even the best statistical evidence cannot speak to the *causes* of observed statistical departures. Statistical deviations do not favor arbitrary pet hypotheses, and statistical evidence cited in support of psi could as easily support other hypotheses as well. For example, if one conducted a parapsychological experiment while praying for above-chance scoring, statistically significant outcomes could be taken as evidence for the power of prayer just as readily as for the existence of psi.

Another key consideration is that parapsychology's putative phenomena are all negatively defined: to claim that psi has been detected, all possible normal influences must be ruled out. However, one can never be certain that all normal influences have been eliminated; the reader of a research report has only the experimenter's word for it.

This point brings us to a related concern. Research reports involve an implicit social contract between experimenter and audience. The reader can evaluate only what has been put into print and must presume that the researcher has followed the best practices of good research. We assume that the participants did actually participate and that they were not allowed to use their cellular telephones during the experiment or to chat with other participants. We assume that they were effectively shielded from cues that might have inappropriately influenced their responses. We assume that the data were as reported—that none were thrown out because they did not suit the experimenter—and that they were analyzed appropriately and in the manner indicated. We assume that equipment functioned as described and that precautions reported in the experimental procedure were carefully followed. We take for granted that the researcher set out to test particular hypotheses and did not choose the hypotheses after looking at the data. We must take all this on faith, for otherwise any research publication might simply be approached as a blend of fact, fantasy, skill, and error, possibly reflecting little more than the predilections of the researcher. Obvious methodological or analytical sloppiness indicates that the implicit social contract has been violated and that we can no longer have confidence that the researcher followed best practices and minimized personal bias. As Gardner (1977) wrote, when one finds that the chemist began with dirty test tubes, one can have no confidence in the chemist's findings and must wonder about other, as yet undetected, contamination. So, when considering Bem's present research, not only do we need to look at the data, but—following the metaphor—we need to assess whether Bem used clean test tubes.

### Bem's Research

Bem describes a series of nine experiments that "test for retroactive influence by 'time-reversing' well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur." His stated goal is "to provide well-controlled demonstrations of psi that can be replicated by independent investigators." He defines *psi* as denoting "anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms."

### EXPERIMENT 1: Precognitive Detection of Erotic Stimuli

Each trial in this experiment involved the presentation of an erotic, negative, or neutral picture. The participant sat in front of a computer screen and was tasked to predict which of two curtains had a picture behind it. Only after the participant had chosen a curtain by depressing a key did the computer select a picture at random and present it behind either the left or the right curtain.

Each participant was presented with thirty-six of these trials and was given feedback on each one. The erotic pictures were considered to be "explicit reinforcement for correct 'precognitive' guesses," although no effort was made to determine whether they were indeed reinforcing anything. The main hypothesis was that participants would be able to identify the position of the hidden erotic picture significantly more often than by chance.

So far, clear enough. But then things become quite messy: we learn that "most" of the pictures used in the experiment were selected from a databank, the International Affective Picture System. Bem then states that each session (a "session" refers to all the trials of an individual participant) involved thirty-six trials of randomly intermixed erotic and non-erotic pictures (eighteen of each). However, we soon learn that not all sessions were conducted in this way: the first forty of the one hundred sessions (that is, those of the first forty participants) involved twelve trials of erotic pictures, twelve of negative pictures, and twelve of neutral pictures! (The distinction between the "non-erotic" pictures seen by the majority of the participants and the "neutral" pictures seen by only the first forty is unclear.) To muddle things even more, Bem then states that the remaining sixty sessions involved "18 trials of erotic pictures and 18 trials of non-erotic positive pictures *with both high and low arousal ratings. These included pictures featuring couples in romantic but non-erotic situations*. . ." (emphasis added). How many were of high or low arousal weighting, or what those terms even mean, he does not say.

What is going on here? Setting aside the confusion about the stimulus, 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.

But that is not all. Bem next indicates that in all the experiments using highly arousing erotic or negative stimuli, "a relatively large number of non-arousing trials must be included to permit the participant's arousal level to 'settle down' between critical trials. This requires including many trials that do not contribute directly to the effect

**We then find that participants were allowed to choose their target set! This is one of the most baffling descriptions of research materials and procedures that I have ever encountered.**

being tested." This leaves us not knowing how many trials were actually run and wondering by what method the researcher determined the number of non-arousing trials that were needed to ensure that even the most randy of participants would "settle down."

So by this point, it is not clear how many trials were actually presented to each participant or even whether they all received an equal number of trials. It is unclear just what the stimulus materials were, and we are faced with a procedure that was changed partway through the experiment.

Just when one thinks that this study cannot be made any more confusing, Bem informs us that he discovered in Experiment 5 (which turns out to have been conducted prior to Experiment 1!) that "women showed psi effects to highly arousing stimuli but men did not." In light of this odd complication, Bem states that "we introduced different erotic and negative pictures for men and women in subsequent studies, *including this one* using stronger and more explicit images from Internet sites for the men. We also provided two additional sets of erotic pictures so that men could choose the option of seeing male-male erotic images and women could choose the option of seeing female-female erotic images" (emphasis added).

By now, a careful reader is totally confused as to what went on in this experiment. Now, we find that participants were allowed to choose their target set! This is one of the most baffling descriptions of research materials and procedures that I have ever encountered.

In reflecting on the extremely unusual change in procedure during the experiment—when the appropriate course would be to run two different experiments—one cannot help but wonder if two experiments were indeed run, and when each failed to produce significant results the data from them were combined with the focus shifted to only the erotic pictures common to all participants. Surely that was not done, for such an action would make a mockery of experimental rigor.

*Data Analysis*: Bem states that "the main psi hypothesis was that participants would be able to identify the position of the hidden erotic picture significantly more often than chance (50%)." At first, this claim is puzzling. Although sixty of the participants completed eighteen trials with erotic pictures and eighteen trials with "non-erotic positive pictures"—therefore making the chance outcome 50 percent of the thirty-six trials—the other forty participants received twelve trials with erotic pictures, twelve with negative pictures, and twelve with neutral pictures. For them, the chance outcome would be 33.3 percent. However, it turns out that Bem combined the data for success or failure, but on the erotic pictures only, from all one hundred sessions (i.e., from all one hundred participants) and then applied t-tests[4] to assess whether identification of the future position of erotic pictures occurred significantly more frequently than the 50 percent rate expected by chance. We are also informed that the hit rate for non-erotic pictures—whether they were neutral, negative, positive, or romantic and non-erotic—did not differ significantly from chance. (This is the first mention of "romantic but non-erotic," which adds to the confusion.)

Now we have learned that the focus of the experiment is on the erotic pictures presented to the participants, but no information is provided regarding how participants with three choices scored on erotic pictures as compared with those who had only two choices; one wonders why this is so.

The data analysis was conducted through multiple t-tests without any correction for that multiplicity. We are informed that there were at least seven such t-tests, and the only significant outcome was that the one hundred participants "identified the future position of erotic pictures significantly more frequently than the 50% hit rate expected by chance: 53.1%." This was stated to be statistically significant at $p = .01$. However, that significance level is simply incorrect. This kind of error (Type I)[5] increases with the number of t-tests conducted, and given that there were at least seven such t-tests with a criterion of $p \leq .01$, the actual probability associated with each of these t-tests is $1 - (.99)^7 = .06$ one-tailed.[6] Thus, none of these t-tests is actually statistically significant, not even at a more generous .05 p value. It is simply unacceptable that Bem did not correct for multiple testing, despite indications later in his report that he is familiar with one such correction technique, the Bonferroni t-test.[7]

Another reason for concern is Bem's deliberate use of one-tailed t-tests, which provide a simpler criterion to meet than the two-tailed tests generally employed by parapsychologists. (Parapsychologists typically interpret both above-chance and below-chance scoring as indicative of psi, and thus they do not make specific predictions about the direction of the extra-chance scoring.) When we say that something is significant at the .01 level two-tailed, this means that we would expect these results to occur by chance alone only 1 percent of the time. But, given that either above-chance or below-chance results are considered to be meaningful, this 1 percent must be distributed in both directions. Thus, above-chance results would be significant at the .01 level two-tailed only if they are so extreme that they would be expected to occur by chance only half of 1 percent, or 0.5 percent, of the time. The same applies for below-chance results.

Bem also reports that he carried out a nonparametric binomial test[8] on the overall proportion of hits on erotic targets across all trials and sessions, but he offers no adequate rationale for using more than one type of significance test for the same data. The test is redundant and offers nothing beyond the t-test.

Then, after having examined the data, he introduces the possibility that introversion/extroversion may play a role in presumed precognitive ability. He suggests that it may be an extrovert's "susceptibility to boredom and the tendency to seek out stimulation" that underlies observed correlations between extroversion and psi performance reported in the literature. However, rather than using existing, well-documented measures of stimulus-seeking, he constructed his own such scale comprising two statements, reversed in scoring: "I am easily bored" and "I often enjoy seeing movies I've seen before." The content and construction of this scale is bewildering. Proper scale construction involves precise and often difficult work, including operationalization of the construct, finding items that can be demonstrated to relate to the construct, endeavoring to ensure that the increments on the response measure are of approximately equal size, and establishing satisfactory reliability and validity of the final scale. Bem has ignored these considerations. As a result, the arbitrary assignment of numbers to participants' responses on this "scale" is unjustified and misleading.

Nonetheless, Bem correlates responses on the scale with the participants' "psi scores" and reports a significant correlation, but only for those participants whose scores on his scale fall above the midpoint. Participants who score below the midpoint on the scale did not score significantly above chance on either erotic or non-erotic trials.

*Overall Evaluation*: Just about everything that could be done wrong in an experiment occurred here. And even if one chooses to overlook that methodological mess, Bem's data still do not support the claimed above-chance effect because of the multiple-testing problem.

It is difficult to have confidence that the other eight experiments, some of which were carried out earlier than the one just described, were conducted with appropriate attention to experimental rigor: We have toured the laboratory; we have found the dirty test tubes and the mislabeled vials; we have observed inappropriate methodology and analysis. We have lost confidence in the chemist, and there seems little need to poke about further.

Nonetheless, go on we must.

### EXPERIMENT 2: Precognitive Avoidance of Negative Stimuli

This study involved 107 female and forty-three male undergraduate students. Using a computer, each participant first responded to Bem's two-item stimulus-seeking scale and then completed a sequence of thirty-six trials in which a "low arousal affectively neutral" picture was presented side by side with its mirror image. The participant depressed a key to indicate which picture he or she liked better. Only after the preference was registered did the computer randomly choose which of the two pictures would be considered the "target." If the participant had chosen this target, the computer thrice flashed a reportedly subliminal "positively valenced picture." If the participant chose the non-target, then a "highly arousing negatively valenced picture" was flashed three times.

A hit was defined as choosing the "target-to-be." However, as in Experiment 1, the description of the situation is difficult to unravel: For the first one hundred sessions (the first one hundred participants), "the flashed positive and negative pictures were independently selected and sequenced randomly." Then there was a change in procedure. For the next fifty participants, "the negative pictures were put into a fixed sequence, ranging from those that had been successfully avoided most frequently during the first 100 sessions to those that had been avoided least frequently." When the participant selected what was to later be designated as the target picture, the positive picture was flashed, subliminally as before, and the negative picture was retained for the next trial. However, when the participant selected the non-target, "the negative picture was flashed and the next positive and negative pictures in the queue were used for the next trial."

This presents the same problem as before—the procedure has been changed partway through the experiment. Bem states that this was done to evaluate the possibility that "the psi effect may be stronger if the most successfully avoided negative stimuli were used repeatedly until they were eventually invoked." It is difficult to get one's head around this justification, and in any case, this should have been examined in a separate study. Again, given the inherent unreasonableness of changing the procedure in an ongoing experiment, one cannot help but wonder if two separate experiments were run and then combined after neither produced significant results on its own.

As in the first experiment, simple t-tests were used to compare partici-

pants' hit rates against the chance hit rate of 50 percent, and a nonparametric binomial test was used to assess the proportion of hits across all sessions. (A third statistic was also calculated; it is said to correct for unequal frequencies of left/right target positions within each session.) In this instance, we are not told how many other t-tests were carried out; if there were other tests, as is likely, this again would have required a correction for multiple comparisons. Of course, because all the data were pooled, we have no information about how many participants actually scored at a level significantly above chance. It seems odd that this information was not of interest.

Bem reports a significant correlation between the score on his two-item stimulus-seeking test and psi performance, but once again the effect was nonsignificant for "low stimulus seekers." (Could it be that Bem has serendipitously invented a two-item scale that predicts psi ability?)

#### EXPERIMENT 3: Retroactive Priming I

This experiment involved a "priming" paradigm borrowed from contemporary psychological research: participants indicate as quickly as possible whether a picture is pleasant or unpleasant, and their response time is measured. Just prior to the presentation of the picture, a positive or negative word (a "prime") is presented briefly ("subliminally") on the screen. This prime has been shown to have an effect in that participants usually respond more quickly when a positive picture is preceded by a positive word, or a negative picture is preceded by a negative word, than when picture and word are incongruent. Bem refers to this as a *contrast effect*.

Bem has taken this procedure and changed it so that the prime is presented after the participant has responded. He reports a significant contrast effect. His data analyses are very complex, involving two transformations as well as outlier[9] cutoff criteria; without access to the actual data, it is difficult to evaluate the adequacy of the analysis. However, it is obvious once again that multiple comparisons were carried out without any control for multiple testing.

#### EXPERIMENT 4: Retroactive Priming II

Experiment 4 is described as a replication of Experiment 3 "with one major change and two timing changes." Similar positive results were reported. Again, one would need access to all the data, including the discarded outliers, before one could properly evaluate the stated conclusions.

#### EXPERIMENT 5: Retroactive Habituation I

After the presentation of the previous four experiments, we are now informed that Experiment 5 was a pilot for the basic procedures used in the other experiments in this article. Why it is presented as the fifth experiment is not explained.

The experiment employed a mere-exposure protocol[10] borrowed from experimental psychology, but Bem "runs it backwards." The participant is presented with two pictures side by side. One is the "habituation target" and the other is "closely matched" to the habituation target. The participant is then instructed to indicate which picture he or she likes better. Only then is the participant repeatedly exposed, subliminally, to a picture of the "habituation target." If it turns out that the habituation target is the one that was earlier chosen, this is considered a "hit;" it is assumed that the effect of the repeated subliminal exposure to the target *after* the participant had made a choice operated backward in time to influence that original choice.

The habituation target was chosen 53.1 percent of the time, which is reported to be significant at the .014 level one-tailed. However, once again multiple t-tests (six) are reported, which means that the actual p values need to be adjusted. (Suppose that Bem had begun with .014 as the criterion value; then the actual Type I error would be $12(12.014)^6 = .08$, which is not significant).

Incidentally, Bem reports that the hit rate was significantly above chance for women but not for men. Nonetheless, he also states that there was *not* a significant sex difference! Though this seeming contradiction can arise statistically, it is up to the researcher to make sense of it—which Bem does not.

#### EXPERIMENT 6: Retroactive Habituation II

Experiment 6 is described as a replication and extension of Experiment 5. Trials with erotic picture pairs were added, and it was hypothesized that the outcome for erotic pictures would be the opposite of that for the negative pictures and that the participants would prefer the habituation picture in fewer than 50 percent of the trials. Bem does not explain his reasoning. There was also another change: on the basis that men may have simply been less aroused than women by the erotic pictures in Experiment 5, thus leading them to not produce a significant effect, it was decided to use stronger and more explicit negative and erotic images obtained from Internet sites for male subjects. Men were also given the choice of male-male erotic images and women the choice of female-female erotic images. (The reader will recall that this was also done in Experiment 1, which was run after Experiment 5.) Such matters should be investigated in further pilot studies rather than incorporated into what is billed as a replication experiment.

Bem also tells us that he had not yet introduced by this point his two-item stimulus-seeking scale into his series of experiments (remember, Experiments 5 and 6 were at the beginning of this series of nine). Instead, he constructed another *ad hoc* scale by converting two items from Zuckerman's (1974) well-known Sensation-Seeking Scale into true-false statements: "I enjoy watching many erotic scenes in movies" and "I prefer to date people who are physically exciting rather than people who share my values." He gives no reason for choosing only these statements, but he does not hesitate to treat them as a reliable and valid measure. While showing no concern for the psychometric properties of these two statements, he then arbitrarily defines only those who endorse both statements as

"erotic stimulus seekers." Thus, an individual who enjoys "many erotic scenes in movies" but prefers to date people who share his/her values was not considered to be an erotic stimulus seeker. This is purely an *ad hoc* and unacceptable procedure, again suggesting a cavalier attitude about the rigors of proper experimentation.

As for the data analysis, once again there were numerous t-tests without any control for multiple testing, thereby rendering erroneous the claimed significance levels.

**Making choices about data analysis after the data are collected introduces unacceptable opportunity for bias and allows selection of a method that suits one's hypothesis.**

#### EXPERIMENT 7: Retroactive Induction of Boredom

The hit rate was not reported to be significant in this experiment. The reader is therefore spared my deliberations.

#### EXPERIMENT 8: Retroactive Facilitation of Recall I

This experiment was an attempt to test the hypothesis that the future rehearsal of a set of words can make them easier to recall in the present. (Students would be delighted if this effect could be verified and harnessed, for they could then do further study following a difficult exam and presumably improve their performance on the examination already taken). The design was simple. Participants were shown a set of words and then were tested for their recall of the words. Subsequently, they were given practice exercises with a randomly selected subset of those words, and the hypothesis was that as a result of this subsequent practice, their performance on the test (in the past, remember) would be enhanced and they would have (in the past) recalled more of the words that were practiced in the present.

The participants were one hundred undergraduates. Again, they first responded to the two stimulus-seeking statements. Next, forty-eight common nouns were presented serially for three seconds each. The participant was then asked to type out all the words he or she could recall. The computer then selected twenty-four words at random, and the participant was now instructed to type each of the selected words. It was hypothesized that these practiced words would turn out to be the ones that had been better recalled (before the practice).

Each recalled word was deemed to be a trial and was scored as either a practice or a control word. The actual difference—recall of practice words minus recall of control words—was not analyzed; only a weighted score was given, which was that difference multiplied by the participant's overall score (on both practice and control words). We are told that this was done to give more weight to the scores of those participants who recalled more words. No appropriate justification is given for this awkward analysis; an analogy is drawn with the practice of weighting studies by their sample size in a meta-analysis, but this is a spurious analogy. The apparently arbitrary weighting of scores, when the more direct-difference scores would offer less ambiguity, renders these findings extremely difficult to interpret. Making choices about data analysis after the data are collected introduces unacceptable opportunity for bias and allows selection of a method that suits one's hypothesis.

Making matters more complicated, Bem then informs us that another twenty-five "control" sessions were run, similar to the sessions outlined above but without any practice sessions. These control sessions were interspersed among the experimental sessions. The overall recall of words in his control sessions was no different than that in the experimental sessions, and so he concluded that "the enhanced recall of practice words came at the expense of diminished recall of control words."

Again, it was found that participants who scored low in terms of his stimulus-seeking scale scored at the chance level in the recall test, while those high in stimulus seeking scored above chance.

#### EXPERIMENT 9: Retroactive Facilitation of Recall II

This is described as a replication of Experiment 8, with one procedural change: a new practice exercise was introduced "immediately following the recall test in an attempt to further enhance the recall of the practice words." Again, weighted scores were calculated, and on this basis a significant result was obtained. However, on this "replication," the stimulus-seeking questions did *not* correlate with psi success. My concerns about the data analysis in Experiment 8 similarly apply in this case.

Overall, then, this is a very unsatisfactory set of experiments that does not provide us with reason to believe that Bem has demonstrated the operation of psi. All that he has produced are claims of some significant departures from chance, and these claims are flimsy given the many methodological and analytical problems that I have touched on in this review. Moreover, Ray Hyman has noted (in my personal

**We have toured the laboratory; we have found the dirty test tubes and the mislabeled vials; we have observed inappropriate methodology and analysis. We have lost confidence in the chemist, and there seems little need to poke about further.**

communication with him) that the correlation of effect size (as well as significance level) with sample size is highly significant across this set of Bem's experiments, but it is in the wrong direction! "Effect size," simply put, refers in this case to the magnitude of the difference between the observed scoring rate and the chance rate. Larger samples provide a better opportunity to detect such a difference if it is truly there, and thus effect size should increase with increased sample size. However, in Bem's experiments, the effect size correlates *negatively* (–.91) with sample size, indicating that the claimed effect is smaller when the sample size is larger.

Statistical power is a related concept that refers to the ability to detect an effect when it is actually there. Hyman notes that while power (he uses the log of significance probability as a proxy for power) should be positively correlated with sample size (technically with the square root of sample size), in this series of studies the correlation is approximately .80—*in the wrong direction* once again. This raises a bright red flag and further erodes confidence with regard to the conduct of this research.

* * *

Having presented his nine experiments, Bem then discusses a number of general issues in parapsychology research and then turns to quantum mechanics! Even if one were to take his interpretation of his results at face value, the claimed results are small and hardly justify an incursion into quantum mechanical theory in the pursuit of accommodation of psi phenomena within modern scientific theory.

While it may seem puzzling that this distinguished psychologist has produced such flawed research, anyone who has read his "Writing the Empirical Journal Article" (published on his website at http://dbem.ws/WritingArticle.pdf) would not be surprised. There he provides advice to students regarding the conduct of research. A few revealing selections (emphasis added):

> *Once upon a time*, psychologists observed behaviour directly, often for sustained periods of time. No longer. Now, the higher the investigator goes up the tenure ladder, the more remote he or she typically becomes from the grounding observations of our science. If you are already a successful research psychologist, then you probably haven't seen a participant for some time. Your graduate assistant assigns the running of a study to a bright young undergraduate who writes the computer program that collects the data automatically. And like the modern dentist, *the modern psychologist rarely sees the data until they have been cleaned by human or computer hygienists*.
>
> To compensate for this remoteness from our participants, let us at least become intimately familiar with the record of their behaviour: the data. Examine them from every angle. Analyze the sexes separately. Make up new composite indexes. If a datum suggests a new hypothesis, try to find additional evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you don't like, or trials, observers, or interviewers who gave you anomalous results, drop them (temporarily). Go on a fishing expedition for something—anything—interesting. . . .
>
> When you are through exploring, you may conclude that the data are not strong enough to justify your insights formally, but at least you are now ready to design the 'right' study. . . . Alternatively, *the data may be strong enough to justify re-centering your article around the new findings and subordinating or even ignoring your original hypotheses*. . . .
>
> Your overriding purpose is to tell the world what you have learned from your study. If your research results suggest a compelling framework for their presentation, adopt it and make the most instructive findings your centerpiece. Think of your data set as a jewel. *Your task is to cut and polish it, to select the facets to highlight, and to craft the best setting for it. Many experienced authors write the results section first.*
>
> But before writing anything, Analyze Your Data!

Reflections of this advice appear to be writ large throughout Bem's research article.

* * *

The publication of this set of experiments will serve no one well. Parapsychology is not honored by having this paper accepted by a mainstream psychology journal. Neither does Bem's paper serve the public well, for it only adds to confusion about the scientific case for the existence of psi. And it does no service to the reputation of the *Journal of Personality and Social Psychology*. Although Bem has failed to demonstrate the existence of mysterious intellectual powers that defy the normal constraints of time and space, there seem nonetheless to have been mysterious intellectual powers at play here. I refer to

the decision by the editors of an esteemed psychology journal to publish this badly flawed research article.

"Think of your data set as a jewel," Bem instructs. However, with these nine experiments, Bem did not end up with a polished jewel. Rather, to extend his metaphor, the jewel cracked under the intense pressure used to try to shape it to fit expectation. One is left with nothing but useless fragments that reflect not the light of knowledge but the biases of the researcher.

Rhine, Schmidt, Targ, Puthoff … the list grows on. *Plus ça change, plus c'est la même chose*. ■


## Acknowledgements

Thanks to Ray Hyman, Scott O. Lilienfeld, Timothy Moore, and Benjamin Wolozin for their very sage comments on an earlier draft of this article.


## Notes

1. My discussion is based on the pre-publication version of Professor Bem's article that appears on his website at www.dbem.ws/FeelingFuture.pdf.

2. The *p value* is the likelihood of having concluded that there is a significant effect when in fact there is not. The lower the p value, the less likely it is that the null hypothesis (that there is no effect) is true.

3. *Meta-analysis* is a statistical process for testing the combined results of a number of studies that were based on similar research hypotheses.

4. The *t-test* is a statistical test typically used either to compare two means or to compare a mean with a theoretical expectation—for example, to assess the difference between an observed average success rate and a hypothetical chance rate of 50 percent.

5. A *Type I error* occurs when the null hypothesis (that there is no effect) is rejected when it is in fact true.

6. In a *two-tailed test,* one assesses the data to see whether they significantly differ from what would be expected by chance in either direction, that is, whether they are greater than or less than what would be expected by chance alone. When we say that something is significant at the .01 level two-tailed, this means that we would expect these results to occur by chance alone only 1 percent of the time. But, given that either above-chance or below-chance results are considered to be meaningful, this 1 percent must be distributed in both directions. Thus, above-chance results would be significant at the .01 level (two-tailed) only if they are so extreme that they would be expected to occur by chance half of 1 percent, or 0.5 percent, of the time or less. The same would apply for below-chance results.

With a *one-tailed test*, one also assesses the data to see whether they significantly differ from what would be expected by chance but in only one direction, that is, whether they are either greater than expected by chance or less than expected by chance, but not both. A one-tailed test is properly used if one has good reason to predict the direction of the data in advance. Again, using the example of the .01 level, for a one-tailed test the data only need be extreme enough that they would be expected by chance alone 1 percent of the time or less (compared to 0.5 percent with a two-tailed test). This makes it much easier to claim statistical significance.

Parapsychologists normally employ two-tailed tests because results that are either significantly above chance or significantly below chance are taken to reflect psi. Although Bem indicates that he predicted that the erotic pictures would lead to above-chance scoring, which could justify using a one-tailed test, what would he have done had the participants scored at a below-chance rate that would have been significant had he predicted that the results would indeed be below chance? Apparently committed to a one-tailed test and having made only the above-chance prediction, he properly would have had to ignore those data—something that parapsychologists do not want to do. By using two-tailed tests, parapsychologists avoid the problem and also avoid any suspicion of having changed the direction of their prediction after having examined the data.

7. The *Bonferroni t-test* is a modified t-test that adjusts for the number of tests being carried out so that the overall likelihood that one of them produces significance by chance alone is kept at a specified level, such as 5 percent.

8. A *nonparametric binomial test* deals with data divided into two categories and examines the statistical significance of deviations from a theoretically expected distribution. It is referred to as "nonparametric" because it does not rely on the parameters of a distribution, such as the mean.

9. An *outlier* is a datum that is numerically distant from all the other data in the sample, either as a result of measurement error or because the data are not distributed in the manner that was assumed.

10. *Mere-exposure protocol* is a research approach in which participants' responses are assessed with the assumption that having simply been exposed (perhaps subliminally) to a stimulus object will cause an effect.

## References


Alcock, J.E. 1988. A comprehensive review of major empirical studies in parapsychology involving random event generators and remote viewing. In Commission on Behavioral and Social Sciences and Education, *Enhancing Human Performance: Issues, Theories and Techniques, Background Papers.* Washington, D.C.: National Academy Press, 601–719. Available online at http://books.nap.edu/openbook.php?record_id=778&page=601.

Bem, D.J., and C. Honorton. 1994. Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin* 115: 4–18.

Bem, D.J., J. Palmer, and R. Broughton. 2001. Updating the ganzfeld database: A victim of its own success. *Journal of Parapsychology* 65: 1–6.

Gardner, M. 1977. ESP at random. *New York Review of Books*, July 14.

Hyman, R. 1985. The ganzfeld psi experiments: A critical appraisal. *Journal of Parapsychology* 49: 3–49.

———. 2010. Meta-analysis that conceals more than it reveals: Comment on Storm et al. [2010a]. *Psychological Bulletin* 136(4): 486–90.

Hyman, R., and C. Honorton. 1986. A joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology* 50: 351–64.

Jahn, R., B. Dunne, G. Bradish, Y. Dobyns, A. Lettieri, R. Nelson, J. Mischo, E. Boller, H. Bosch, D. Vaitl, J. Houtkooper, and B. Walter. 2000. Mind/machine interaction consortium: PortREG replication experiments. *Journal of Scientific Exploration* 14: 499–555.

Jeffers, S. 2003. Physics and claims for anomalous effects related to consciousness. In J.E. Alcock, J. Burns, and A. Freeman (Eds.), *Psi Wars: Getting to Grips with the Paranormal.* Exeter, UK: Imprint Academic, 135–52.

Marks, D., and R. Kamman. 1978. Information transmission in remote viewing experiments. *Nature* 274: 680–81.

———. 1980. *The Psychology of the Psychic.* Buffalo, NY: Prometheus Books.

Milton, J., and R. Wiseman. 1999. Does psi exist? Lack of replication of an anomalous process of information transfer. *Psychological Bulletin* 125: 387–91.

PEAR (Princeton Engineering Anomalies Research). n.d. Experimental research. Available online at www.princeton.edu/~pear/experiments.html.

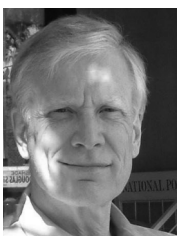
Rhine, J.B., and W. McDougal. 1934/2003. *Extra-Sensory Perception*. Whitefish, MT: Kessinger Publishing.

Storm, L., P.E. Tressoldi, and L. Di Risio. 2010a. Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin* 136: 471–85.

———. 2010b. A meta-analysis with nothing to hide: Reply to Hyman. *Psychological Bulletin* 136: 491–94.

Targ, R., and H. Puthoff. 1974. Information transmission under conditions of sensory shielding. *Nature* 251: 602–4.

Zuckerman, M. 1974. The sensation seeking motive. In B.A. Maher (Ed.), *Progress in Experimental Personality Research* (Vol. 7). New York, NY: Academic Press, 79–148.


**James E. Alcock** is professor of psychology at York University, Toronto, Canada. He is author of *Parapsychology: Science or Magic?* and co-editor of *Psi Wars: Getting to Grips with the Paranormal.* He is a member of the Committee for Skeptical Inquiry's executive council and of the SKEPTICAL INQUIRER editorial board. He may be reached via e-mail at jalcock@glendon.yorku.ca.


A response to this article from Daryl Bem plus author Alcock's detailed reply to Bem's response are on our website at

**www.csicop.org**