

Skepticism and Negative Results in Borderline Areas of Science

J.E. Kennedy

(Unpublished Manuscript, 1981)

When researchers who are skeptical of the validity of a hypothesis fail to replicate the significant results obtained by those more favorable to the hypothesis, the skeptics often explicitly or implicitly interpret the positive results as being due to some type of experimental error.

The purpose of this paper is to address the other side of the coin, the possibility that, at least sometimes, biased errors by the skeptics play a decisive role in producing their negative results and conclusions. To this end, some cases in which skeptics either carried out research or evaluated the work of others are examined for errors, and then some implications of these cases are discussed. The presentation here is not intended to be a state of the art summary of the research areas of these cases, but rather an examination of the strategy and methodology used in the examples. Before examining the cases, some background matters need to be dealt with.

Most people who consider themselves "scientific" sincerely believe that their judgments are based on objective evaluations of the evidence rather than on personal biases. This controversial (perhaps absurd, in light of recent work in the history and sociology of science -- see e.g., Barber, 1961; Brush, 1974; Kuhn, 1963) view of their underlying motivations will not be specifically challenged here.

For the purpose of this discussion, the term skeptic is used to refer to those who have, for whatever reason, a strong expectation that a particular hypothesis will not be verified when objectively investigated. Those who are irrationally hostile to a phenomenon are, of course, also included within the domain of the term.

This paper is primarily concerned with "borderline" areas of science. The term "borderline" is used here in a very general sense, referring to any research on topics not firmly accepted by the scientific community. My primary experience has been in the field of parapsychology and most of the examples will be drawn from that field. However, because these considerations apply to a much broader range of topics and since a recent investigation of astrology (which I consider not to be in the realm of parapsychology) is particularly apropos, the term "borderline" was selected. The reader can include under this term whatever topics he or she desires with little effect on the points made here.

Because I am critically reviewing criticisms by others, it may be useful to discuss some aspects of scientific criticisms per se. Perhaps the most extreme form of criticism is a blanket accusation of fraud for any result that cannot readily be dismissed on methodological grounds. Some critics of research into borderline areas of science do not require that actual evidence for fraud be found in order to disregard a study; rather, they dismiss any study for which a

mechanism of fraud can be hypothesized. However, with some ingenuity, speculations of fraud can be raised for any study by proposing schemes involving some unreported or unverifiable details of the procedure.

The pastime of dreaming up possible mechanisms of fraud has been called the "Hume Game" by Palmer (1978) since, following the logic of the philosopher David Hume, some critics of parapsychology have rejected all evidence for psi on the grounds that it could have been fraudulently obtained. When this argument is taken to the extreme of proposing that even successful replications were also fraudulently produced, the skeptic is in a completely unassailable position. Such a skeptic is also, of course, out of the arena of science and into the realm of fanatical dogma.

While this extreme position is clearly unacceptable in science, the exact number and rate of successful replications that can be considered as providing compelling evidence for a finding is a controversial and subjective matter. Also, although some precautions against intentional errors should be incorporated into experimental designs, the amount of effort that should be taken is very debatable.

A less extreme variation of the Hume Game is to dream up experimental modifications that might have improved the precautions against unintentional or intentional errors. Here, too, no matter how carefully an experiment is designed, other features that could have been employed can always be imagined. The game the skeptic plays is to present a few of these potential modifications and then conclude that the experiment was incompetently designed because these features were not incorporated. Not surprisingly, those who follow this course feel that none of the successful experiments have been adequately designed. For parapsychology, C.E.M. Hansel (1966, 1980) is the most well-known Hume Game player.

Those who have not yet developed an appreciation for the unlimited power of these methods for rejecting experimental results may want to carry out a couple of enlightening exercises. First, create a fictitious experimental report describing conditions and results that would provide satisfactory evidence for a hypothesis (e.g., evidence for ESP). Then subject the report to the level of criticism found in Hansel's writing. The impossibility of carrying out and reporting acceptable experiments quickly becomes obvious. Another interesting exercise is to subject studies with chance results to the same level of criticism, thus casting doubts upon the validity of these studies. It is surprising how easily (trivially) any study can be dismissed once one gets the hang of these types of criticisms. As might be expected, such criticisms are usually only raised for experiments with results which contradict the critic's biases.

Again, while these extreme forms of criticism are unacceptable, the exact demarcation between legitimate versus unreasonable concerns about methodology is controversial.¹ This difficulty of demarcation applies not only to criticisms by skeptics, but also to criticisms of their work. I believe, however, that the issues discussed in the examples below are sufficiently straightforward that such demarcation problems do not arise.

"Fads and Fallacies"

That some individuals are strongly biased against borderline areas is well known and sometimes openly acknowledged. For example, when discussing the work of Dr. J. B. Rhine, Martin Gardner (1957) commented, in his book *Fads and Fallacies in the Name of Science*:

There is obviously an enormous, irrational prejudice on the part of most American psychologists -- much greater than in England, for example -- against even the possibility of extrasensory powers. It is a prejudice which I myself, to a certain degree, share. Just as Rhine's own strong beliefs must be taken into account when you read his highly persuasive books, so also must my own prejudice be taken into account when you read what follows (pp. 299-300).

Numerous examples of the type of erroneous reporting that Gardner was presumably warning about can be found in his discussion of parapsychology. One instance concerns the "negative results" of Coover (1975). Gardner stated:

Professor John E. Coover, of Stanford University, made extensive and carefully controlled ESP tests which were published in detail in 1917, in a 600-page work, *Experiments in Psychical Research*. Recently Rhine and others have gone over Coover's tables, looking for forward and negative displacement, etc. They insist that ESP is concealed in his figures... You can always find patterns in tables of chance figures if you look deep enough (p. 308).

Gardner's description is erroneous on at least three accounts:

(1) Coover's experiments were not carefully controlled. When confronted with the later-discovered significant statistics, Coover (1939) took this position himself and it was also noted by Rhine, Pratt, Stuart, Smith, and Greenwood (1940, P. 147) in their survey of ESP experiments. One of the problems with Coover's experiments was the possibility of recording errors -- the same reason that Gardner used to dismiss much of Rhine's work. In Gardner's view, Rhine's early work (which Rhine interpreted as evidence favorable to ESP) was carried out under unacceptable conditions while Coover's work (which Coover interpreted as unfavorable to ESP) was "carefully controlled." In fact, the conditions of Coover's experiments were as bad or worse than many of the early experiments by Rhine. (Actually, most of the experiments carried out by anyone prior to the middle 1930's were done under conditions that are loose by today's standards.) It would appear that Gardner's rating of "carefully controlled" is related to the agreement of the conclusions with his biases rather than to the actual conditions of the experiment.²

(2) The idea that Coover's results could not be attributed to chance (which is different from concluding that the results were due to ESP) did not come from a scrounging of the data. Coover compared two conditions in his experiment, a telepathy condition in which an agent knew the target card, and a "control" condition in which no one knew the target. One of Rhine's first areas of interest in parapsychology was the investigation of clairvoyance (ESP without an agent knowing the target). In light of this work, Coover's second condition was actually a clairvoyance condition rather than a control. Rhine's early work indicated that ESP could operate equally well in both telepathy and clairvoyance situations and thus, Coover's comparison might not be

expected to give significant results. When the now-obvious step of pooling the data for the two conditions in Coover's experiment was taken, the overall result was statistically significant ($p < .005$). The data were, in fact, in line with the ESP hypothesis, however, the methodological weaknesses prevented Coover's experiments from being of clear evidential value.

(3) Rhine certainly did not "insist" that ESP was concealed in Coover's data. Rather, he took a much more cautious tone as evidenced by his early conclusion concerning Coover's work: "While, then, Prof. Coover did not prove anything at all, perhaps he unwittingly opened up some very interesting suggestions, which might profitably have been followed up" (Rhine, 1935, p. 27).

Gardner's description misrepresents the quality of Coover's work and gives a clearly erroneous picture of Rhine's attitude towards that work. While Gardner's misrepresentations may be somewhat excused because he did give the readers fair warning of his "irrational prejudice" and because he is a writer rather than a research scientist, the practices of certain other individuals are less pardonable.

The Wheeler Incident

Perhaps the most inexcusable error by a skeptic in recent years was made by physicist John Wheeler at the annual meeting of the AAAS on 8 January 1979 in Houston. Some physicists have suggested that the solution of the very perplexing problems related to the concept of "observation" in physics may require an explicit role for "consciousness." In a panel session on "Physics and Consciousness," Wheeler presented a paper which argued against these ideas and called for the study of brain functioning and consciousness to be kept separate from questions about the concept of observation in physics. He did not indicate that he had solutions to the observation problems in physics; rather, his strong views were apparently based on his personal conviction that his ideas will provide answers while other approaches will only impede progress. (The dubious nature of such a position is amply documented by the history of science – for example, the resolution in 1901 of the Council of the Royal Society which requested that mathematics be kept separate from biological studies (B. Barber, 1961)).

In two appendices to his paper Wheeler called for the disaffiliation of the Parapsychological Association from the AAAS. That this was not one of the more objective, carefully reasoned, and calmly articulated presentations at a AAAS convention can readily be seen from the title of the first appendix, "Drive the Pseudos Out of the Workshop of Science." (Reprinted in the 17 May 1980 issue of the *New York Review of Books*; the word "Drive" was later incorrectly given as "Put" when Wheeler gave the reference in *Science* - see below.) Some of Wheeler's arguments included: "There's nothing that one can't research the hell out of." "Research guided by bad judgment is a black hole for good money." "Where there is meat there are flies." And concluded with: "Now is the time for everyone who believes in the rule of reason to speak up against pathological science and its purveyors."

While the credibility of Wheeler's appendices is obviously limited, he did make a very serious charge in the discussion following their presentation. (Tape recordings of the full session, including the paper, the appendices, and the discussion were distributed by the AAAS.) The appendices presented parapsychology only in terms of sensationalized, popular topics. When

asked to comment on the actual experimental work, Wheeler stated that J.B. Rhine, as an assistant to William McDougall in some animal experiments on heredity, had been exposed as intentionally producing spurious results and that Rhine "had started parapsychology that way." This comment was obviously intended to provide a basis for dismissing much of the experimental work in parapsychology.

Suffice it to say, the statements about Rhine were completely untrue, a fact verified by both Rhine and the man who was allegedly involved in his exposure. Wheeler has subsequently published in *Science* a (somewhat opaque) "correction" acknowledging the erroneous nature of his story about Rhine (Wheeler, 1979; also see comments by Rhine, immediately following Wheeler's letter).

If fraud is the most inexcusable act in science, worthy of the harshest condemnation, then attempting to discredit a person or an area of research by using fabricated accounts of data manipulation must be equally unacceptable. In the present case, it is not clear who actually invented the story; Wheeler says only that it was "second-hand." Wheeler was obviously undiscerning in his sources of information about parapsychology and his invalid statements will likely have a lasting, detrimental effect upon the proper scientific evaluation of Rhine's work.³

Experiments by a Skeptic

Two ESP experiments reported in the *Journal of Social Psychology* by Warner Wilson (1964) were intended to replicate "with modification" previous work by Schmeidler (1952) which found a difference in ESP scores between those who believed in ESP (sheep) and those who were skeptical (goats). Wilson's first experiment tested 621 subjects with each subject making ten calls. Statistical evaluations were done with chi-square tests using the subject as the unit of analysis (skeptics vs. believers by above vs. below chance). The results clearly were in line with the previous work; he sheep scored significantly above chance ($p < .005$) and significantly better than the goats ($p < .005$). This situation forced Wilson to comment:

Although the writer made an attempt to be reasonably objective in the introduction, the reader can, no doubt see, that he is skeptical rather than optimistic about the reality of ESP. The reader, therefore, can imagine the writer's consternation when he noted that the data of the first experiment lend considerable comfort to the parapsychological position (p. 382-383).

Under the circumstances, one would normally expect the researcher to verify his previous results by carrying out as nearly an exact replication as possible. However, Wilson's second experiment involved fundamental changes in design that obviously biased the experiment against significant results.

The most glaring change was reducing the sample size to 90 subjects with each subject making five calls, for a total of 450 trials (as compared to 6210 in the first study). Given the magnitude of the effect in the first experiment, this decrease in sample size almost guaranteed chance results on the second experiment. As would be expected, the results were at chance which led Wilson to comment:

The results of Experiment two offer more comfort to the skeptic and seem definitely embarrassing to the parapsychologist (p. 384).

After pooling the data for both experiments, Wilson reported that the overall result was not significant and concluded that "the results taken as a whole...offer no support to the parapsychologist" (p. 387).

However, Wilson pooled the results by using the subject as the unit of analysis in the first study and the trial as the unit of analysis in the second, thus weighting the second study over the first by a factor of approximately ten. This method of pooling results is clearly inappropriate and if a uniform analysis had been used on both sets of data, the pooled result would probably have been significant.

While Wilson was quick to condemn incompetence on the part of parapsychologists, the methodology of his own studies was far below the common standards of the time. The details of his experimental procedures were not described so it is not possible to determine whether sensory cues or cheating by the subjects could have entered into the results. (In the first study, the subjects apparently received unsealed envelopes which contained the target sequences. The amount of supervision of the testing sessions was not stated.)

The statistical techniques Wilson employed were dubious because the same target sequences were used by more than one subject, thus leading to a lack of independence between subjects. This problem was more severe in the second study since all of the subjects used the same target sequence while only three or four subjects used the same target sequences in the first experiment. Further, in the second study the trial was used as the unit of analysis; yet, many trials were discarded because the "scores" were at chance rather than above or below -- a situation that makes no sense to me since each trial should have been either a hit or a miss. Wilson's paper contains numerous other errors, ambiguities, and misrepresentations, particularly with regard to discussions of previous ESP research, but further documentation here would serve no purpose.

The Mars Effect

In many cases the misuse of statistical methods to discredit experimental results is less obvious. The evaluation by Zelen, Kurtz, and Abel (1977) of part of Michel Gauquelin's work on the "Mars effect" (an effect that fits into the general category of astrology) demonstrates the type of difficulties that can arise. Gauquelin (1967) reported that sports champions tend to be born when the planet Mars is in certain positions relative to the horizon. This conclusion was based on data for 1553 sports champions from France and 535 from Belgium, but the exact details of the analyses and findings need not concern us here.

In response to some legitimate (in my opinion) questions raised about the validity of the statistical methods Gauquelin employed (e.g., Committee Para, 1976a, 1976b), Professor Marvin Zelen of Harvard proposed a statistical method that was acceptable to all parties (Zelen, 1976). However, the "Zelen test" required collecting more birth data as controls and would have been impractical to apply to the entire sample. To overcome this difficulty, Zelen suggested that a random sample be selected from the group of 1553.⁴ Because an adequate number of control births could not be obtained for sports champions born in small towns, Gauquelin did not randomly select cases from the overall group but rather used all the cases that came from larger towns. This resulted in a sample of 303 cases.

When evaluated by the Zelen method, the sample significantly ($p < .03$) displayed the Mars effect (Gauquelin and Gauquelin, 1977), a finding which the Gauquelins, of course, accepted as clearly supporting their previous conclusions.

In their analyses and evaluation, Zelen, Kurtz, and Abel (1977) partitioned the sample by sex, geographical region, and data source. After excluding the females from the sample, the 294 male cases (yielding $p < .04$) were categorized according to three geographic regions: (1) Paris, (2) France, excluding Paris, and (3) Belgium. When the Mars effect was evaluated for each group separately, the Paris sample was significant ($N = 40$, $p = .01$), the rest of France was suggestive ($N = 189$, $p = .09$) and the Belgium sample was slightly in the opposite direction ($N = 65$, $p = .52$). Zelen et al. felt that these outcomes, combined with the non-random sampling from the larger group, raised doubts about the validity of the Mars effect. So that there will be no ambiguities in paraphrasing their position, the relevant part of their paper is quoted below:

This decision for the data collection means that large cities will have contributions not related to their population proportions in the original data set of 1,553 sports figures. Hence cities like Paris or Marseille have larger weights than they would have if a random sample of sports figures had been drawn. If the Mars effect does not recognize geographic boundaries, then confining the subsample to large localities may be satisfactory. However, if there is a geographic distinction in the demonstration of the Mars effect, then one can draw conclusions associated with large geographical regions, when in truth the significant contributions come from a few localities.

This is exactly the situation we are confronted with. The statistical evidence from the Paris data is strong. No other geographic area shows a demonstrable statistical difference [between the athletes and controls]. Hence one concludes that the experiment carried out by the Gauquelins has shown a statistical effect for Paris. However, it is not so for the entire group.

As the Gauquelins point out, if the Mars effect is a weak one, then large samples would be required to demonstrate the effect. Among the sample of 294 male champions, 42 [should be 40⁵] were from Paris. One would have expected that the data, excluding Paris, which consists of 252 [should be 254] male champions, would also have demonstrated a statistical difference. This was not so. Thus a sample six times as large as that of Paris failed to demonstrate a statistically significant difference. This failure is not likely to be attributed to the small sample size.

Finally, what are we to conclude from this experiment? If the Mars effect is real, why can it not be demonstrated over a larger geographical locality than Paris? Another possible interpretation of the Paris results is that indeed one has observed a rare event. In looking at many data sets, one will occasionally conclude the existence of a real difference when in fact none really exists. If a significance level of .05 is adopted as indicating a real difference, then 5 per cent of such statements in which a difference is declared can be in error, and in fact no difference may really exist. The interpretation of the Gauquelin results would have been conclusive if the difference could be consistently demonstrated. The

fact that it could only be demonstrated for Paris is disconcerting. (Zelen, et al. , 1977, p. 38).

First, let us consider the matter of the non-random sample from the group of 2088. (The Gauquelins actually selected from the overall group of 2088 rather than the 1553 indicated by Zelen, et al.) If it is assumed that the group of 2088 adequately represents France and Belgium, then a random sample is desirable if the conclusions are to generalize to all of France and Belgium. A sample drawn from the larger cities in these countries, then, strictly allows inferences about only the more populated areas rather than the entire countries. I think everyone will agree that the exact generality of Gauquelin's data (i.e., whether they represent all of France and Belgium or just the more populated areas) is a minor point. The more important question with regard to generalization is: will the Mars effect appear in data from other countries? Everyone will also agree, I believe, that this important question can only be resolved by collecting data from other countries. Thus, the non-random sampling has only a minor impact upon the already limited generality of the existing data.

The arguments of Zelen, et al. that the Mars effect occurred only for the Paris data⁶ and that this raises doubts about the reality of the effect, also require further consideration. Their main argument hinges around the existence of a "geographic distinction" of the Mars effect. What is the evidence that the Mars effect is different for Paris versus other areas? In essence, the argument seems to be that the Paris data are independently significant while the other data are not, hence, the Mars effect is restricted to the Paris data.

However, this is not a proper way to make inferences about differences between groups. Unless statistical evidence is provided that the Paris data are significantly different from the other data, the argument for "geographic distinctions" is unfounded. Zelen, et al., have not reported the required analyses. If appropriate statistical tests were carried out, I expect that they would find the comparison of Paris with Belgium to be significant but the comparison of Paris with the rest of France to not even remotely approach significance. Likewise, I doubt that the comparison of Paris with the rest of France and Belgium combined would come out significant.

The most serious doubts about the comments of Zelen, et al. arise when the overall strategy they have used is examined. In brief, we find this situation: a body of data with a small but significant effect overall was first divided into two groups (males and females) and then a small part of the data (the females) was discarded, noting "our analyses show no demonstrable effect" for this group. The remaining data were divided into three subgroups (location) and then into two other subgroups (source of data).

After examining the results of these various subgroups, the one with the strongest effect was selected out and, as if it was not a highly selected, post hoc situation, the fact that the rest of the data were not significant when this subgroup was removed was presented as a worrisome issue. In fact, given the size of the effect Gauquelin expected, it is not particularly surprising that a sample of 254 would not reach significance. The implication that this sample should be significant given the strong result in the Paris data also overlooks the fact that the Paris data were selected post hoc.

A subgroup specifically selected post hoc because it has the strongest effect certainly does not represent the magnitude of the effect that can be expected in the rest of the data. Also, under these circumstances, the speculations about differences between groups (even if the speculations were supported by statistical analyses) are unconvincing because of the post hoc multiple comparisons.

Finally, the comment by Zelen, et al., that "the experiment carried out by the Gauquelins has shown a statistical effect for Paris. However, it is not so for the entire group" is a remarkably poor choice of words. The one thing we do know is that the Mars effect was significant ($p < .05$) when the entire group (303 total or 294 males) was evaluated.

The point of this section is that just as multiple analyses and post hoc selection of results can be used to artifactually provide support for a hypothesis, the same dubious techniques can be used to try to weaken or discredit results. When advocates of a hypothesis find nonsignificant results overall, they can often discover post hoc significant differences between subsets of the data and then argue (with, of course, legitimate-sounding reasons for subdividing the data) that the effect is real, but "less general" than previously realized. When skeptics find significant results overall, they can just as often discover post hoc significant differences between subsets of the data and then argue (with the same kind of "legitimate" reasons for dividing the data) that the effect either does not exist or is so specific that it is of questionable interest.

DISCUSSION

A not uncommon strategy for evaluating a topic is to discuss a few instances that are in line with the reviewer's biases, be the biases favorable or unfavorable, and then draw general conclusions with a statement to the effect that the examples are typical of all the work. Following this strategy, I could now note that the examples given above are typical of the work of skeptics and call for a sweeping dismissal of all other work by skeptics. However, such a generalization would, of course, be dubious. What can be concluded is that since these kinds of biased errors do sometimes occur, the possibility of their occurrence should always be considered when evaluating analyses by skeptics.

The discussion so far has of necessity dealt only with errors that can be identified by critical evaluations of published information. As might be expected, there is every reason to believe that skeptics sometimes make biased errors that cannot be identified from the published reports. The available evidence indicates that recording and computational errors tend to occur in line with the researcher's expectation about the results.

Just as those favorable to a hypothesis tend to make errors in line with their expectations, those who expect negative results tend to make errors that bias the results towards chance. For reviews of the evidence, see T.X. Barber (1976) or Rosenthal (1976).

Thus, negative evaluations or findings by skeptics must be submitted to the same methodological scrutiny as positive findings by staunch proponents of a hypothesis. In particular, the argument that poor methodology is irrelevant when negative results were obtained is generally not a sound position – particularly if the work was done by skeptics. The negative results can be considered legitimate only to the extent that the researchers could not have made errors or otherwise biased the results to sabotage a genuine effect. The Wilson ESP study in

particular and to a lesser extent the Zelen, Kurtz, and Abell critique of the Gauquelins' work demonstrate that some skeptics will take rather extreme measures in attempting to neutralize positive results. As can be seen from the examples of erroneous statements by Gardner and Wheeler, careful evaluation is required for all aspects of work by skeptics, including their comments on the personal competence and integrity of researchers as well as their efforts to collect, analyze, and interpret data.

It may be valuable to conclude by mentioning a dilemma that arises when one tries to place people into categories of "skeptics" or "believers." A problem occurs when a person modifies his or her position on the basis of empirical studies. Since a person who is initially skeptical or neutral may become a "believer" after obtaining positive results, the criticism that only "believers" have obtained positive results in borderline areas of science has a degree of intrinsic (and trivial) truth. Likewise, since a person initially favorable to a hypothesis may become skeptical after obtaining negative results, the idea that skeptics generally have gotten chance results has an equal degree of intrinsic (and trivial) truth.

Classifying researchers according to initial beliefs in order to speculate upon the nature of the errors they might have made will certainly be a difficult and thankless task. Since strong a priori beliefs often seem to be an inevitable aspect of scientific work, the best course of action is to require high methodological standards across the board, independent of the researchers' attitudes or the experimental outcomes.

FOOTNOTES

1. One way to establish whether a skeptic is in the realm of fanaticism rather than science is to consider whether it would be possible to design, carry out, and report in a reasonable scientific format experiments that the skeptic would accept as sound evidence. If not, the scientific status of the skeptic should be questioned.

2. Gardner is, of course, not alone in letting his judgments about the quality of methodology be influenced by the agreement of the results with his biases. See, for example, the study of bias in experimental evaluations by Goldstein and Brazis (1970).

3. With Dr. Rhine's recent death, this case also forces the sobering realization that if similar stories are fabricated and irresponsibly propagated, Dr. Rhine and/or the other parties involved will not be in a position to force the record to be set straight – by implied threat of legal action if nothing else. It can only be hoped that if these situations do arise in the future, people will remember that the falsity of such reports was clearly established when correction was mandatory.

4. Prof. Zelen originally suggested a sample size of 100 to 200 cases, but Gauquelin and Gauquelin (1977) noted that such a sample would not be large enough that a significant result would be expected given the magnitude of the effect in the group of 2088.

5. In Zelen, et al. (1977) the discrepancies between the figures given in the text and in Table 1 raise some ambiguities. I am assuming the figures in Table 1 are correct.

6. Technical questions can be raised about the p value calculated for the Paris data. The statistic Zelen proposed is a normal approximation to a discrete distribution and, as such, is

appropriate only for large sample sizes. It is likely that the size of the Paris data ($N = 40$) is in the range where the approximation starts breaking down. Consider, for example, the similar case of obtaining 13/40 with $P = 1/6$ for the binomial distribution. The exact two-tailed probability of this outcome is $p = .021$. A normal approximation with a continuity correction gives $Z = 2.47$, $p = .014$, two-tailed. Without the continuity correction, the normal approximation gives $Z = 2.69$, $p = .007$, two-tailed, which is exaggerated by a factor of three. The Zelen statistic, as proposed and apparently employed, does not include a continuity correction.

REFERENCES

- Barber, B. Resistance by scientists to scientific discovery. *Science*, 1961, 134, 596-602.
- Barber, T.X. *Pitfalls in human research: Ten pivotal points*. New York: Pergamon Press, 1976.
- Brush, S.G. Should the history of science be rated X? *Science*. 1974, 183, 164-1172.
- Committee Para. The Committee Para replies to Gauquelin. *The Humanist*, 1976, 315(1), 31. (a)
- Committee Para. The Committee Para's reply to Gauquelin. *The Humanist*, 1976, 36_(3), 32-33. (b)
- Coover, J.E. *Experiments in psychical research*. Palo Alto, Calif.: Stanford University Press, 1975. (reprint)
- Coover, J.E. Reply to critics of the Stanford experiments on thought transference. *Journal of Parapsychology*, 1939, 3., 17-28.
- Gardner, M. *Fads and fallacies in the name of science*. New York: Dover, 1957.
- Gauquelin, M. *The cosmic clocks: From astrology to a modern science*. Henry Regnery, 1967.
- Gauquelin, M., & Gauquelin, F. The Zelen test of the Mars effect. *The Humanist*, 1977, 37.(6), 30-35.
- Goodstein, L.D., & Brazis, K.L. Psychology of scientists: XXX. Credibility of psychologists: An empirical study. *Psychological Reports*, 1970, 27, 835-838.
- Hansel, C.E.M. *ESP: A scientific evaluation*. Charles Scribner's Sons, 1966.
- Hansel, C.E.M. *ESP and parapsychology: A critical re-evaluation*. Buffalo, NY: Prometheus Books, 1980.
- Kuhn, T.S. *The structure of scientific revolutions*. Chicago: University of Chicago Press, 1962.
- Palmer, J. Extrasensory perception: Research findings. In S. Krippner (Ed.), *Advances in parapsychological Research 2: Extrasensory Perception*. New York: Plenum Press, 1978.
- Rhine, J.B. *Extra-sensory perception*. Boston: Humphries, 1935.
- Rhine, J.B., Pratt, J.G., Stuart, C.E., Smith, B.M., & Greenwood, J.A. *Extra-sensory perception after sixty years*. Boston: Humphries, 1940.
- Rosenthal, R. *Experimenter effects in behavioral research*. New York: Irvington, 1976.

Schmeidler, G.R. Personal values and ESP scores. *Journal of Abnormal and Social Psychology*, 1952, 47, 757-776.

Wheeler, J.A. Parapsychology -- a correction. *Science*, 1979, 205, 144.

Wilson, W.R. Do parapsychologists really believe in ESP? *Journal of Social Psychology*, 1964, 64, 379-389.

Zelen, M. Astrology and statistics: A challenge. *The Humanist*, 1976, 36(1), 32-33.

Zelen, M., Kurtz, P., & Abell, G. Is there a mars effect? *The Humanist*, 1977, 37(6), 36-39.

Return to: [Paranormal Phenomena Articles](#)