http://dl.dropbox.com/u/8290411/ResponseToAlcock.pdf

Response to Alcock's "Back from the Future: Comments on Bem

http://www.csicop.org/specialarticles/show/response_to_alcocks_back_from_the_future_comments_on_bem

Daryl J. Bem

Cornell University

On December 3, 2010, James Alcock published an essay on this site (http://www.csicop.org/specialarticles/show/back_from_the_future/) critiquing my article "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect," which has been accepted for publication in the *Journal of Personality and Social Psychology* (*JPSP)*. (A prepublication copy is available at http://dl.dropbox.com/u/8290411/FeelingFuture.pdf). Even though the article will not appear in print for several months, it began to receive widespread media coverage after the experiments were described by a blogger for *Psychology Today*.

Alcock begins his critique by noting that

> "What has made this report so particularly newsworthy is both the academic stature of the author, a respected Professor of Psychology at Cornell University, and the fact that it is published in the American Psychological Association's (APA) *Journal of Personality and Social Psychology,* the world's preeminent social psychology journal."

I believe that Alcock has also put his finger on what is so particularly newsworthy about his critique: the striking contrast between his harsh assessment of my work and the collective assessment of the two editors and four reviewers who vetted it for the Journal. *JPSP* in one of the most rigorously refereed journals in the entire field of psychology, with a rejection rate of 82% in 2009. Moreover, authors' names and other identifying information are removed from a manuscript before it is sent to reviewers so that their

evaluations will be based purely on its merits and not be influenced by knowledge of an author's reputation or the prestige of his or her institutional affiliation.

The contrast between the assessments of Alcock and the Journal's editors and reviewers is also particularly newsworthy because it is not simply a reprise of the familiar disagreement between skeptics and proponents of psi (ESP). Like Alcock, several of the reviewers expressed various degrees of skepticism about the reality of psi, while still urging the article's acceptance. Unlike Alcock, however, they are all active researchers who regularly contribute to the mainstream experimental literature in psychology and cognitive science. Their task was to evaluate the logic and clarity of the article's exposition, the soundness of its experimental methods, and the validity of its statistical analyses. They did not have to agree with my conclusions regarding psi to make those assessments. As Joachim Krueger, an experimental psychologist at Brown University, put it so charmingly: "My personal view is that this is ridiculous and can't be true. Going after the methodology and the experimental design is the first line of attack. But frankly, I didn't see anything. Everything seemed to be in good order (quoted by Peter Aldhous in the *NewScientist* 16:29, November 11, 2010)."

The Research

My article reports nine experiments, involving more than 1,000 participants, that test for precognition or retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before rather than after the stimulus events occur. Each time-reversed experiment tests the straightforward hypothesis that we should observe the same effect that we normally observe in the standard (non-psi) version of the experiment. Five different effects are tested in this way; and, to bolster confidence in the results, four of the nine experiments are actually replications of

the other experiments in the article. Across all nine experiments, the combined odds against the findings being due to chance are greater than 70 billion to 1.

The Critique

Alcock challenges both my experimental procedures and my statistical analyses. His article is quite lengthy, and so I will here focus only on his two most frequently recurring criticisms, one concerning experimental procedures and one concerning statistical analyses. (I will not here address Alcock's lengthy preamble in which he imaginatively rewrites the history of psi research. That has already been done by Dean Radin on his Internet blog at http://deanradin.blogspot.com/.)

Alcock's major procedural criticism concerns my selection and deployment of the pictorial stimuli used in six of the nine experiments. As explained in the article, they are drawn primarily from the widely used International Affective Picture System (IAPS), a set of 820 digitized photographs that have been rated by both male and female raters on numerical scales for their emotional tone (extremely negative to extremely positive) and their arousal level (non-arousing to highly arousing).

Male and female raters differed markedly in their ratings of negative and erotic pictures. Male raters rated every one of the negative pictures as less negative and less arousing than did the female raters, and they gave more positive ratings than the female raters to the most explicit erotic pictures. Possibly reflecting this sex difference, female participants showed significant psi effects with negative and erotic stimuli in my earliest experiment but male participants did not. Accordingly, I decided to introduce different sets of pictures for men and women in subsequent experiments, choosing more extreme and more arousing pictures for the men. As a result, no sex differences in psi performance appeared in any of the later experiments. In addition, the computer program gave

participants the choice of being shown either opposite-sex or same-sex erotic pictures, without having their choice divulged to the experimenter.

Although the *JPSP* reviewers had no problems with any of this, Alcock clearly does: "Now we find that participants were allowed to choose their target set! This is the most baffling description of research materials and procedures that I have ever encountered." I am surprised by Alcock's reaction here. Because he had post-doctoral training in clinical psychology and has served as a member of the Council for Scientific Clinical Psychology and Psychiatry, I would have expected him to be familiar with several well-known clinically-oriented experiments on reactions to threat in which different sets of threatening stimuli were assembled for groups of participants with different psychiatric diagnoses (e.g., homosexually-toned materials for male patients diagnosed with paranoia). Some of those experiments even constructed tailor-made sets of stimuli for each individual participant. This is all kosher. The conceptual hypotheses in those experiments concerned the ways in which participants responded to stimuli threatening to *them*. Similarly, the hypotheses in my experiments concern the ways in which participants respond to stimuli that are erotically arousing for *them*.

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. This example also illustrates a more general problem with Alcock's critique: A failure to distinguish between potential flaws in an experiment that would illegitimately produce false positive results and potential flaws that would actually work against the experimental hypothesis by introducing noise into the data. The first kind of flaw is fatal and constitutes grounds for rejecting the probative value of an experiment. The second kind of flaw simply produces

weak or nonsignificant results and is a setback only for the experimenter. For example, if using different sets of stimuli for men and women were a flawed procedure, it would have been a flaw of this second kind and would have militated against positive results.

The Precognitive Detection of Erotic Stimuli

Alcock made his comment about being baffled by the description of research materials in his discussion of the first experiment reported in the article, and I presume he meant it to apply to other features of that experiment as well. The experiment was designed to test the hypothesis that individuals can precognitively detect the future location of an erotic picture.

There were 100 sessions in this experiment, and on each of 36 trials, participants saw images of two curtains side-by-side on the computer screen. They were told that a picture would be behind one of the curtains and a blank wall would be behind the other. Their task on each trial was to click on the curtain they felt concealed the picture. After they made their selection, the selected curtain opened, revealing either a picture or a blank wall. Unknown to participants at the time, the computer did not actually select the picture to be shown or determine its left/right position until after they had already made their decision. This procedure thus tested a participant's ability to anticipate a future event, a test of precognition.

On randomly selected trials, the picture was erotic; on other trials, it was nonerotic, and the participant had no (non-psi) way of knowing which kind of picture would be used on any given trial. Because there were two alternatives on each trial—left curtain or right curtain—the probability that the participant would correctly select the location of the picture by chance was always 50%. (Alcock apparently misunderstood the procedure, concluding somehow that trials with nonerotic pictures had a 33% chance probability of success. When he says he is baffled, I believe him. Fortunately, the reviewers for *JPSP*

were not baffled.) Accordingly, the experimental hypothesis being tested in this experiment was that on trials using erotic pictures, participants would select the correct curtain on significantly more than 50% of the trials.

The hypothesis was supported: Participants successfully detected the future location of the erotic pictures on 53.1% of the trials. This result was evaluated for statistical significance by a *t* test, which evaluated the probability that a 53.1% success rate across 100 sessions could have arisen by chance. By convention, psychologists are permitted to call a result "statistically significant" if it could have arisen by chance less than 5% of the time. This particular result could have occurred by chance less than 1% of the time.

I further analyzed the data to see if participants could also detect the future locations of nonerotic pictures. It might well be that there is nothing unique about erotic pictures beyond their high arousal value and positive emotional tone. Using the IAPS numerical ratings, I defined four kinds of nonerotic pictures: emotionally negative pictures, emotionally neutral pictures, emotionally positive pictures, and romantic-but-nonerotic pictures (e.g., a kiss between a bride and groom at their wedding).

To accommodate so many different kinds of nonerotic pictures, I divided the 100 sessions into two parts. Forty sessions included trials with negative and neutral pictures and sixty sessions included trials with positive and romantic pictures. By design, this yielded 600 positive trials and 480 each of negative, neutral, and romantic trials—enough of each to permit separate statistical testing. A *t* test across all sessions revealed that participants did no better than chance on nonerotic pictures, and separate *t* tests further revealed that they did no better than chance on any of the subsets of nonerotic pictures.

The Problem of Multiple Statistical Tests

This brings us to the statistical criticism that Alcock raises repeatedly throughout his critique. As he correctly notes, 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.

For example, as noted above, multiple *t* tests demonstrated that participants did no better than chance on any of the subcategories of nonerotic pictures. It is here that Alcock first complains about my performing multiple tests without adjusting the significance level for the number of tests performed.

In this case, Alcock is almost right. Suppose that in testing each of the four subcategories of nonerotic pictures, I had found that one of them (e.g., romantic pictures) showed a significant precognitive effect. Because this finding would have emerged post hoc, only after I had first performed separate tests on four different picture types, I would have had to adjust the significance level to be less significant. If I did not, I would be illegitimately capitalizing on the likelihood that at least one of the four tests would have yielded a positive result just by chance. But there was no psi effect on any of the subcategories of nonerotic pictures. Perhaps Alcock wants me to change my conclusion that there were no significant effects on nonerotic pictures to the conclusion that there were *really really* no significant effects on nonerotic pictures.

In choosing to test the main hypothesis about erotic pictures in this experiment with a *t* test, I was aware that particularly cautious or skeptical readers might worry about the

mathematical assumptions that underlie this common statistical test. So, I demonstrated that the same result would be obtained by using an alternative test—called a "nonparametric" test—that did not rest on those assumptions. I did this throughout the article, showing in every experiment that the same conclusions are reached no matter which kind of test is used. These multiple tests were thus aimed at showing that the same conclusion arises from different statistical treatments of the same data. This is very different from conducting several exploratory tests on different portions of the data and then concluding post hoc that one of them showed a significant effect. We shall now see further evidence that Alcock seems not to understand the difference.

The Retroactive Priming Experiments

Two of my experiments tested retroactive priming, a time-reversed version of a popular procedure in contemporary cognitive and social psychology. In a typical (non-psi) priming experiment, participants are asked to judge as quickly as they can whether a picture is pleasant or unpleasant, and their reaction time is measured. Just before the picture appears, a pleasant or unpleasant word (e.g., *beautiful*, *ugly*) is flashed briefly on the screen; this word is called the prime. Individuals typically respond more quickly when the prime and the picture are both pleasant or both unpleasant than when one is pleasant and the other is unpleasant. In my time-reversed version of the procedure, the prime did not appear until after participants made their judgments of the pictures.

Alcock's objections to these experiments are that my

> …data analyses are very complex, involving two transformations as well as outlier cut-off criteria, and 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.

With regard to the complexity of the data analysis, it is true that reaction time data require specialized treatment, and I adopted the analytic procedures that are now considered standard for priming studies. The associate editor and two of the reviewers of my article are experts in priming studies and major contributors to the priming literature. Had I not performed the standard analyses of the data, the reviewers would have required me to do so before they accepted the article. At least one expert in priming experiments has also argued that one should always perform several analyses using different transformations and different cut-off criteria to ensure that the priming effects hold up across these variations. That is precisely what I did. Unlike Alcock, the reviewers understand both the statistical treatment of priming data and why the multiple tests strengthen the conclusions drawn.

Multiple Tests—One More Time

In two of my experiments, I was concerned about potential bias or nonrandomness in the computer's successive left/right placements of the target pictures, so I presented four different data analyses, each one controlling in a different way for possible bias in the randomization process. Again, Alcock robotically invokes his mantra about multiple tests, failing to realize that the whole point of multiple tests in these experiments was to demonstrate in several converging ways that my conclusions were not compromised by bias in the random placement of target pictures.

Ironically, one purpose in reporting multiple tests throughout the article was to counter a charge often made by skeptics who are tempted to explain away psi data on the grounds of experimenter dishonesty: This is the charge that an experimenter might have tried out several statistical tests and then cherry-picked among them to report only the one that worked. Alas, when dealing with Alcock, no good deed goes unpunished.