Statistical Problems in ESP Research

Persi Diaconis

Summary. In search of repeatable ESP experiments, modern investigators are using more complex targets, richer and freer responses, feedback, and more naturalistic conditions. This makes tractable statistical models less applicable. Moreover, controls often are so loose that no valid statistical analysis is possible. Some common problems are multiple end points, subject cheating, and unconscious sensory cueing. Unfortunately, such problems are hard to recognize from published records of the experiments in which they occur; rather, these problems are often uncovered by reports of independent skilled observers who were present during the experiment. This suggests that magicians and psychologists be regularly used as observers. New statistical ideas have been developed for some of the new experiments. For example, many modern ESP studies provide subjects with feedback—partial information about previous guesses—to reward the subjects for correct guesses in hope of inducing ESP learning. Some feedback experiments can be analyzed with the use of skill-scoring, a statistical procedure that depends on the information available and the way the guessing subject uses this information.

Is modern parapsychological research worthy of serious consideration? The volume of literature by reputable scientists, the persistent interest of students, and the government's funding of ESP projects make it difficult to evade this question. Over the past 10 years, in the capacity of statistician and professional magician, I have had personal contact with more than a dozen paranormal experiments. My background encourages a thorough skepticism, but I also find it useful to recall that skeptics make mistakes. For example, the scientific community did not believe in meteorites before about 1800. Indeed, in 1807 when a meteorite shower fell in Weston, Connecticut, an extended investigation was made by Professors Silliman and Kingsley of Yale. When Thomas Jefferson—then President of the United States and scientist of no small repute—was informed of the findings, he reportedly responded, "Gentlemen, I would rather believe that those two Yankee Professors would lie than to believe that stones fell from heaven" (*1*).

Critics of ESP must acknowledge the possibility of missing a real phenomenon because of the difficulty of designing a suitable experiment. However, the characteristics which lead many to be dubious about claims for ESP—its sporadic appearance, its need for a friendly environment, and its common association with fraud—require of the most sympathetic analyst not only skill in the analysis of nonstandard types of experimental design but appreciation of the differences between a sympathetic environment with flexible study design and experimentation which is simply careless or so structured as to be impossible to evaluate.

In this article I use examples to indicate the problems associated with the generally informal methods of design and evaluation of ESP experiments—in particular, the problems of multiple end points and subject cheating. I then review some of the commentaries of outstanding statisticians on the problems of evaluation. Finally, as an instance of using new analytic methods for nonstandard experiments, I give examples of some new statistical techniques that permit appropriate evaluation of studies that allow instant feedback of information to the subject after each trial, an entirely legitimate device used to facilitate whatever learning process may be involved.

Informal Design and Evaluation

A common problem in the evaluation of ESP experiments is the uncertainty about what outcomes are to be judged as indicative of ESP. Sometimes the problem can be dealt with by setting up a second experiment to verify the unanticipated but interesting outcome of a first experiment.

In a much discussed card-guessing experiment reported by Soal and Bateman (*2*), a receiving subject tried to guess the name of a card that was being thought about by a sending subject. When the data were first analyzed, no significant deviations from chance were observed. Several years later, the experimenters noticed that the guessing subject seemed to name not the card the sender was thinking about but rather the card two cards down in the deck (an example of precognition). Once this hypothesis was clearly formulated, the data were reanalyzed and new data were collected. The results stood up. The publication of Soal and Bateman's book touched off a series of lively articles (*2, 3*). The validity of Soal's experiment is still being debated [there are claims that the records are unreliable (*4, 5*)], but that he subjected the data to reanalysis after finding an unusual pattern seems acceptable to almost everyone. Whatever the view about reanalysis, the design and evaluation of the later experiments fall squarely within the domain of familiar scientific practice. The problems are more acute in the next example.

Three papers in the papers of the *Journal of Parapsychology* (*6*) describe experiments with a young man called B.D. These experiments took place at J. B. Rhine's Foundation for Research on the Nature of Man in Durham, North Carolina. The effects described, if performed under controlled conditions, seem like an exciting scientific breakthrough. In May of 1972, I witnessed a presentation by B.D., arranged by the Psychology Department of Harvard University. I was asked to observe as a magician, and made careful notes of what went on. Although the experiments were not controlled, I believe they highlight many problems inherent in drawing inferences from apparently well-controlled experiments.

Most of the demonstrations I witnessed B.D. perform involved playing cards. In one experiment, two onlookers were invited to shuffle two decks of cards, a red deck and a blue deck. Two other onlookers were asked to name two different cards aloud; they named the ace of spades and the three of hearts. Both decks were placed face down on a table. We were instructed to turn over the top cards of each deck simultaneously and to

The author is an assistant professor in the Department of Statistics, Stanford University, Stanford, California 94305.

0036-8075/78/0714-0131$01.50/0

continue turning up pairs in this manner until we came to either of the named cards. The red-backed three of hearts appeared first. At this point, B.D. shouted, "Fourteen," and we were instructed to count down 14 more cards in the blue pack. We were amazed to find that the 14th card was the blue-backed three of hearts. Many other tests of this kind were performed. Sometimes the performer guessed correctly, sometimes he did not.

Close observation suggested that B.D. was a skilled opportunist. Consider the effect just described. Suppose that, as the cards were turned face upwards, both threes of hearts appeared simultaneously. This would be considered a striking coincidence and the experiment could have been terminated. The experiment would also have been judged successful if the two aces of spades appeared simultaneously or if the ace of spades were turned up in one deck at the same time the three of hearts was turned up in the other. There are other possibilities: suppose that, after 14 cards had been counted off, the next (15th) card had been the matching three of hearts. Certainly this would have been considered quite unusual. Similarly, if the 14th or 15th card had been the ace of spades, B.D. would have been thought successful. What if the 14th card had been the three of diamonds? B.D. would have been "close." In one instance, after he had been "close," B.D. rubbed his eyes and said, "I'm certainly having trouble seeing the suits today."

A major key to B.D.'s success was that he did not specify in advance the result to be considered surprising. The odds against a coincidence *of some sort* are dramatically less than those against any prespecified *particular one* of them. For the experiment just described, including as successful outcomes all possibilities mentioned, the probability of success is greater than one chance in eight. This is an example of exploiting multiple end points. To further complicate any analysis, several such ill-defined experiments were often conducted simultaneously, interacting with one another. The young performer electrified his audience. His frequently completely missed guesses were generally regarded with sympathy, rather than doubt; and for most observers they seemed only to confirm the reality of B.D.'s unusual powers.

Subject Cheating

In the experiments at Harvard, B.D. occasionally helped chance along by a bit of sleight of hand. During several trials, I saw him glance at the bottom card of the deck he was shuffling. He then cut the cards, leaving a quarter of an inch step in the pack. This fixed the location of the card he had seen. The cards were then spread out and a card was selected by one of the onlookers. When the selected card was replaced in the deck, B.D. secretly counted the number of cards between the card he had seen and the selected card. B.D. named a "random" card (presumably the card he had glanced at) and asked someone to name a small number. He disregarded the first number named and asked someone else to name another small number—this time the difference in location between the card B.D. had seen and the selected card. One of the observers counted down in the pack until he came to the "randomly" named card. Addressing the observer who originally selected a card, B.D. asked, "What card are you thinking of?" Sure enough, when the second small number was counted off, the selected card appeared. When presented in the confusing circumstances I have described, the trick seemed impossible. About ten of the observers were psychology faculty, the remaining five were graduate students. When they tried to reconstruct the details of this presentation, they could not remember exactly who had thought of the number and who had selected the card. They muddled the circumstances of this particular test with those of previous tests. I call this blending of details the "bundle of sticks" phenomenon. It is a familiar element in standard magic tricks: An effect is produced several times under different circumstances with the use of a different technique each time. When an observer tries to reconstruct the modus operandi, the weak points of one performance are ruled out because they were clearly not present during other performances. The bundle of sticks is stronger than any single stick.

B.D.'s performance went on for several hours. Later, some of the observers realized that B.D. often took advantage of the inevitable lucky breaks. However, his performance must have made quite an impression on some of the observers because the 13 July 1973 issue of *Science* reported that B.D. had been given a grant from Harvard "to explore the nature of his own psychic ability." My personal curiosity about the possibility of B.D. having powers that upset the known physical laws is fully satisfied—in the negative. This position is further discussed below.

Another exposé of which I have first-hand knowledge concerns Ted Serios. Serios claimed that he could create psychic photographs on Polaroid film in cameras he had never seen before. A group of scientists in Chicago and Denver had become convinced that there was no trickery involved; indeed, they believed that Serios had extraordinary psychic abilities. I became involved when Eisenbud's book, *The World of Ted Serios* (*7*), was being considered for review by *Scientific American*. A team of experienced magicians went to Denver to take a close look at Serios' performance. When we arrived, Serios was attempting to produce psychic images on TV film at a Denver TV station. Conditions were chaotic. Several news teams were present, each team having brought its own Polaroid film. After a short time, I managed secretly to switch about 20 boxes of their film with marked film we had brought along. We wanted to determine whether their film had been previously exposed. It had not been. The fact, however, that it had been so easy for me to switch the film by sleight of hand clearly indicated that the investigators did not have adequate control over the essential materials. Conditions remained like this during our several days' stay, and our observation revealed irreparable methodological flaws in all phases of the experiments. Serios openly used a small paper tube which he placed on his forehead pointing toward the camera "to help focus the thought waves." I observed that he occasionally placed this tube in front of the camera lens. On one trial, I thought I saw him secretly load something into the tube. When I asked to examine the tube, pandemonium broke loose. Several of the Denver scientists present jumped up, shouting things like, "You can't do that!" Serios hastily put the tube in his pocket. He was not searched. We were later able to duplicate Serios' pictures in several ways. After our exposé (*8*) of how we believe Serios obtained his results, *Life* magazine published an article about Serios' psychic powers, with no mention of our findings. Paranormal claims tend to receive far more media coverage than their exposés.

There are many other reports of subject cheating in ESP experiments. For example, Gardner (*9*) figured out how Russian women "saw" with their fingertips and, in a recent paper (*10*), exposes Uri Geller's supposedly "foolproof" alteration of the internal memory of several pieces of Nitinol wire. Nitinol is an alloy of nickel and titanium which has a memory. Under intense heat, a piece of Nitinol wire can be given a shape. When

cold, it can easily be reshaped between the fingers. After being heated, it snaps back to the original shape. One of the most persistently quoted proofs of Geller's paranormal powers is Eldon Byrd's claim that "Geller altered the lattice structure of a metal alloy in a way that cannot be duplicated." As usual, there is a story of amazing feats performed under test conditions (*11*). Gardner's competent detective work reveals the usual tale of chaotic conditions and bad reporting. There is an interesting twist here. Supporters of Geller argue that the event is amazing, even in light of chaotic conditions, since Geller could not have had access to a heat source of about 500°C, "the only known way to get this result" (*11*). Gardner found he could easily alter the memory of a piece of Nitinol wire with a pair of pliers or even by using his teeth.

Unfortunately, a nonmagician's memory of a magic feat is unreliable. For example, Hyman, a psychologist and magician, has described his visit to the Stanford Research Institute, during which Geller demonstrated many of his psychic feats (*12*). Hyman reports observing sleight of hand performed under uncontrolled conditions, much at variance with the published report (*13*) of the SRI scientists involved. Geller probably ranks as the most thoroughly exposed psychic of all times (*12, 14, 15*); yet the parascience community continues to defend him as a psychic who is often genuine, even though he occasionally cheats.

Some Conclusions

Rejecting the claims of a psychic who has been caught cheating raises thorny scientific problems. I am sure that B.D. used sleight of hand several times during the performance I witnessed. Yet, as one of the other observers remarked, "The people who introduced B.D. never said he didn't do card tricks; they just claimed he had extraordinary powers on occasion." During my encounter with Serios, a psychologist present put it differently: "Suppose he was only genuine 10 percent of the time; wouldn't that be enough for you?" My position is conservative: the similarity of the descriptions of the controlled experiments with B.D. and Serios to the sessions I witnessed convinces me that all paranormal claims involving these two performers should be completely discounted.

The fact that a trained observer finds reason to discredit two psychics is not, of course, sufficient evidence to discredit the existence of ESP or the integrity of other potential psychics. However, the pervasiveness of fraud in so many claims for ESP makes it extremely difficult for the disinterested observer to identify evidence worthy of credit. Whether Houdini was a disinterested witness, as he claimed, is hard to judge (*16*). But his tireless investigation and exposure of spiritualists in England and America (*16*) give powerful evidence of the extent of fraud in this domain and of the difficulties of detecting it. Randi, also a professional magician, has recently undertaken a detailed exposé of Uri Geller. Randi repeatedly documents the discrepancy between actual circumstances and those reported in newspapers and scientific journals (*14*).

Even if there had not been subject cheating, the experiments described above would be useless because they were out of control. The confusing and erratic experimental conditions I have described are typical of *every* test of paranormal phenomena I have witnessed. Indeed, ESP investigators often insist on nonnegative observers and surroundings. Because of this, skeptics have a difficult time gaining direct access to experimental evidence and must rely on published reports. Such reports are often wholly inadequate. According to Davey (*17*), Hansel (*5*), and others, it is not easy to notice crucial details during ESP experiments. For example, each of the studies referred to above describes experimental conditions beyond reproach. My own observation suggests that the conditions were not in control. Some of these problems can be overcome by insisting that expert magicians and psychologists, skilled at running experiments with human subjects, be included in study protocols.

Statisticians and ESP

The only widely respected evidence for paranormal phenomena is statistical. Classical statistical tests are reported in each of the published studies described above. Most often these tests are "highly statistically significant." This only implies that the results are improbable under simple chance models. In complex, badly controlled experiments simple chance models cannot be seriously considered as tenable explanations; hence, rejection of such models is not of particular interest. For example, the high significance claimed for the famous Zenith Radio experiment is largely a statistical artifact (*18*). Listeners were invited to mail in their guesses on a random sequence of playing cards. The proportion of correct guesses was highly significant when calculations were based on the assumption of random guessing on the part of each listener. It is well known (*19*) that the distribution of sequences produced by human subjects is far from random, and hence the crucial hypothesis of independence fails in this situation. More sophisticated analysis of the Zenith results gives no cause for surprise.

In well-run experiments, statistics can aid in the design and final analysis. The idea of deliberately introducing external, well-controlled randomization in investigation of paranormal phenomena seems due to Richet (*20*) and Edgeworth (*21*). Later, Wilks (*22*) wrote a survey article on reasonable statistical procedures for analyzing paranormal experiments popular at the time. Fisher developed new statistical methods that allow credit for "close" guesses in card-guessing experiments (*23*). Good (*24*) continues to suggest new experiments and explanations for ESP. The parascience community, well aware of the importance of statistical tools, has solved numerous statistical riddles in its own literature. Any of the three best known parascience journals is a source of a number of good surveys and discussions of inferential problems (*25*).

The actual circumstances of even well-run ESP tests are sufficiently different from the most familiar types of experiment as to lead even able and well-regarded analysts into difficulty; and the statistical community has a mixed record, with errors in both directions. On one hand, the celebrated statement by the Institute of Mathematical Statistics (*26*) was widely regarded as an endorsement of ESP analysis methods, a position that seems hard to justify. As an example of unjust criticism of ESP, consider Feller's review (*27*) of the methodology of ESP research (*28*).

Feller was an outstanding mathematician who made major contributions to the modern theory of probability. He attacked some of the statistical arguments used by J. B. Rhine and his co-workers (see *27*). It appears now that several of Feller's criticisms were wrong. To give one instance: a standard ESP deck consists of five symbols repeated five times each to make up a 25-card deck. Feller found published records of the order of ESP decks before and after shuffling. He noticed that one could match up long runs of consecutive symbols in the two orders and took this as evidence of "unbelievably poor results of shuffling" (*29*). In a follow-up article, Greenwood and Stuart (*28*) pointed out that such runs of matching symbols did not prove poor

mixing. Since each symbol is repeated five times, long runs of matching symbols are inevitable. Feller had no respect for their remarks: "Both their arithmetic and their experiments have a distinct twinge of the supernatural," he wrote years later (*29*).

I believe Feller was confused. As proof of this, consider one of the experiments that Greenwood and Stuart (*28*) carried out to prove their point: they simulated two arrangements of ESP decks from a table of random numbers, and they showed that random arrangements exhibited long runs of matching symbols. Feller completely misunderstood this experiment; he thought that Greenwood and Stuart chose a sample of 25 from a set of five symbols *with replacement*. If the simulation were done by sampling with replacement, only those outcomes that had exactly five of each symbol would be useful. Since these are rare, the time required to complete the simulation reported by Greenwood and Stuart would have been lifetimes long. Thus, Feller found the report of the resulting samples "miraculously obliging." The comments of Feller that I have quoted, suggesting that the investigators were at best incompetent, persisted through three editions of his famous text. I have asked students and colleagues of Feller about this, and all have said that Feller's mistakes were widely known; he seemed to have decided the opposition was wrong and that was that.

Feedback Experiments

If ESP phenomena are real, we still do not know a reliable method for eliciting them; and any serious exploration of the subject requires that as much leeway as possible be provided for experimental designs that seem likely to produce an effect. In their search for replicable experiments, psychic investigators have modified the classical tests of ESP. Important changes include the use of targets of increasing complexity such as drawings or natural settings and greater use of feedback, either telling the subject whether the guess was right or wrong, or, in a card-matching experiment, what the last target card actually was. Unfortunately, the statistical tools for evaluating the outcome of more complex experiments are not available, and the ad hoc tests created by researchers are often not well understood. An article on remote viewing (*30*) provides an example. Apparently, in a typical phase of the experiment, nine locations (a local swimming pool, tennis court, and others) were selected from a list of 100 locations chosen to be as distinct as possible. A team of sending subjects went to each of the nine locations in a random order. A guessing subject tried to describe where they were. After each guess the guessing subject was given feedback by being taken to the true location. This is clearly a complex experiment to evaluate, and there are several reasons to discount the findings presented in (*30*). I give some of these reasons at the end of the next section. I first focus on the analysis of simpler feedback experiments.

Feedback of some sort is a much-used technique in modern ESP research (*31*). The appropriate analysis of a feedback experiment is easy in some simple cases but not at all clear in other cases. The assessment of such experiments requires new methods. Graham and I have explored some of the problems in a situation simple enough to allow mathematical analysis (*32*), and the following examples are drawn from that research.

Let us consider an experiment that involves a sending subject, a receiving subject, and a well-shuffled deck of 52 cards. The sending subject concentrates on each card in turn, and the receiving subject attempts to guess the suit and number of the card correctly.

No information case. If no additional information is available to the receiving subject, the chance of a correct guess at any point in the experiment is 1 in 52; thus, the expected number of correct guesses in a single run through the 52-card deck is 1. If we do not accept ESP as possible, it can easily be shown that any system of guessing leads to one correct guess on the average. However, the distribution of the number of correct guesses can vary widely as a function of the guessing strategy: if the same card is guessed 52 times in succession, then exactly one guess will be correct. It has been shown that the variance of the number of correct guesses is largest when each card is called only once (*33*).

Complete feedback case. Next, let us consider an experiment that includes giving information to the guesser. After each trial he is shown the card he has attempted to identify. The most efficient way the guesser can use this information is always to name a card he knows to be still in the deck. This strategy leads to an expected number of correct guesses of

$$\frac{1}{52} + \frac{1}{51} + \frac{1}{50} + \cdots + 1 \doteq 4.5$$

in a single run through the deck, much larger than the one correct we expect with no information.

Partial information case. A third situation is created by giving only partial information. The guesser is told only if each guess is correct or not. In this situation, it can be shown that the guesser's optimal strategy is to name repeatedly any card—for example, the ace of spades—until he is told his guess is correct. After he is told that he has guessed correctly, he then repeatedly calls any card known to be in the deck until that card is guessed correctly or the run through the deck is completed. The expected number of correct guesses, if this optimal strategy is used, is

$$\frac{1}{52!} + \frac{1}{51!} + \frac{1}{50!} + \cdots + 1 \doteq e - 1 \doteq 1.72$$

where e is the base of the natural logarithms. A subject given partial information can minimize the expected number of correct guesses by naming cards without repeating the same card until a correct guess is made. The guesser then repeatedly calls the card known not to be in the deck for the remaining calls. The expected number of correct guesses in this situation is well approximated by

$$1 - \frac{1}{e} \doteq .632$$

Similar analysis can be carried out with the standard 25-card ESP deck, consisting of five different symbols repeated five times. If no feedback information is given to the guessing subject, then, under the hypothesis of chance guessing, each guess has probability 1/5 of being correct. In a run through the 25-card deck, five correct guesses are expected. In the case of complete feedback, the best strategy is to guess the most probable card at each stage. This leads to 8.65 as the expected number of correct guesses, as shown by Read (*34*). In the case of partial information—telling the guesser only if each guess is right or wrong—things are more complicated. For example, the optimal strategy no longer is to choose the most probable card for each guess. It is easy to give a simple strategy that gets six cards correct on the average: Guess a fixed symbol until told that five correct guesses have been achieved, and then guess a second symbol for the remaining cards. There seems to be no simple closed-form expression for the optimal strategy; but the expected number of correct guesses, if the optimal strategy is used, satisfies a multivariable recurrence that makes dynamic programming techniques available. Gatto at Bell Laboratories succeeded in putting this problem on the

computer and, by solving the recurrence, showed that the expected number of correct guesses is 6.63, if the optimal strategy is used (*35*). The result took about 15 hours of CPU (central processing units) time on a large computer.

These examples show that feedback can drastically change the expected number of correct guesses.

Simple Guessing Experiments with Feedback: Scoring Rules

Available evidence (*19*) suggests that subjects do not use their best possible strategies in simple probabilistic experiments. In more complicated situations—for example, if the experimenter uses a deck of cards with values repeated several times and gives the subject feedback as to whether his guess is "close" or not—the most efficient strategy may be very difficult to compute. Tart (*31*) gives references to the use of scoring rules that range from not taking into consideration the amount of information available to including the assumption that the subject is using the optimal strategy. Both of these approaches seem unnecessarily crude. The former might give an untalented subject a high score, while the latter might penalize a skillful subject who does not make efficient use of the information available to him.

For problems of this type, there exists a class of scoring rules which depend on the amount of information available to the subject and on the way the subject uses the information given. The idea is to subtract at the i^{th} stage the probability of the i^{th} guess being correct, given the history up to guess i. For example, if a guesser names a card he knows not to be in the deck, no penalty is subtracted. More formally, if G_i is the subject's guess on the i^{th} trial and Z_i is one or zero as the i^{th} guess is correct or not, then the skill-scoring statistics for n trials is defined by

$$S = \sum_{i=1}^{n} \{Z_i - E(Z_i|G_1, G_2, \cdots, G_i, Z_1, Z_2, \cdots, Z_{i-1})\} \quad (1)$$

The conditional expected values that appear in Eq. 1 can be calculated for any past history with the use of new combinatorial formulas related to problems of permutations with restricted positions (*32*). The statistic S is related to the skill-scoring rules used to evaluate weather forecasters (*36*). S has the property that, in the absence of skill (that is, ESP or talent), the expected score is zero for any guessing strategy, optimal or not.

Table 1. Card guessing with ten cards and partial feedback. Column 1 is trial number; column 2 is subject's guess; column 3 is feedback to subject; column 4 is the probability of the i^{th} guess being correct, given the history up to time i; for example, subject guessed card 9 on trial 2 after being told that the guess on trial 1 was wrong, penalty = probability (9 on trial 2 given that the guess was wrong on trial 1) = 8/81; column 5 is the actual card in i^{th} position. $S = 3 - 1.0874 = 1.9126$

Trial	Guess	Feedback	Penalty	Card
1	1	Wrong	0.1000	3
2	9	Wrong	0.0988	4
3	6	Wrong	0.0976	8
4	3	Wrong	0.0965	6
5	2	Right	0.0955	2
6	1	Wrong	0.1189	10
7	4	Wrong	0.1031	9
8	7	Right	0.1019	7
9	6	Wrong	0.1282	5
10	1	Right	0.1470	1
Total			1.0874	

Let us consider an example made explicit in Table 1. A deck of ten cards numbered from 1 to 10 was well mixed. A sender looked at the cards in sequence from the top down, and a guesser guessed at each card as the sender looked at it. After each trial the guesser was told whether she was correct or not. There were three correct guesses. If one ignores the availability of partial information, one comes to the conclusion that this response was two more than could be expected by chance. If one assumes that the guesser used the optimal strategy outlined in the partial information example in the previous section, then one would compare the number of correct guesses with 1.72, the expected number of correct guesses under the optimal strategy. Thus, one would conclude that the score of 3 was 1.28 higher than "chance." The guesses which were actually made are far from the optimal strategy. For example, on the second trial the optimal guess was 1, not 9; on the third trial the optimal guess was 1 or 9, not 6. In this case, the skill-scoring statistic scores this experiment as 1.91 higher than chance. Skill-scoring statistics can be tested by using an appropriate normal approximation available via Martingale central limit theorems (*32*).

Skill-scoring provides an example of how mathematical statistics can be used to evaluate experiments under nonstandard conditions. Clearly, experiments designed to include both feedback and sampling with replacement will be far easier to evaluate. The problems dealt with above—dependent trials coupled with feedback—arise in practice. For example, the analysis can be applied for reassessing experiments where subjects were seated within sight or hearing of one another, and an investigator suspects that unconscious sensory cuing has taken place. To be specific, a sender might, by his behavior, unconsciously indicate to the receiver whether his last guess was correct or not. This assumes, of course, that right and wrong were the only information cues transmitted. If the investigator thinks that the sender cued the guesser with information about each card as he looked at it, no statistical analysis can salvage the data.

One problem with feedback experiments is that they seem highly sensitive to clean experimental conditions. If the conditions break down, it will be hard to make sense of the data. For example, if a random number generated in an experiment with feedback is faulty, it may be that subjects can learn something of the pattern from the feedback (*37*). In the remote viewing experiment (*30*) referred to above, subjects included reports of where they had been taken during a "feedback trip" in the description of a current target. When a judge is given the subjects' nine transcripts, the judge is told which nine targets were visited but not the order of the visits. Information within a transcript allows a judge to rule out some of the potential targets and renders analysis of the results impossible. This is only one of many objections to the findings in (*30*). Because of inadequate specification of crucial details (*38*), I find it impossible to interpret what went on during this experiment.

Conclusions

To answer the question I started out with, modern parapsychological research *is* important. If any of its claims are substantiated, it will radically change the way we look at the world. Even if none of the claims is correct, an understanding of what went wrong provides lessons for less exotic experiments. Poorly designed, badly run, and inappropriately analyzed experiments seem to be an even greater obstacle to progress in this field than subject cheating. This is not due to a lack of creative investigators who work hard but rather to the difficulty of finding an appropriate balance between study designs which both permit analysis and experimental results. There always seem to be many loopholes and loose ends. The same mistakes are made again and again. The critiques and comments of Davey (*17*) and Hall (*39*) seem as relevant for modern studies as they did at the turn of the cen-

tury. Regrettably, the problems are hard to recognize from published records of the experiments in which they occur; rather, these problems are often uncovered by reports of independent skilled observers who were present during the experiment.

There have been many hundreds of serious studies of ESP, and I have certainly read and been told about events that I cannot explain. I have been able to have direct experience with more than a dozen experiments and detailed secondhand knowledge about perhaps 20 more. In every case, the details of what actually transpired prevent the experiment from being considered seriously as evidence for paranormal phenomena.

References and Notes

1. As quoted in H. H. Nininger, *Our Stone-Pelted Planet* (Houghton Mifflin, Boston, 1933).
2. S. G. Soal and F. Bateman, *Modern Experiments in Telepathy* (Yale Univ. Press, New Haven, Conn., 1954).
3. G. R. Price, *Science* **122**, 359 (19 ➡ S. G. Soal, *ibid.* **123**, 9 (19 ➡ J. B. Rhine, *ibid.*, pp. 11 and ➡ P. E. Meehl and M. Scriven, *ibid.*, p. ➡ P. W. Bridgman, *ibid.*, p. ➡ G. R. Price, *ibid.*, p. ➡ *ibid.* **175**, 359 (1972).
4. C. Scott and P. Haskel, *J. Soc. Psych. Res.* **118**, 220 (1975).
5. C. E. M. Hansel, *ESP a Scientific Evaluation* (Scribners, New York, 1966).
6. E. F. Kelly and B. Kanthanani, *J. Parapsychol.* **36**, 185 (1972); B. Kanthanani and E. F. Kelly, *ibid.* **38**, 16 and 355 (1974).
7. J. Eisenbud, *The World of Ted Serios* (Morrow, New York, 1967).
8. D. Eisendrath and C. Reynolds, *Pop. Photogr.* **61** (No. 4), 81 (1967); J. Eisenbud, *ibid.* **61** (No. 5), 31 (1967).
9. M. Gardner, *Science* **151**, 654 (1966).
10. ———, *The Humanist* **37** (May/June), 25 (1977).
11. E. Byrd, in *The Geller Papers*, C. Panati, Ed. (Houghton Mifflin, Boston, 1976).
12. R. Hyman, *The Humanist* **37** (May/June), 16 (1977).
13. H. Puthoff and R. Targ, *Mind Reach* (Delacorte, New York, 1977).
14. J. Randi, *The Magic of Uri Geller* (Ballantine, New York, 1976).
15. D. Marks and R. Kamman, *Zetetic* **1** (No. 2), 3 (1977).
16. H. Houdini, *A Magician Among the Spirits* (Harper, New York, 1924).
17. S. J. Davey, *J. Soc. Psych. Res.* **3**, 8 (1887).
➡ L. D. Goodfellow, *J. Exp. Psychol.* **23**, 601 (1938).
➡ P. Slovic, B. Fischoff, S. Lichenstein, *Annu. Rev. Psychol.* **28**, 1 (1977); A. Tversky and D. Kahneman, *Science* **185**, 1124 (1974).
20. C. Richet, *Rev. Philos.* **18**, 41 (1884).
21. F. Y. Edgeworth, *Proc. Soc. Psych. Res.* **3**, 190 (1885); **4**, 189 (1885). For a historical review see M. McVaugh and S. H. Mauskopf [*Isis* **67**, 161 (1976)].
22. S. S. Wilks, *N.Y. Statistician* **16** (No. 6), (1965); **16** (No. 7), (1965).
23. R. A. Fisher, *Proc. Soc. Psych. Res.* **34**, 181 (1924); *ibid.* **38**, 269 (1928); *ibid.* **39**, 189 (1929).
24. I. J. Good, *Parasci. Proc.* **1** (No.2), 3 (1974), and references given therein.
25. For a useful survey of this literature, see D. S. Burdick and E. F. Kelly, "Statistical methods in parapsychological research," in *Handbook of Parapsychology*, B. Wolman, Ed. (Van Nostrand, New York, 1977).
26. B. H. Camp, *J. Parapsychol.* **1**, 305 (1937) (statement in notes section).
27. W. Feller, *ibid.* **4**, 271 (1940).
28. J. A. Greenwood and C. E. Stuart, *ibid.*, p. 299.
29. W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, ed. 3, 1968), pp. 56 and 407.
30. H. E. Puthoff and R. Targ, *Proc. IEEE* **64**, 329 (1976).
31. C. Tart, *Learning to Use ESP* (Univ. of Chicago Press, Chicago, 1976), chaps. 1 and 2.
32. P. Diaconis and R. L. Graham, "The analysis of experiments with feedback to subjects," *Ann. Stat.*, in press.
33. J. A. Greenwood, *J. Parapsychol.* **2**, 60 (1938) and references therein.
➡ R. C. Read, *Am. Math. Mon.* **69**, 506 (1962).
35. M. A. Gatto, personal communication.
36. H. R. Glahn and D. L. Jorgensen, *Mon. Weather Rev.* **98**, 136 (1970).
37. M. Gardner, *N.Y. Rev. Books* **24** (No. 12), 37 (14 July 1977).
38. R. Hyman, *The Humanist* **37** (November/December), 47 (1977); D. M. Stokes, *J. Am. Soc. Psych. Res.* **71**, 437 (1977).
39. G. S. Hall, *Am. J. Psychol.* **1**, 128 (1887).

40. I thank Tom Cover, Bradley Efron, David Freedman, Martin Gardner, Mary Ann Gatto, Seymour Geisser, Judith Hess, Ray Hyman, William Kruskal, Paul Meier, Lincoln Moses, Frederick Mosteller, David Siegmund, Charles Stein, Stephen Stigler, Charles Tart, and Sandy Zabell for comments on earlier versions. Partially supported by NSF grant MPS74-21416.

NEWS AND COMMENT

NATO Builds a Better Battle Tank But May Still Lose the Battle

The battle tank is still the principal weapon of a modern army. Far from driving the tank into extinction, technological developments such as the antitank missile have only hastened its rate of evolution. For the past 15 years the United States has stumbled from one fiasco to another in its attempts to design a new main battle tank, but seems at last to have a winner.

Both the failure and success of the tank development program are integrally related to a central crisis of the NATO alliance, the lack of cooperation in designing, developing, and producing new weapons. Through failure to standardize, the NATO allies at present field 31 different antitank weapons and seven different tanks. Such diversity causes a formidable logistics problem. It is the product of duplicative national research programs which waste about a third of the alliance's general purpose R & D budget. It is a principal factor in the alarming paradox that the backward economies of the Warsaw Pact can outproduce advanced NATO economies in tanks by a ratio of 4 to 1. NATO has recently cut the ratio to 2 to 1 yet still has only 7000 tanks deployed in Europe against the Warsaw Pact's 19,000. Nor does the quality of NATO tanks offset the gross deficiency in numbers. Germany's Leopard 1 and America's M60 are only about as capable as the Soviet T-72, not by any means its superior.

Though everyone agrees on the importance of NATO standardization, the commonly proposed remedies often seem worse than the disease. European countries, already fretful that they buy $8 of military equipment from the United States for every $1 they sell, view calls for standardization as another pressure to buy American. To offset its lack of appetite for European weapons, the United States has tried to develop weapons jointly with its allies, but with notable lack of success. Nowhere have the inherent problems of standardization been more vividly brought to light than in the Sisyphean attempts by America, Germany, and Britain to cooperate in developing a new main battle tank.

Tank technology is one of the military arts in which the United States does not possess a commanding lead; the Soviet, German, and British traditions of tank design have probably been superior. A British designed gun, the 105-mm cannon, is used by the tanks of all three NATO nations, and a revolutionary method of tank protection, known as Chobham armor, is also a British invention. German tanks, with their superior range and accuracy, were generally predominant in World War II until outnumbered. In part because of German expertise, Secretary of Defense Robert McNamara in 1963 initiated a German-American project to build a new main battle tank for the 1970's, the MBT-70.

The designers of the MBT-70 produced a tank that could squat, so as to lower its silhouette. They put the driver in the turret, instead of the hull, and kept him facing forward when the turret turned by a counter-rotating cylinder. "It was an all singing, all dancing, thing. Everybody thought it was absolutely marvelous but far too expensive and far too complicated for any crew to handle," says one NATO observer. As the cost approached $1 million a tank, Congress killed the MBT-70 in 1969. Both sponsoring countries went their separate ways, the Germans starting work on the Leopard 2 and the American

0036-8075/78/0714-0136$01.00/0