1982

MATHEMATICAL INFERENCE STRATEGIES VERSUS PSI: INITIAL EXPLORATIONS WITH THE PROBABILISTIC PREDICTOR PROGRAM

Charles T. Tart
Eugene Dronek
University of California

Classical testing procedures in parapsychological research involve (a) the randomization of a sequence of numbers or items (the target sequences); (b) responses by a percipient as to what he believes the order of the target sequence is in the lack of any known sensory cues (the calls); (c) a comparison of the number of correspondences between corresponding temporal calls and targets in the target and response sequences (scoring); and (d) a computation of the probability of getting the obtained or a greater number of such correspondences or hits, given the a priori chance probabilities of a hit on any particular trial and the standard assumptions made for the statistical tests used.

While it has long been recognized that a given percipient's response sequence is usually quite non-random, this lack of randomicity in the response sequence is irrelevant to testing the null hypothesis of no information transfer about the target sequence, via ESP, to the percipient. Randomicity of the target sequence, on the other hand, is a necessary assumption for the standard statistical tests used to be valid, so it has generally been believed that lack of randomicity in the target sequence may invalidate the results of our statistical tests.

The effects of non-randomicity in the target sequence may manifest

in either of two ways. First, consider the case where there is no feedback given about individual target identity to a percipient before he or she has completed all of his calls to a target sequence from a uniformly biased source. It is possible that if a particular percipient's global response biases just happen to match the biases in the target sequence, a 'statistically significant' hit score will result, but this is not due to any ESP but rather to what we, in retrospect, realize is an incorrect specification of the a priori probability of hits. Control scoring can usually handle this global bias problem. In the second case, if feedback of target identity is given to the percipient after each call, we can postulate that an occasional percipient may gradually assess what the biases in the target sequences are by keeping track of what previous targets have been and adjust his response pattern accordingly, that is, he may develop a mathematical inference strategy, thus artifactually inflating his score and invalidating the standard statistical test of the null hypothesis (note 1). The specifics of this second possibility have not been adequately explored to date in the literature.

RANDOMICITY

To assess the practical significance of a mathematical inference strategy for artifactually inflating hits, we need to examine the concept of randomicity more closely. Randomicity is conventionally taken as meaning a lack of order, but this definition can be misleading without further examination. It is generally agreed among mathematicians (see, for example, Chatian, 1975) that you may take any sequence of numbers generated by a truly random generator, and in retrospect, construct a mathematical algorithm that, when expanded, gives the exact set of numbers seen in this particular sequence. The sequence may be any length, from half a dozen digits to millions. While constructing the algorithm becomes much more complex for longer sequences, it can, in principle, be done. We postdict the sequence. This seems to imply that the sequence was not generated in a random fashion after all, but by a complex but completely deterministic and lawful generator whose operating characteristics are described by the particular algorithm.

The catch, however, is that the algorithm you have constructed for a sequence of N numbers will not predict subsequent numbers obtained from the same random source on trials N+1, N+2, etc. with any greater

accuracy than would be predicted by chance. To put it in more general terms, we may come up with a set of concepts which seem to account for the generation of any observed set of events or numbers in retrospect, but while we can always imagine such concepts, they do not necessarily have anything to do with the actual mechanism of generation of the events or numbers. This is the mathematical version of the psychological concept of rationalization: we can always imagine patterns in retrospect. The concept of randomness being a lack of order or pattern per se then is not really accurate, as some kind of pattern can always be projected in retrospect.

The more important property that we imply by randomicity in parapsychological (and general scientific) research is that randomicity means lack of predictability. Given a sample of numbers of length N from a random number generator (RNG), if your generator is truly random, your knowledge of these N numbers does not help you predict the output on further trials N+1, N+2, etc., with greater than chance expectancy. This lack of predictability is what is crucial for parapsychological testing. Given feedback to the percipient about previous target sequences, if we see any above chance scoring we want to be able to attribute it to some kind of ESP, rather than to some kind of mathematical inference strategy which the percipient might have worked out, consciously or unconsciously, from the sample of target numbers presented to him to date. Postdiction, of course, is not prediction.

Ordinarily, if the target sequence used in an experiment meets the conventional tests of randomicity we usually assume non-predictability. These conventional tests of randomicity would usually be a test for equiprobability of each possible target number (target singlets), and of serial independence of target generation, measured by tests of the equiprobability of each possible sequential pair of numbers of the target sequence (target doublets), each possible sequential triplet (target triplets), etc.. Given the usual experimental ways of generating random numbers, such as thorough shuffling of cards or electronic RNGs, where we do not have any theoretical reason to suspect higher order patternings at the triplet, quadruplet, and higher levels, testing for singlet and doublet bias is ordinarily considered sufficient. Pseudo-random generators require more complete testing, and are rarely used in parapsychological research.

REJECTING DATA

Naturally we should always strive to have acceptably random target sequences, but suppose we carry out a parapsychological experiment and discover, in retrospect, that some of our target sequences were significantly nonrandom by the above-mentioned tests of equiprobability of singlets and doublets? A frequent reaction is that such lack of randomicity of the target sequence somehow totally invalidates the experiment, so all the data are lost. Psychologically, this is an understandable reaction, for parapsychologists have often been subjected to such severe (even if often unfounded) methodological criticisms that they have become hypersensitive to departure from accepted experimental requirements.

We will argue that we need not necessarily reject data from experiments where a small degree of target non-randomicity is retrospectively discovered. We suggest conducting a powerful, empirical test of the degree to which this non-randomicity might allow a mathematical inference strategy by the percipients to account for results. If the results attainable by this test are considerably smaller in magnitude than those that were actually obtained in the experiment, and if the remaining data are significantly extra-chance, it is legitimate to argue that ESP was operating. We shall describe the algorithm we have developed for such a test, which we call the 'Probabilistic Predictor Program', and apply it to the results of an actual experiment.

MATHEMATICAL INFERENCE STRATEGIES

In an experiment where a percipient receives trial by trial feedback, and where there are small but statistically significant biases in the target sequence, how would he go about making use of these biases to raise his hit score? To simplify things, let us assume that the percipient never uses any kind of ESP, but only mathematical inference. Let the target sequence consist of the numerals zero to nine. We will postulate that our random generator contains small biases, such that one or more singlet targets have a probability slightly greater than 10%, and that some doublets, triplets, etc., may have biases resulting from singlet biases and/or independent biases in addition to those resulting from a reflection of singlet biases (note

2).

For his first call, our percipient's response must be a pure guess: he has no information yet about the target RNG to make any estimate of its biases from. Having made his guess response, he obtains feedback and finds out what his first target was.

To make his second call, his estimate of the biases in the target RNG must be based on a sample size of one, and we would not expect him to be able to make a very useful estimate of small target sequence biases from such a small sample. As he sees more and more of the target sequence, we can imagine him more or less accurately remembering whether some particular target singlet or singlets seem to appear very frequently, whether some particular target doublets seem to appear very frequently, etc.. By the time he has a sample of say, 100 targets, he can make a better estimate of possible biases in the target sequence than when he had a sample of one. By the time he has a sample of 1,000 targets, his estimates are even better, etc. (ignoring for the sake of simplicity, the possibility of forgetting and confusion in his memory). The larger the number of trials he has had feedback on, the more adequately he might be able to work out some sort of strategy where he might usually call, say, a five, because so far he has observed fives occurring with a frequency of two-tenths rather than one-tenth. He might supplement his singlet strategy by observing and deciding that threes were frequently followed by sevens, etc.. If his estimations of the biases in the target source are accurate he can, in principle, score above mean chance expectation.

Note that the potential success of such a mathematical inference strategy is related to the degree of bias in the target RNG and the accuracy of the percipient's perception of these biases. If there are very strong biases, such as a particular singlet target having a probability of one-half, and all the other possible singlets having much smaller corresponding probabilities, or if, for example, twos are always followed by ones, then our percipient should be able to begin scoring significantly above chance with a relatively small sample of the target sequence, say, after 25 or 50 trials. If, at the other extreme, the departures from an equiprobability and serial independence model in the target sequence are very small, say, one or two singlet targets having probabilities of .15 rather than .10, it would take a much longer sample of the target sequence to begin to get any accurate idea of these departures, and it is quite possible that forgetting and confusions in memory over very long samples of the

target sequence might effectively rule out any practical use of such a mathematical inference stategy for many percipients. We shall examine this in more detail below.

In attempting to estimate how much percipients in general might be able to actually use some sort of mathematical inference strategy, we get into a murky area. At one extreme, we find rare cases of mathematical or memory prodigies or idiot savants who can do very striking and unusual things with very complex numbers, which leads us to think that almost any kind of conceivable mathematical computing ability might be possessed by some person somewhere. Balanced against this extreme possibility is the fact that in almost all cases of people we actually come in contact with, they can seldom remember more than 10 digits or so accurately. The limit of the 7-digit telephone numbers, which are difficult for quite a number of people to remember, is a practical example of this. Thus in practice we would expect most percipients to have considerable difficulty keeping very accurate track of even the singlet distribution of the target sequence, much less the doublet, triplet, and higher order distributions, after more than a few dozen or so trials. Actual experimental data on what people can judge about large masses of serial, numerical data would, of course, be useful in estimating human capacities more precisely.

INADEQUACY OF CLASSICAL BIAS TESTS

Before considering our predictor program, it is important to note that standard tests for non-randomicity of sequences do not necessarily indicate how reliably a mathematical inference strategy could provide significant results in an ordinary length experiment. This will be illustrated by the following examples. We will use only singlet level bias, but the same arguments hold for higher level biases.

Suppose a percipient works in an experiment involving guessing the numbers one to ten, and he makes a total of 200 responses. We shall deliberately use a biased target source, such that outputs 1, 2, 3, 4, and 5, all have a probability of .15, while the outputs 6, 7, 8, 9, and 10 have a probabilty of .05, instead of all targets being equiprobable. The observed distribution of targets in the first half of our experiment (100 trials) would look like this, deliberately giving it a perfect reflection of the target bias pattern for

simplicity of illustration:

TARGETS	1	2	3	4	5	6	7	8	9	10
FREQUENCY	15	15	15	15	15	5	5	5	5	5

The standard Chi-square test for equal frequency of observed targets would tell us that this sample of 100 is not from a random source, as we get a Chi-square of 25.00 with 9 degrees of freedom, $p<.01$, one-tailed.

Suppose our percipient takes the first 100 trials to catch on to this singlet bias pattern, so that while he has only scored the 10 hits expected under our assumed equiprobability model in the first 100 trials, he will use a mathematical inference strategy, based on his new knowledge, for the remainig 100 trials of the experiment. His best strategy is to guess a 1, 2, 3, 4, or 5 on every trial. It does not matter whether he picks one of the high five and always guesses it, or randomly alternates among the high five: we would expect him to score about 15 hits in the second 100 trials. If we assess the significance of this score under our assumption of equiprobability, we compute a CR of 1.67, $p<.05$, one-tailed. For the whole experiment of 200 trials, we now have (10+15)=25 hits with an associated CR of 1.18, which, while not reaching statistical significance, might suggest to an experimenter that something was happening.

If the experiment was longer than 200 trials and the bias pattern and mathematical inference strategy were consistent, the percipient could obviously attain convential levels of significance as he went further. For 300 trials, for example, we would have (10+15+15)=40 hits, for a CR of 1.92, $p<.05$, one-tailed. We shall stay with a 200 trial experiment for now, however, to illustrate certain points.

Now consider another target source with a quite different sort of bias, where we observe the following distribution of targets in the first 100 trials:

TARGETS	1	2	3	4	5	6	7	8	9	10
FREQUENCY	24	10	10	8	8	8	8	8	8	8

This generator is highly biased toward producing ones, with no other large biases. The Chi-square test for equiprobability of singlets for

this distribution gives a value of 22.40, $p<.02$, one-tailed. If we mistakenly assumed that the magnitude of the Chi-square values reflected the degree of predictability of these target sources for a mathematical inference strategy, we would think this target source was equally or slightly less predictable than the source in the previous example. We would be quite wrong.

Again assume that it takes our percipient the first 100 trials to catch on to the bias, so he only scores 10 hits in the first 100. Now he follows the optimal strategy of calling a one for every one of the remaining 100 trials, and scores about 24 hits! For the second 100 trials alone, this gives a CR under the equiprobability model of 4.67, $p<.0001$, one-tailed. For the whole experiment of 200 trials, we have $(10+24)=34$ hits, with a CR of 3.30, $p<.0005$, one-tailed.

For equal Chi-square values in tests of bias, two sequences may differ enormously in usefulness for a mathematical inference strategy. Further, there will be far more possible bias patterns of less usefulness for a mathematical inference strategy than there will be highly useful ones for a given Chi-square value: there are many ways to rearrange the bias pattern in our first example without giving it the single number peak bias pattern of our second example that is so useful in a mathematical inference strategy.

Consider a third example where the following frequencies of targets are observed in the first 100 trials:

TARGETS	1	2	3	4	5	6	7	8	9	10
FREQUENCY	18	10	9	9	9	9	9	9	9	9

The standard Chi-square test of equiprobability tells us that this is not a biased sequence, for Chi-square is only equal to 7.20, which would occur more than half the time by chance alone. Yet if our target source is really biased toward ones in this way, and our percipient decided to call all ones in the second 100 trials of the experiment, he could make 18 hits there, for a CR of 2.67, $p<.01$, one-tailed, and a total of $(10+18)=28$ hits for the whole experiment, CR=1.89, $p<.05$, one-tailed. It is especially interesting to note that the entire sequence of 200 targets for this percipient, with the bias continuing through the second 100 trials, still does not show any significant bias: Chi-square=14.40, $p<.10$, one-tailed.

We may conclude the following for standard Chi-square tests of bias. For long to infinite length experiments:

(1) lack of significance Chi-square values in bias tests probably indicates lack of significant predictability by a mathematical inference strategy; and

(2) the presence of significant Chi-squares in bias tests indicates some degree of predictability by a mathematical inference strategy, but the magnitude of Chi-square does not indicate the degree of predictability.

For short to moderate length experiments of the type frequently carried out, however:

(3) a significant Chi-square indication of bias does not necessarily indicate that a significant overall score can be obtained through a mathematical inference strategy;

(4) the magnitude of the obtained Chi-square is a very poor measure of the magnitude of results that can be obtained with a mathematical inference strategy; and

(5) a mathematical inference strategy may produce significant results from a biased source which does not appear to be significantly biased by Chi-square evaluation.

The shortcomings of standard Chi-square measures of bias in realistic length experiments further illustrate why more direct measures of predictability by mathematical inference need to be developed.

THE PROBABILISTIC PREDICTOR PROGRAM

Our interest has been in carrying out an extreme test of the hypothesis that human percipients might use mathematical inference strategies to capitalize on non-randomicities in a target sequence instead of using ESP, so we set out to develop a mathematical inference strategy which, as far as we can tell, would be far superior to anything we could expect from a typical or even an exceptional human being. Our strategy program, for example, (a) never forgets or get confused about what previous targets in a target sequence to date have been; (b) has precise and accurate ways of estimating such things as whether, e.g., five to seven occurrences of one particular doublet in the first 100 trials should carry more weight as to what to call than a 15% proportion of a specific singlet appearing; (c) never gets tired or becomes prone to error; and (d) is not intellectually or

emotionally attached to a previous successful strategy, but constantly updates itself against new information from the target sequence.

We cannot claim that the human mind could not have some unknown, fantastic computing powers that might be superior to our inference strategy, but we do claim that our procedure looks to be far more powerful than anything we could reasonably expect to find in a typical human being. Until a more powerful mathematical inference strategy is actually modeled and demonstrated to be superior to ours, the procedure we present below constitutes an extremely powerful test of the hypothesis that percipients might artifactually inflate their hit scores through a mathematical inference strategy.

We have named the mathematical estimator strategy we have developed the 'Probabilistic Predictor Program' (PPP), and have implemented it on the Control Data Corporation Model 6400 computer at the University of California Berkeley campus. Its basic mode of operation on each trial is to (a) consult storage registers about what targets have been on all trials previous to the current one; (b) compute the probabilities of the observed target frequencies to date given an assumed equiprobability and sequential independence model; (c) use the largest deviation from this model to estimate the most likely possible biases in the target RNG and respond accordingly; and (d) update all these steps with the feedback on target identity for that trial in preparation for the next trial. We shall illustrate the operation in detail for seven trials, using an exceptionally biased source which generates only the sequence

12341234123412341234123..... etc.

even though we assume an RNG which generates the digits zero through nine independently and with equal probability. Using such an exceptionally biased source for our example allows the features and power of the PPP to be demonstrated in a few trials.

Trial One

Since there is no data in the PPP storage registers about the target distribution, except the assumption that any of the target digits zero to nine are equally probable, the PPP consults an RNG program within the computer and produces a digit from zero to nine as its first response. It might, for example, make a response of five. It now consults the appropriate storage register to pull out the first digit

of the target sequence, the equivalent of immediate feedback to a human percipient in a comparable situation. In our example the first target was a one. The PPP scores its response against the target, in this case a miss, and it notes in its singlet storage register that the digit one has occurred once in a total of one trial.

Trial Two

In order to estimate possible deviations from the assumption of every target digit having an equal probability of .1 and serial independence, from this trial on, the PPP begins consulting its storage registers for information about the target sequence on previous trials. For the second trial, the only register with any data in it is the singlet register, and it tells the PPP that one has been the most frequent singlet to date. Because it will be necessary in later trials for making decisions, although it is not really necessary on this second trial, the PPP computes the exact cumulative, one-tailed binomial probability of the most frequently observed target to date (a one in this case) having appeared one time in one trial, given the assumption of a probability of .1 for any possible digit. In this case the exact binomial probability is .1, one-tailed. The general decision rule now followed by the PPP is to always use the most significant departure from the equiprobability and serial independence model as the estimator of the most likely target to come up next in a target RNG that is actually biased. The lowest probability observation, the maximum departure from the model among several comparisons, is the most important. We have only one figure this time, namely the probabilty of .1 for a one, so the PPP responds with a one as its call for the next target.

The PPP now consults the second digit of the target series for its immediate feedback. In our example this digit was a two. The PPP again scores a miss, stores the fact that a two has occurred once in the singlet storage register (which is already storing the fact that a one has occurred once), and also stores the fact that a two has followed a one in a total of one (doublet) trial in the doublet register.

Trial Three

The PPP again consults its storage registers. In its singlet register it finds that both a one and a two are tied for the highest frequency of occurrence, namely each at one occurrence in two trials. The exact binomial probability of any one specific number occurring one or more

times in two trials, when the assumed probability of any particular number occurring is .1, is .19, one-tailed. The PPP consults then the doublet register to see if there is any relevant information about doublet frequencies, given that a two has just occurred on the previous trial. While the doublet register indicates that, given a one, a two has followed in one (doublet) trial, it has no relevant information about previous occurrences given a two as the first pair of a doublet. The call decision will have to be based solely on information from the singlet register. Since both one and two are tied for frequency of occurrence (and probability of occurrence) to date as singlets, the PPP randomly chooses one of them; for our example, the number one is its response.

The PPP now consults the third digit of the target series for immediate feedback, which for our example is a three. Its response of one was a miss and is scored appropriately. The fact that a three has occurred is stored in the singlet register, the fact that a two has been followed by a three is stored in the doublet register, and the fact that a one, two has been followed by a three is stored in the triplet register. Our target and response sequence to date is as follows:

Targets:	1	2	3
PPP Responses:	5	1	1

Trial Four

On consulting the singlet register, the PPP finds that the target digits one, two, and three have occurred with equal frequency. The exact binomial probability of any particular digit occurring one or more times in three trials, given our assumptions, is .27, one-tailed. There are no relevant doublet registers and no relevant triplet registers. Since the digits one, two, or three all present equally improbable departures from the model, it randomly chooses one of them, a two in this case. The PPP then consults the fourth digit of the target series for its immediate feedback. The target was a four. A miss is scored; a four is stored as having occurred once in the singlet register; a three, four is stored as having occurred in the doublet register; a two, three, four is stored as having occurred in the triplet register; and a one, two, three, four is stored as having occurred in the quadruplet register. Results to date look as follows:

Targets:	1	2	3	4
PPP Responses:	5	1	1	2

Trial Five

The PPP consults the storage registers and finds that one, two, three, and four have all shown an equal frequency of occurrence, and computes that the exact binomial probability of any particular digit having occurred one or more times in four trials, given an assumed probability of .1, is .34, one-tailed. There are no relevant doublet, triplet, or quadruplet registers, so since the singlet targets one, two, three, and four are equiprobable a random choice is made, a four in this case. The fifth digit of the target series is consulted for feedback: it is a one in this case, and the PPP response of four is scored as a miss. The fact that a one has just occurred is stored in the singlet register and high level strings to date are stored in high level registers, as previously illustrated, to the quintuplet level now. Our targets and responses to date are as follows:

Targets:	1	2	3	4	1
PPP Responses:	5	1	1	2	4

Trial Six

The way more complex decisions are made by the PPP is illustrated as we reach the sixth trial. The PPP first consults the singlet storage registers, and finds that one particular singlet target has occurred more frequently than any others, namely a one has occurred twice in five trials. There are no ties for the most frequent singlet to date. Given our assumed probability of .1, the exact binomial probability of any particular digit having occurred two or more times in five trials is .08, one-tailed. There is now relevant information in the doublet storage register: given a one, a two has followed it on one occasion in a total of four (doublet) trials. Assuming that any particular doublet has a probability of .01 (.1 times .1, given our model of serial independence), then the exact binomial probability of a particular doublet having occurred one or more times in four trials is .05, one-tailed. There are no relevant triplet, quadruplet, or quintuplet indications.

The PPP computes that the probability of .05 is a greater departure from the assumed model than a probability of .08, and so uses the

doublet information to make its decision: a one has occurred which was followed by a two before, a one has occurred again, so a response of two is given. It now consults the target storage register for feedback as to the identity of the sixth target, which is a two in our example, given our repeating target sequence, and it scores this as a hit. It then records the occurrence of a two in the singlet register, of a one, two in the doublet register; of a four, one, two in the triplet register, of a three, four, one, two in the quadruplet register; of a two, three, four, one, two in the quintuplet register; and of a one, two, three, four, one, two in the sextuplet register. Our results to date are as follows:

Targets:	1	2	3	4	1	2
PPP Responses:	5	1	1	2	4	2
						HIT

While in principle one could go on constructing higher registers above the sextuplet level, this obviously involves considerable increases in computer space used and computational time, so we made a practical decision to not construct higher registers than the sextuplet one.

Trial Seven

The PPP consults its singlet register and finds that the target singlets one and two are tied for highest frequency of occurrence. The exact binomial probability of any particular digit occurring two or more times in six trials, given our assumed model, is .11, one-tailed. Consulting the doublet register, given that a two has been the previous target, a two has been followed by a three once before. The probability of one particular doublet occurring one or more times in six (doublet) trials is .06, one-tailed. There is also a relevant triplet, namely given a one, two, a three has occurred before. The exact binomial probability of any particular triplet occurring one or more times in five (triplet) trials is .01, one-tailed. There are no other relevant storage registers for this trial.

Since the p value of .01 represents the greatest departure from the equiprobability and serial independence model and is thus our best estimate of possible biases in the target sequence, the PPP chooses to respond on the basis of the triplet register with a three. The extreme probability values that are quickly generated by doublet and higher order biases make the PPP capable on capitalizing on short term sequential biases, as well as long term biases. Consulting the target

register for feedback on the identity of the seventh target, we find that it is a three so a hit is stored, appropriate entries are made in the appropriate registers, etc.. Our results to date are as follows:

Targets:	1	2	3	4	1	2	3
PPP Responses:	5	1	1	2	4	2	3
						HIT	HIT

There is no need to continue to illustrate the functioning of the PPP at this point, as all the basic principles have been outlined. Notice that for this highly biased target sequence of a repetition of the series 123412341234..... the PPP has started hitting by the sixth trial and will make a perfect hit score from this point on.

APPLYING THE PPP TO EMPIRICAL DATA

Two kinds of sources for the generation of sequences to be tested by the PPP have been used to date, both drawn from an experiment (Tart, 1975; 1976a; 1977c), the first Training Study (TS), designed to assess the effects of immediate feedback on GESP performance (note 4). One source of sequences was the target sequences used in that experiment, which were generated by an electronic RNG; the other was the response sequences of the 10 percipients in that experiment. These 10 percipients were preselected for ESP talent from a much larger pool.

The electronic RNG was of the 'electronic roulette wheel' type, where an oscillator operating at a frequency of approximately five megaherz drives a zero-to-nine counter over and over again. The length of time that the experimenter holds down a pushbutton determines how long the oscillator is connected to the counter, and thus the final output that the counter stops at. Since the speed of the oscillator/counter combination is at least four orders of magnitude faster than human reaction time, as well as much faster than random variations in pushbutton pushing performance we would expect (from factors like variations in neural firing times and muscular reaction patterns) that the output of this type of generator would be random, with each output equiprobable and the outputs serially independent. A more complete description of the RNG is given in the original publications (Tart, 1975; 1976), although the circuit diagram there is incorrectly drawn: anyone wishing to build this RNG should contact C.T.T..

The zero-to-nine output of the RNG was displayed om a seven-segment Litronix Data Lite 10, which was then read by an experimenter and copied down on a data sheet. Occasional transcription errors might occur, although we would expect them to be rare for such a straightforward task. Responses of the percipients were also recorded manually on the data sheets by the experimenters, so occasional transcription errors were also possible here. We do not believe any significant number occurred, however, for the number of hits in a run of 25 trials was recorded automatically on a electro-mechanical counter, and the hit total of this counter had to be equal to the number of hits recorded by hand on the data sheet.

We would expect the response sequences of the percipients to be far more biased than anything generated by an electronic RNG, especially in terms of sequential biases, so they will constitute an especially interesting test sequence for the PPP.

The experimental design called for each of the ten percipients to make exactly 500 responses (in 20 runs of 25 trials each), but the number of usable targets for PPP analysis of target and response sequences varied slightly from this nominal 500 in a number of cases. On a few occasions the written record of a particular trial or response was ambiguous, so it was deleted. Also, the percipients sometimes used the Pass option (note 5) on the Ten-Choice Trainer and did not make a response to a particular target, so we would have a target digit but no response digit. Although the percipients did not receive feedback on Passes, we conservatively allowed our PPP to use the data obtained from Pass trials as part of its estimation strategy, thus giving the PPP a small advantage (the larger the N, the better the estimate of biases) that the percipients did not have.

RESULTS

In this analysis of PPP performance on target sequences, as well as the following one on guess sequences, we present only the results when the PPP was operated with a memory span of six, that is, it based its decisions on the singlet, doublet, triplet, quadruplet, quintuplet, and sextuplet levels, as this was the most powerful operation of the program in principle. We shall compare performance with shorter length memory spans than the sextuplet later.

The performance of the PPP on the target sequence of the ten percipients is shown in table 1, along with the degree of singlet and doublet bias in the target distributions, as assessed by the magnitude of the appropriate Chi-square tests for equal frequency and serial independence. The target source sequences are ordered from the highest to the lowest level.
As table 1 shows, the target sequences of two percipients, P5 and P3, showed statistically significant singlet bias at the level. The primary source of this bias, by inspection of the raw data, was a lack of XX doublet, i.e., a lack of 11s, 22s, 33s, etc.. Only one of the ten target sequences (for P3) is significant for general doublet bias if the XX doublet lack contribution is deleted. Retrospective questioning revealed that this XX bias was due to experimenter error. Pushing the button to activate the RNG did not have any kind of obvious mechanical snap to the switch, the switch just got more resistant to pushing toward the end of its travel. When an experimenter pushed the switch, let it out, and saw that the target was the same as on the previous trial, an XX doublet, he sometimes thought that he had not pushed the switch in far enough to activate the RNG, so he pushed it again, thus leading to a systematic depletion of the XX doublets. Triplet bias was not tested for, as an N of 500 is far too low to meet the assumptions for a valid Chi-square test.

Although doublet bias that is independent of singlet bias is presented in table 1, it is important to note that for the PPP any high doublet bias is potentially useful for prediction, regardless of whether it is significant independently of singlet bias or only a reflection of it.

The fourth column of table 1 is the mean number of hits per run of the PPP, when the expected chance value is 2.50. Although N (given in the last column) was nominally 500, it was slightly high or low in each case due to occasional extra trials being run by an experimenter, incomplete data, or passes by the percipient, so the means presented in the fourth column were adjusted to a standard run length of 25 trials to facilitate comparison. The fifth column presents the Z-score for hitting of the PPP program. As can be seen, the PPP was able to score significantly above chance on three of the target sequences, but not on the other seven. Only one of these three was a distribution that showed significant singlet or doublet bias by the standard Chi-square test.

Table 2 presents the performance of the PPP on the response

TABLE 1
Performance of Probabilistic Predictor Program
on target sequences

Sequence Source	Singlet Chi-square	Doublet Chi-square	Mean Hit per Run+	Z	N++
P5	38.81	99.08	3.32	2.50*	495
P3	35.53*	137.70*	2.55	.15	510
P4	17.04	97.37	2.95	1.40	525
P2	16.34	107.84*	2.78	.84	513
P14	15.08	104.70*	2.95	1.30	500
P1	13.47	87.81	3.30	2.46*	530
P32	12.12	76.23	3.05	1.70*	508
P17	11.83	101.43	2.30	-.60	501
P11	5.13	75.66	2.48	-.30	501
P7	2.64	93.32	2.48	-.10	528
			MCE=2.50		

* $p<.05$, one-tailed.
** Doublet tests, given the observed singlet distribution.
\+ These means have been adjusted to a standard run length of 25 trials.
++ N was nominally 500, but is slightly higher or lower in each case due to incomplete data or Passes by percipients. The mathematical transformations which base means and Ns on 25 trial runs and 500 trial series facilitate comparison, but make it difficult to retrieve raw data from this and subsequent tables. Researchers needing raw data should contact C.T.T..

sequences of the percipients. We expect human beings to show much more biased response patterns than an electronic RNG, and this is certainly the case.
As in table 1, the response sequences are ordered from highest to lowest singlet bias magnitude. The PPP was able to score as high as

TABLE 2
Performance of Probabilistic Predictor Program
on response sequences

Sequence Source	Singlet Chi-square	Doublet** Chi-square	Mean Hit per Run+	Z	N++
P5	146.47*	122.42*	4.90	7.23*	510
P11	105.53*	129.23*	3.89	4.15*	501
P4	79.53*	120.68*	4.26	5.37*	522
P17	71.00*	89.29	4.20	5.07*	500
P1	65.47*	94.80	3.44	2.87*	523
P32	48.39*	131.19*	2.81	.91	499
P14	40.88*	119.03*	2.80	.89	500
P3	39.35*	131.35*	3.28	2.34*	511
P7	28.88*	108.67*	2.90	1.19	500
P2	15.00	90.27	2.77	.81	515
			MCE=2.50		

* $p<.05$, one-tailed
** Doublet tests given the observed singlet distribution.
\+ These means have been adjusted to a standard run length of 25 trials.
++ N was nominally 500, but is higher or lower in each case due to incomplete data or extra Pass trials.

4.90 hits per run, against a mean chance expectation of 2.50, for a corresponding Z of 7.23. The PPP could significantly predict response patterns for six of the ten percipients.

We would expect that the degree of hitting by the PPP would be roughly proportional to the degree of bias by the Chi-square measures, but only roughly, for the reasons discussed earlier. For the target sequences alone, r=+.40 for the Z of PPP performance versus singlet bias, which is not statistically significant for an N of ten. For the response sequences, the corresponding r=+.89, which is significant at

the .001, one-tailed level for an N of ten. The correlation is low in the target sequences, for example, as a quite unbiased (by Chi-square) sequence for P1 turns out to be significantly predictable. Similarly, one of the less biased response sequences, that of P3, turns out to be significantly predictable, but the correlation is not so attenuated for the response sequences as a whole because of the wider range of bias in these sequences.

The imperfect correlation of PPP performance with Chi-square measures of degree of non-randomicity is to be expected. Chi-square measures reflect the overall degree of deviation of the observed target distributions from a model of equiprobability and sequential independence. Much of the kind of deviation from the model that would contribute to raising the Chi-square, however, would not be of any use in an inference strategy aimed at making hits. Much of the significance of a Chi-square test of singlet distribution, e.g., might come from the fact that half of the possible targets were too infrequent, given the model, but in making predictions you want to know what is the more frequent singlet, so you can give it as a response. At the doublet level, the lack of the XX doublets observed in the target sequences contributes greatly toward making the Chi-square test significant, but knowing not to call X when X has just appeared is much less useful information than knowing, as an example, that if X has appeared you ought to call Y because the observed frequency of Y following X in previous trials has been much higher than expected. The target distribution for P1 is a good illustration of this: neither the singlet nor doublet Chi-square measures of lack of randomicity approaches statistical significance, but the PPP managed to significantly predict this sequence.

The PPP was designed to be sensitive not only to relatively consistent biases to a target sequence at various levels, but also to pick up short-lived, but potentially important, higher order biases. The compilation and consultation of doublet through sextuplet registers is particularly important in creating this sensitivity of the higher order and transient biases such as might occur from temporary malfunctioning of the RNG. To empirically examine the actual usefulness of potential higher order biases in this empirical data, we modified the PPP so it could be run separately with memory lengths that included only the singlet level, the singlet plus doublet levels, etc., from the singlet through the sextuplet levels. Table 3 presents the performance of the PPP at each of these six memory levels, tabulated separately for target and response sequences. The number in

the body of the table is the Z-score of the PPP performance.

TABLE 3
Effects of various memory lengths on the performance of the Probabilistic Predictor Program

Sequence Source	Target Sequences Memory Length						Response Sequences Memory Length					
	1	2	3	4	5	6	1	2	3	4	5	6
P1	2.5	2.5	2.5	2.5	2.3	2.5	2.6	2.7	2.9	2.9	2.9	2.9
P2	.1	.3	.3	.8	.8	.8	1.1	.5	.4	.7	.7	.8
P3	-.1	-.3	.1	.1	.1	.1	1.6	2.3	2.5	2.3	2.5	2.3
P4	1.1	1.1	1.2	1.2	1.4	1.4	4.2	5.2	5.4	5.4	5.4	5.4
P5	2.2	2.2	2.3	2.5	2.5	2.5	6.9	7.5	7.2	7.2	7.5	7.2
P7	11.0	-1.3	-.6	-.1	-.3	-.1	.3	1.2	.6	.9	1.2	1.2
P11	.6	.1	.4	-.3	-.3	-.3	3.0	3.3	3.7	4.2	4.2	4.2
P14	1.6	1.3	1.6	1.6	1.3	1.3	1.6	1.1	.9	.9	.9	.9
P17	-.8	-.9	-.9	-.8	-.6	-.6	5.2	4.8	4.9	4.9	5.1	5.1
P32	1.4	1.8	2.1	2.0	1.7	1.7	1.2	1.2	1.2	.9	.9	.9
Mean CRs	.76	.68	.90	.95	.89	.93	2.77	2.98	2.97	3.03	3.13	3.09

Note: Figures in body of table are standard Z-scores

As can be seen, there was considerable variation in how much success the higher level registers added across various sequences. Consider the three target sequences (for P1, P5, and P32) where the PPP was able to score significantly: in two of these three cases the PPP had essentially gained its significance at the singlet level with very little added at higher levels. For some of the target and response sequences, the addition of higher level decision making actually resulted in some decrease in performance, as, for example, in the target sequence from P11. One effect of adding higher decision making levels, which carry higher weight, is that erroneous, chance fluctuations that suggest higher level biases can keep the PPP from

responding to a singlet bias if that is the only bias that is actually in the sequence. That is, the power of PPP which makes it sensitive to real higher order biases also makes it somewhat sensitive to being thrown off by artifactual indication of biases of this sort, just as human beings might be.

The differences between the various memory lengths are small, and generally not well suited to formal statistical analyses, but we present them primarily for their suggestive value. They do suggest that as far as the target sequences of this experiment are concerned, most of the predictive power that might be used in an estimation strategy is available at the singlet level, with only a slight amount added by higher level biases. For the ten target sequences in general, there is actually a non-significant (by t-test for related pairs) drop when the doublet memory length is added in, and a rise by the time the triplet and higher levels are used. The gain from the singlet to the triplet level is suggestive ($p<.10$, one-tailed), but in general, differences between adjacent levels are non-significant. Even when higher level registers are consulted, inspection of the PPP output shows that almost all of the hits arise from singlet level decisions. We interpret these results as suggesting that high level biases are of little consequence in this particular data. While it is always conceivable that a very high level bias (greater than the sextuplet level) might exist, we regard this as quite unlikely, for not only can we not envision a mechanism whereby an electronic RNG of this type would generate such a bias, but we also think it would be quite likely that if such a high level bias existed, components of it might very well show up as lower levels (sextuplet level or less) bias that would be detected and capitalized on by the PPP.

DID PERCIPIENTS IN THE FIRST TRAINING STUDY INFLATE THEIR SCORES WITH A MATHEMATICAL INFERENCE STRATEGY?

The question we shall now deal with is the degree to which percipients in the first TS might have used mathematical inference strategies to increase their hit scores. Three hypotheses will be examined: (1) no mathematical inference strategies were used, all the above chance-scoring was due to ESP; or (2) all of the above chance-scoring was due to (unconscious) mathematical inference strategies, with no need to invoke ESP; or (3) percipients may have used both ESP and some degree of mathematical inference strategies to

increase their hit scores.

How do scores of percipients and the PPP compare?

Table 4 presents the mean hit/run and associated Z-scores that the percipients and the PPP (memory span of six) (note 6) made on the target sequences of the feedback experiment. The PPP had more targets to work with than the percipients, for reasons discussed earlier, giving it a small advantage and so making our comparison more rigorous. Overall, the PPP made 577 hits in 5,136 trials for a Z of 2.95 and an associated p=.003, two-tailed. In terms of means, the PPP averaged 2.82 hits per run of 25, compared to the value of 2.50 per run expected by chance. The percipients, on the other hand, made a total of 722 hits in 5,000 trials for a Z of 10.46 and an associated p=2x10(E-25), two-tailed. In terms of means, the percipients averaged 3.61 hits per run instead of the chance-expected 2.50.
As far as peak results per target sequence are concerned, the PPP's best Z-score was 2.50, but five of the ten percipients had Z's of 4.17 and higher, with the highest at 11.03. Obviously the percipients performed enormously better than the PPP.

The relationship between percipient performance and PPP performance is illustrated in figure 1. As can be seen, the two performances are often highly divergent. The best percipient, for example, had a Z of 11.03, while the PPP could only score at the quite chance Z of .15 on her target sequence. The only case where the percipients's and PPP's scores are essentially equal was a case of chance level scoring for both. Overall, the Z's of the percipients and PPPs were uncorrelated (r=+.08, n.s.).

The enormous differences in performance magnitude of the actual percipients and the PPP strongly argue against hypothesis two, that all of the percipients' scoring was due to a mathematical inference strategy. A further argument against this hypothesis is the complete lack of correlation between the magnitudes of PPP and actual percipient performance.

Hypothesis three was that the results might be due to a mixture of ESP and mathematical inference strategies. Insofar as any other mathematical inference strategy we have been able to conceive to date would have been even less efficient than the PPP, however, this leaves

TABLE 4
Performance of real percipients versus
the Probabilistic Predictor Program

Sequence Source+	PPP Mean hits per run	PPP Z	Percipients Mean hits per run	Percipients Z
P5	3.32	2.50*	5.15	7.90*
P3	2.55	.15	6.20	11.03*
P4	2.95	1.40	4.05	4.62*
P2	2.78	.84	4.00	4.47*
P14	2.95	1.30	2.85	1.04
P1	3.30	2.46*	3.90	4.17*
P32	3.05	1.70	2.35	-.45
P17	2.30	-.60	2.95	1.34
P11	2.48	-.30	1.95	-1.64
P7	2.48	-.10	2.70	.60
Means	2.82		3.61	
Mean chance expectation	2.50		2.50	

\+ listed in order of decreasing singlet bias
* $p<.05$, one-tailed

even more than at least two-thirds of the results attributable to ESP. Suppose we consider, nevertheless, that results were due to this mixture; we can then use the PPP results as an empirical estimate of the true probability of a hit by ESP after mathematical inference results are subtracted out, and recalculate the significance of the ESP results. The results are only trivially different from the original results.

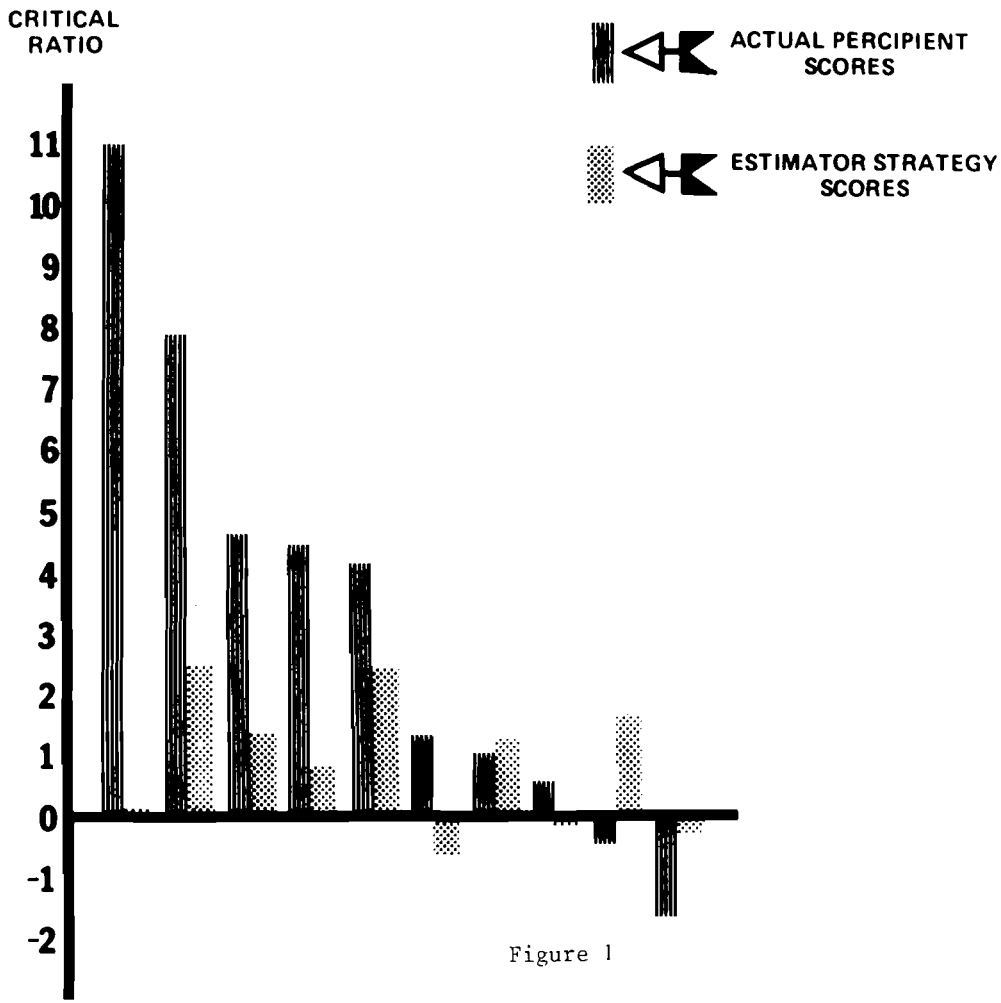
CRITICAL RATIO
ACTUAL PERCIPIENT SCORES
ESTIMATOR STRATEGY SCORES
11
10
9
8
7
6
5
4
3
2
1
0
-1
-2

Figure 1

CONCLUSIONS

Given our discussion of the ways in which conventional statistical measures of bias fail to adequately deal with the question of predictability of a target sequence, and our demonstration that the PPP deals much more adequately with this issue, we propose that the PPP (or some superior version of it) be adopted as the standard measure of predictability, and be used in all psi experiments in which percipients receive feedback on target identity before the end of the experiment. To make this more feasible, one of us (C.T.T.) is working on developing a simplified version of the PPP that will run on an Apple II microcomputer, since that computer is apparently being adopted as a standard in parapsychological laboratories.

For the first TS whose data have been examined here, the analyses strongly support the hypothesis that ESP is the best way to account for the hitting, with mathematical inference strategies playing little or no significant part in scoring. This conclusion is, of course, specific to the particular PPP strategy used here. While it may be possible for someone to develop a more powerful mathematical inference strategy, it is not clear to us how this could be done, and we present our results as a challenge to others to develop and empirically demonstrate a superior strategy.

We believe our results are important for parapsychological research in general, as well as having application to general psychological studies of human decision making. For example, there may be some parapsychological studies with useful results that have not been submitted for publication because of a retrospective discovery of some degree of non-randomicity in their target sequences. We suggest that rather than an all-or-none reaction to non-randomicity, experimenters with such data reevaluate them along lines similar to those used here (note 7). Useful data may be forthcoming.

ABSTRACT

With the increasing use of immediate feedback of target identity in parapsychological research, the question of departures from randomicity (equiprobability and serial independence) in target generators becomes important, as it is possible that some percipients

might identify such departures and develop a mathematical inference strategy for predicting targets, thus artifactually inflating their scores. The key aspect of randomicity of relevance is not lack of pattern per se, but the predictability of the generator. It is shown that standard Chi-square tests of randomicity are poor measures of predictability in short to moderate length experiments. A direct approach to the predictability of a possibly biased target source has been developed, the 'Probabilistic Predictor Program' (PPP), which is probably much more powerful than most human percipients could be. The operation of the PPP is described in detail. The PPP is then applied to both the target and response data of Tart's first Training Study, where some small departures from randomicity were found in the electronically generated target sequences and, of course, in the percipient generated response sequences. The PPP was found to occasionally score significantly on the target sequence, but far less successfully than the actual percipients did. The more biased response sequences were predicted quite significantly by the PPP. Examination of the internal displacement scoring patterns of the PPP was also compared with the patterns of the actual percipients and found to be drastically different. For these two reasons, it was concluded that use of mathematical inference strategies of the PPP sort could have only accounted for a trivial portion of the extremely high target scoring of the percipients in the first Training Study. While we should normally strive for completely random target sequences, the PPP is offered as a powerful approach to the question of predictability when departures from randomicity do occur, and can be of use in working with other experimental data.

NOTES

1. We may further subdivide the second category of feedback after calls into situations with closed decks and those with open decks. The closed deck situation has recently been discussed by Thouless (1977). We will not consider it further here, but confine our attention to the open deck situation.

2. If an RNG has significant singlet bias, this will automatically affect doublet, triplet, etc., distributions. A singlet excess of 7s, for example, will produce excesses of 71s, 72s, 73s, etc.. A Chi-square test for equiprobability of all observed doublets, where

the expected frequencies in each cell are based on the square of the assumed, unbiased singlet hit probability, will usually give significant results, as a result of such singlet bias. An appropriate Chi-square test for doublet bias that is independent of singlet bias would use expected frequencies in each cell that were calculated from the marginal totals.

3. This section draws heavily on material published in the January 1978 'Journal of the American Society for Psychical Research' (pp. 50-53) and is reproduced here by permission of the Journal.

4. The first Training Study has generated considerable critical comment (Gardner, 1977; Gatlin, 1978a; 1978b; 1979; O'Brien, 1976; Stanford, 1977a; 1977b) and response (Tart, 1976b; 1977a; 1977b; 1978; 1979) concerning a variety of issues. This paper deals only with the question of mathematical inference strategies as hypotheses for explaining the results of that study, and the reader should see the above references for other considerations.

5. Percipients did not receive feedback as to target identity on these Pass trials, and Passes were not counted in the formal scoring of hits and misses in evaluating the number of total hits obtained. Another minor difference was that this version of the PPP broke ties by choosing the lowest of the tied numbers, rather than randomly selecting among them, but this should not have any important effects.

6. The output of the PPP with a memory span of six is used here since this was theoretically expected to be the most sensitive to higher order biases, but, as an inspection of table 3 will show, the pattern of results reported here would have been the same if any other memory span results had been used.

7. We regret we cannot offer the use of our PPP program to other researchers, but our present version is quite expensive to run. The principles of operation, described above, should allow implementation of the program on other computers by skilled programmers.

ACKNOWLEDGMENT

We wish to thank Laura Dale, Aaron Goldman, J. Gaither Pratt, and Gertrude Schmeidler for helpful suggestions in preparing the final version of this paper.

REFERENCES

Chaitan, C. 'Randomness and mathematical proof', Scientific American, 1975, 232, 47-52.

Gardner, M. 'ESP at random', New York Review of Books, August 14, 1977.

Gatlin, L. 'Correspondence: Comments on the critical exchange between Drs. Stanford and Tart', J. A.S.P.R., 1978, 72, 77-81. (a)

Gatlin, L. 'Correspondence: Reply to Dr. Tart', J. A.S.P.R., 1978, 72, 294-296. (b)

Gatlin, L. 'A new measure of bias in finite sequences with application to ESP data', J. A.S.P.R., 1979, 73, 29-43.

O'Brien, D. 'Review of application of learning theory to ESP performance by C.T. Tart', J.o.P., 1976, 40, 76-81.

Stanford, R. 'The application of learning theory to ESP performance: a review of Dr. Tart's monograph', J. A.S.P.R., 1977, 71, 55-80. (a)

Stanford, R. 'The question is: Good experimentation or not?' J. A.S.P.R., 1977, 71, 191-200. (b)

Tart, C.T. 'The application of learning theory to ESP performance', New York, Parapsychology Foundation, 1975.

Tart, C.T. 'Learning to use extrasensory perception', Chicago, University of Chicago Press, 1976. (a)

Tart, C.T. 'Correspondence: Reply to O'Brien's review of Application of learning theory to ESP performance', J.o.P., 1976, 40, 240-246. (b)

Tart, C.T. 'Toward humanistic experimentation in parapsychology: A reply to Dr. Stanford', J. A.S.P.R., 1977, 71, 81-102. (a)

Tart, C.T. 'Psi and science', New York Review of Books, 1977, October 13. (b)

Tart, C.T. 'Toward conscious control of psi through immediate feedback training: Some considerations of internal processes', J. A.S.P.R., 1977, 71, 375-408. (c)

Tart, C.T. 'Correspondence: Reply to Dr. Gatlin', J. A.S.P.R., 1978, 72, 81-87.

Tart, C.T. 'Randomicity, predictability, and mathematical inference strategies in ESP feedback training experiments', J. A.S.P.R., 1979, 73, 44-60.

Thouless, R. 'The effect of information given to the subject in card guessing experiments', J. A.S.P.R., 1977, 49, 429-433.

Charles T. Tart
Department of Psychology
University of California at Davis
Davis, California 95616
U.S.A.

Eugene Dronek
University of California at Berkeley
Berkeley, California 94720
U.S.A.