

RESEARCH ARTICLE

Tricking the Trickster: Evidence for Predicted Sequential Structure in a 19-Year Online Psi Experiment

DEAN RADIN

Institute of Noetic Sciences, Petaluma, California, USA

Submitted January 3, 2019; Accepted August 19, 2019; Published December 30, 2019

DOI: <https://doi.org/10.31275/2019/1429>

Copyright: Creative Commons CC-BY-NC

Abstract—From August 2000 to December 2018, two online psi experiments based on a five-target, forced-choice protocol together collected 114 million trials from an estimated 200,000 people around the world. The hit rate combined across both experiments was consistent with a null effect. However, a planned secondary analysis, designed to detect a predicted sequential pattern in the data, resulted in a small magnitude but statistically unambiguous outcome. With a chance expected rate of $p_o = 0.32$, the combined observed $p_i = 0.320502 \pm 0.000044$, $z = 11.28$, $p = 1.7 \times 10^{-29}$. Control tests found no evidence that this small deviation, independently observed in two separate experiments, was due to human behavior such as optional stopping or response biases, nor to computational or randomization errors that might have provided participants with subtle cues. If analysis of other forced-choice psi tests replicates this effect, it would suggest the existence of an unconscious process that tends to obscure accuracy in psychic perception. Gaining a better understanding of that process might lead to more robust results in psi experiments.

Introduction

One of the most puzzling aspects of psi is its apparently capricious nature (Beloff, 1994; Hansen, 2001; Kennedy 2003). This refers to the oft-reported difficulty of repeating successful pilot studies in formal replications, or, worse, finding that positive effects obtained in pilot studies significantly reverse in followup attempts. These fickle effects are sometimes half-seriously attributed to the “trickster,” i.e. a mischievous mythological character found across cultures and throughout history (Radin, 1969).

Some propose that the evasive nature of psi is an inherent aspect of the phenomenon, dashing the hopes of experimentalists who hope to develop

easily replicated psi effects (von Lucadou, 1995). But there is another possibility: If one imagines that a “trickster” is responsible for hiding psi effects, then with clues about how the trickster operates—assuming some regularity rather than completely unpredictable methods—we may be able to trick the trickster and reveal what was hidden. The present paper explores this theme in a simple psi task.

One of the earliest and simplest designs for a psi experiment is the forced-choice task, such as the ESP card test popularized by J. B. Rhine in the 1930s. Meta-analysis of 145 reports of ESP card tests published from 1882 to 1939 indicates that those studies produced a small but significant and repeatable effect (Pratt & Rhine, 1967). Selective reporting was only beginning to be recognized as a problem during that early era, so it is difficult today to provide an accurate estimate of the effect size obtained in those studies. A more recent meta-analysis of 309 forced-choice studies conducted after Rhine’s heyday, published from 1935 to 1987, again showed a small but non-chance repeatable effect (Honorton & Ferrari 1989). A meta-analysis of 72 even more recent forced-choice tests, published from 1987 to 2010, confirms that the forced-choice design continues to be a simple but effective way to study psi effects (Storm, Tressoldi, & Di Risio 2010). In sum, hundreds of forced-choice psi experiments indicate that the technique works, but it is also highly inefficient because the effect size is minuscule.

Experimental protocols yielding very small effect sizes require substantial statistical power to reliably detect signal from noise. Historically, the three primary ways researchers have sought to achieve large sample sizes have involved long-term, single-lab efforts (Jahn, 1982), tests held over the radio or in magazines (Brier, 1969; Rhine, 1962), or combining studies with meta-analysis. A fourth approach, increasingly popular since the rise of the Internet, has been use of online psi experiments.

Online Experiments

While helping to solve the statistical power problem, publicly accessible online experiments are not immune to their own design challenges. Of particular relevance to the issue of statistical power, “big data” collected under unsupervised conditions can easily amplify tiny human and computational biases. In addition, because data collected in the real world never exactly conform to the theoretical null hypothesis, if enough data are collected it is possible, at least in principle, to obtain a *p*-value as small as one wishes (Kaplan, Chambers, & Glasgow, 2014, Sullivan & Feinn, 2012). As Cohen (Cohen, 1990, p. 1306) put it,

A little thought reveals a fact widely understood among statisticians: The null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), is always false in the real world. It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then, a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection.

Another problem with unsupervised online experiments is optional stopping, which can occur when a participant receives trial-by-trial feedback and is performing poorly. Online attention is often measured in seconds, so those who become dissatisfied with their ongoing test score are likely to quit before the pre-defined run- or session-length. Others who perform well may be motivated to continue to the end of the planned session or beyond. Such biases affect the interpretation of experimental results, depending on which portions of the data are examined. Such biases can be avoided by including tasks with no feedback, such as implicit or hidden tasks.

Origin of the Present Approach

The purpose of the present analysis was to study an unusual hit pattern predicted to arise in forced-choice psi experiments. The predicted pattern was investigated in data generated in online tests written by the author starting in August 2000. As of December 2018, two of those tests, both using similar forced-choice designs, had accumulated 114 million trials contributed by more than 200,000 individuals around the world. For the first 17 years of these experiments, the analyses discussed here were long forgotten and were not part of the original analytical plan. Then, in August 2017, while reviewing some old files, the author ran across an unpublished paper describing an experiment he conducted decades earlier with a colleague (Radin & Cross, 1990). The present study was motivated by re-reading that paper and realizing that the effect found in that earlier experiment could be tested with the present set of “big data.”

The pattern of interest was proposed by an empirically oriented group of Christian Scientists, known as Spindrift, who were active from the 1970s to 1990s (Sweet, 2007). The Spindrift researchers proposed that the mind has three components, all simultaneously engaged in modulating psi performance. They called these an *ordering force*, a *perceptive ability*, and a *defense mechanism*. The ordering force was described as an ego-less, “they will be done” state of consciousness. This was said to induce equilibrium, coherence, negentropy, or similar concepts associated with production of order, into any labile system capable of expressing degrees of order.

Spindrift's concept of a perceptive ability was what parapsychologists call psi, and their notion of a defense mechanism was an unconscious mental effect that served to hide psi effects.

Spindrift researchers tested their ideas in many ways, often reporting highly positive effects. One such experiment, which they called "visual image, unconscious response," was a simple binary ESP card test. The test involved a deck consisting of 12 copies each of two images. To begin the test, a participant would select one image that they liked from a pool of images, and then a second image they did not like. Cards containing copies of these images were placed into separate opaque envelopes, shuffled, and then the participant guessed what they thought each envelope contained. Later, the participant or an independent experimenter would record the resulting sequence of hits and misses, and the test was repeated.

To evaluate the results, it was assumed that perceptual psi would accurately perceive the images, but that unconscious defense mechanisms would mask that accuracy by intentionally causing the number of hits and misses to be about the same. It was also assumed that attention focused on the task would evoke the ordering effect and cause the hits and misses to be distributed in an orderly way. A sequence of 12 hits in a row followed by 12 misses would represent one form of order, but that might be too obvious, and as such it would defeat the purported masking action of the unconscious defense mechanism. Another type of order, one hidden from casual inspection, would be to alternate hits and misses within each response type (this is explained in more detail below).

In 1990, the author and a colleague attempted to replicate the Spindrift protocol and method of analysis (Radin & Cross, 1990). We invited eight individuals to guess images in a 22-card binary deck and to repeat that task five times. Given the binary design, the chance hit rate for the direct task was $p_o = 0.50$, and our final direct hit rate was non-significantly below that. However, application of Spindrift's sequential analysis resulted in $p_t = 0.525$, which was close to what Spindrift had reported in their tests. A 2.5% effect over chance is not especially impressive, but after being amplified with more than 10,000 trials, as was common in Spindrift's experiments, they were able to report consistently large positive z scores.

In the analyses reported here, the sequential hit rate effect was much smaller than 2.5%. But it was also based on a five-choice task, which requires a more complex scheme for "hiding the results" as compared to a binary task. In addition, the data were contributed online in an unsupervised context, so it is to be expected that the magnitude of the resulting effect size would be much smaller.

Most parapsychologists ignored Spindrift's experiments, partially due

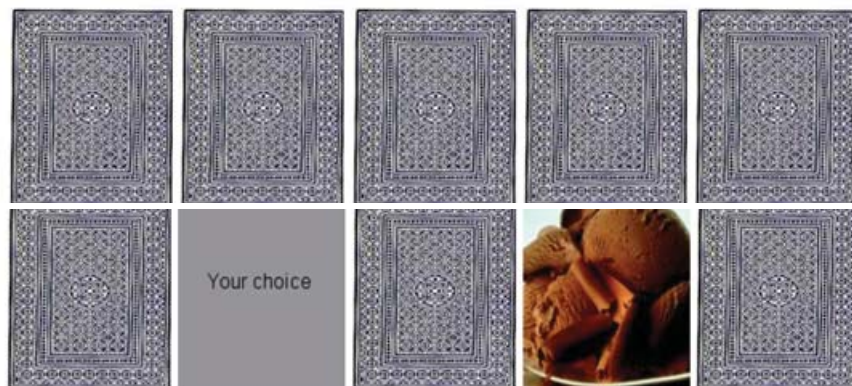


Figure 1. (Top) Five “cards” displayed in the browser. (Bottom) Participant’s choice indicated along with the randomly selected target. This trial would be recorded as a miss.

to suspicions about Spindrift’s religious affiliation, but also because the reported results seemed to be too good to be true. It was also problematic that their analytical methods were not described well and hence were difficult to understand. These discomforts contributed to an assumption that the results were possibly due to sloppy methods or one or more analytical artifacts. After we obtained results similar to what Spindrift had reported, we were less sure that such dismissals were valid. But we did continue to worry about analytical artifacts because, among other things, their test design used a closed deck, which complicates calculation of the chance-expected hit rate. Based on these concerns, we placed our study in the file drawer, where it sat patiently for nearly three decades. The large forced-choice database from our online psi experiments provided motivation to revisit the Spindrift hypothesis.

Method

Tasks

The two forced-choice tests are part of a suite of psi tests located at GotPsi.org. The first, referred to as the *Card* test, consists of five card images displayed on a web browser screen (Figure 1, top). A participant selects one of the cards, then the web server randomly selects and displays one card, along with the participant’s choice (Figure 1, bottom). A correct choice is recorded as a hit, otherwise it is a miss. The chance-expected hit rate is thus 1 in 5, or $p_o = 0.20$. The experiment was originally coded in Perl



Figure 2. Quick remote viewing test, showing the participant's choice below (the eagle) and the computer-selected target above. This trial would be recorded as a hit.

(<https://www.perl.org/about.html>), hosted on several different web servers over the years, and the Perl *rand* and *srand* functions were used to pseudo-randomly generate the targets. No pseudo-random generator (PRNG) can provide perfectly random sequences (Ferrenberg, Landau, & Wong, 1992), although to detect nonrandom patterns requires analysis of very long sequences, and in the present case only a few random numbers were used in each successive trial, and then the PRNG was reseeded with the current clock time, thus adding an effectively truly random element to the target generation process.

The second experiment, described as a “quick remote viewing” test, and referred to here as *QRV*, used a design similar to the card test except that instead of displaying cards, five photos were selected at random from a large pool of photos, and newly random images were used as targets in each successive trial (Figure 2). The task was the same as the card test, thus the chance-expected hit rate was again, $p_o = 0.20$.

Analyses

Data produced in these tests were stored as one line of information per trial per person, and all trials contributed per day were stored in a single file in chronological order. A “day” was defined according to local time in the time zone of the web server that hosted the tests. The information of interest for the present analysis was the response (R) and target (T) in each trial, along with the username of the individual who contributed the trial. Direct *hits*

R	T	H	A	B	C	D	E	F
1	1	X	X	}	0	}	1	}
2	3	0	X					
3	3	X	X	}	1	}	1	}
1	1	X	X					
2	2	X	0	}	1	}	0	}
2	3	0	0					
1	2	0	0	}	1	}	0	}
3	1	0	X					
1	1	X	X	}	1	}	1	}
2	3	0	0					
2	2	X	0	}	1	}	1	}
2	2	X	X					
1	3	0	0	}	1	}	1	}
2	1	0	0					
3	3	X					X	
		7/14	3/4		4/5		2/2	

Figure 3. Example of direct hit and sequential analyses for a 3-target, forced-choice task. See text for explanation.

referred to the number of trials where $R = T$, and the resulting *hit rate* was simply $p_i = hits/N$, where N was the number of trials contributed by that user. A z -test was used to find the p -value for a test of the null hypothesis, namely that the probability of a hit was $p_i = 0.2$ versus the alternative that p_i was not equal to 0.2, with $z = (p_i - p_o) / \sqrt{p_o q_o / N}$, where $p_o = 0.2$ and $q_o = 1 - p_o$. This analysis reflected performance on what might be called the “surface” or explicit task, and it is the conventional way of analyzing the hit rate in a forced-choice psi experiment.

The analysis of interest here was not about overall hits and misses, but a certain *sequence* of hits and misses. Figure 3 illustrates this analysis for one person’s hypothetical data in a three-target, forced-choice design. The illustration is intentionally simplified for the sake of expediency. Once the analytical process is understood, it is a simple matter to expand it to a five-target, forced-choice design.

The first three columns of Figure 3 show 14 trials, labeled response R,

target T, and hit H. In the H column, X refers to a hit and O to a miss. An ordinary hit/miss analysis with a null hypothesis (in this case $p_o = 1/3$) would compare p_o against the observed 7 hits in 14 trials, or $p_i = 1/2$. But what we are interested in is column A, which is formed by extracting only those trials where the participant's response was $R = 1$. There are 5 such cases in the example. The values in column A consist of either an X or an O, taken from the associated lines in H. Column C is where $R = 2$, and the contents of C are again either an X or O taken from column H. Likewise for column E where $R = 3$.

Now to form column B, we take overlapping pairs of values in column A. If a pair consists of XX or OO, it is assigned a 0. If a pair consists of XO or OX it is assigned a 1. In this way, based on the values in column B, we see a total of 4 values, of which 3 are 1s and one is 0. To create columns D and F we follow the same procedure. Now we count the total number of pairs and 1s. This example has a total of 11 pairs tested, of which 9 are 1s, for a hit rate of $p_i = 9/11$. This is the hit rate of interest in this analysis; we refer to it as hr_{seq} . It is a measure of the number of alternating pairs of hits and misses *with respect to each type of response*.

One may wonder why the sequence of hits and misses are measured within each response type, rather than the more straightforward sequence independent of the response. There is no obvious reason; it was simply postulated by the Spindrift researchers as a way that the unconscious might act to sabotage perceptual psi.

To determine how much p_i deviates from chance expectation, note that the probability of obtaining the paired-sequence [0 1] or [1 0] in a three-target test is $[(1 - p_o) \times p_o] + [p_o \times (1 - p_o)] = [0.67 \times 0.33] \times 2 = 0.44$. This is the case because the targets are randomly selected, so each sequential pair of hits or misses can be considered an independent event.

For the five-target test of interest in the present analysis, where $p_o = 0.20$, the expected hr_{seq} over the long term is $[0.8 \times 0.2] + [0.2 \times 0.8] = 0.32$. The appropriate statistical test is $z = (p_i - p_o) / \sqrt{p_o q_o / N}$ where $p_o = 0.32$ and N is the number of hit-pairs examined. In these tests, note that R corresponds to the position of the target on the computer screen, i.e. in Figure 1 the value 1 refers to the left-most target and $R = 5$ to the right-most target.

Results

Card Test

From August 2000 to December 2018, 87.4 million trials were contributed in the Card test over 6,472 days by 234,000 unique usernames (of which

an estimated 90% were different individuals; this estimate was based on a limited number of cases where it was known that the same individual had used multiple usernames). Of these trials, 83.9 million were contributed where the participant provided at least 2 trials in one session. A “session” refers to all trials contributed by a unique individual over the course of a day, and a minimum of two trials was required because the measure of interest involved pairs of trials.

Of the 83.9 million trials of interest, the direct hit rate was $hr = 0.199949 \pm 0.000050$, associated with $z = -1.19$, $p = 0.24$ (all p-values reported are two-tailed). By contrast, the sequential analysis resulted in $hr_{seq} = 0.320411 \pm 0.000051$, $z = 8.07$, $p = 6.9 \times 10^{-16}$.

We might note here that the Card test was not as simple as it appeared to be. In a typical, forced-choice test, the participant assumes that each card has the same likelihood of being selected for the future target. This was not the case in this experiment. On each successive trial the cards were assigned different probabilities of being selected. This feature was hidden from the user, so it could not bias their choice, but it may have been responsible for the overall negative hit rate, and it may also have suppressed the magnitude of the sequential effect. These possibilities will be explored in a future publication.

QRV Test

From April 2005 (when this test was first launched) to December 2018, a total of 26.8 million QRV trials were contributed on 4,769 days by more than 60,000 unique usernames (again, roughly 90% were estimated to be unique individuals). The direct hit rate was a modest $hr = 0.200169 \pm 0.000092$, $z = 2.188$, $p = 0.03$. By contrast the sequential analysis resulted in $hr_{seq} = 0.320825 \pm 0.000092$, associated with an impressive $z = 8.70$ ($p = 3.2 \times 10^{-18}$).

Combined Results

Because the Card and QRV tests both used the same five-target, forced-choice design, their databases could be combined. With a total of more than 114 million trials, the direct hit rate was $hr = 0.200001 \pm 0.000044$, $z = 0.025$, $p = 0.98$. Thus, from a direct-hit perspective the combined results were almost exactly as expected by chance. However, the combined $hr_{seq} = 0.320503 \pm 0.000044$, $z = 11.28$, $p = 1.7 \times 10^{-29}$. This unambiguously indicates either the presence of a genuine sequential pattern, or one or more artifacts. To explore the alternative possibilities, a variety of control tests were conducted.

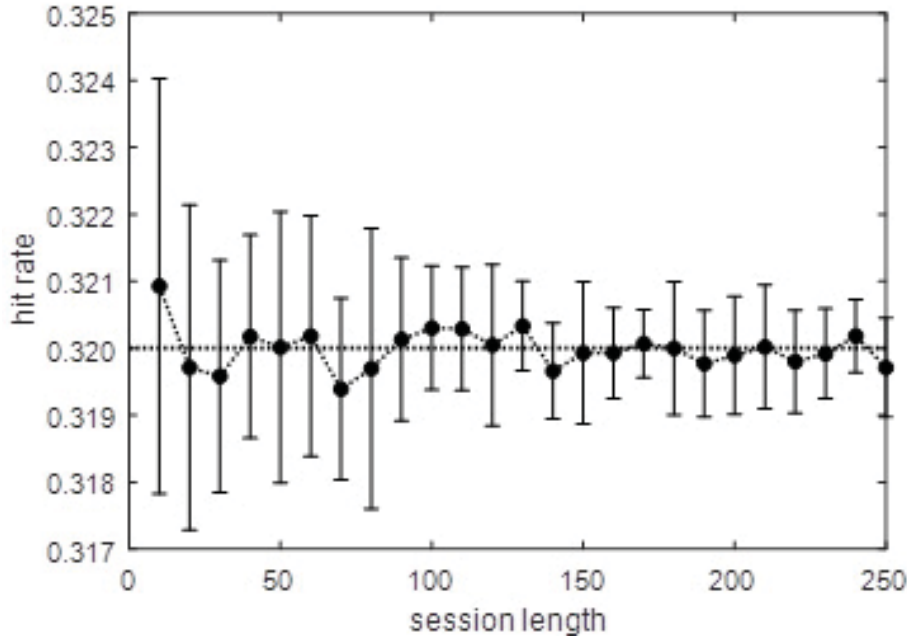


Figure 4. Means and 95% confidence limits of simulated five-target, forced-choice experiment with users contributing from 10 to 250 trials per user per session, with each session repeated 100,000 times. The calculated chance hit rate of 0.32 is shown as a dotted horizontal line.

Simulation

The first item to check was if the calculated chance sequential hit rate of $p_o = 0.32$ was correct. To do this, I programmed a simulated five-target, forced-choice experiment (in Matlab, 2018b) to generate 100,000 repetitions of 10 trials per session, and then repeated this process by incrementing by 10 trials up to 250 trials per session. The result, shown in Figure 4, with mean hit rates and 95% confidence intervals, indicates that the theoretically expected hit rate of 32% is indeed valid.

Target Frequency Distribution and Sequential Runs

To test if the targets were distributed uniformly at random, a chi-square test was performed on each day's distribution of targets, and then the distribution of resulting p -values was tested using a second chi-square test for uniform distribution across 10 bins. The result of the second chi-square test for the

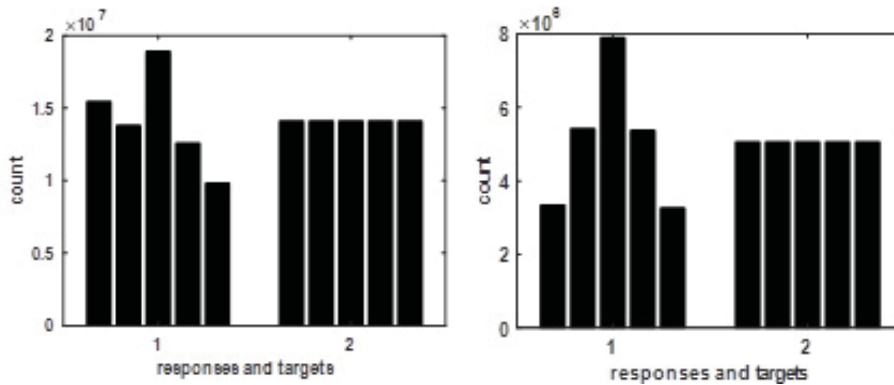


Figure 5. Distribution of responses and targets for the Card test (left) and QRV tests (right).

Card test was $p = 0.58$, and for the QRV test $p = 0.45$. To test for sequential randomness of the targets, a runs test was performed on each day's target data using the Matlab (R2017b) function, *runstest*. The p -values resulting from the runs tests were then tested for uniform distribution of p -values across 10 bins. The result for the Card test was $p = 0.97$, and for the QRV test $p = 0.94$. In sum, no obvious non-random structure was detected in the targets.

Response Biases

One might imagine that the sequential analysis result was biased by the non-random distribution of users' selection of the targets, i.e. by their responses. Figure 5 shows the distribution of responses and targets in the two databases, indicating as expected that the middle target was the most frequently favored choice. Figure 6 shows the results of the sequential analysis per response choice (i.e. $R = 1, 2, 3, \dots$) in the two databases. Out of 10 tests, we see that in 9 of 10 cases the results were substantially above chance expectation of $p_o = 0.32$. This consistency suggests that the sequential analysis result was not due to response biases.

Optional Stopping

Optional stopping behavior was clearly evident in each of the two databases (see Figure 7). As expected, participants tended to quit the test if they were performing poorly, but they continued to contribute trials up to the predefined

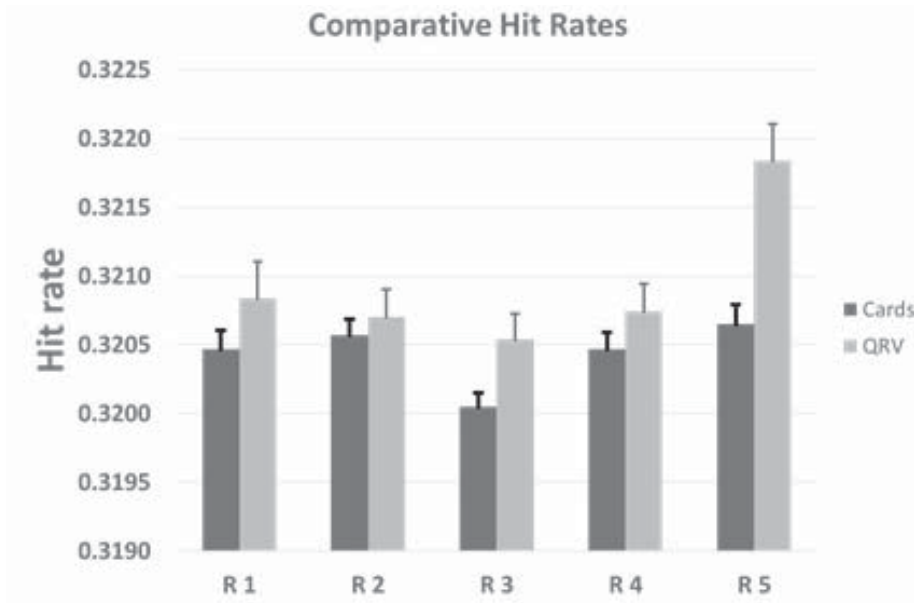


Figure 6. Mean sequential hit rates in the Card and QRV databases, with one standard-error bar; the chance-expected hit rate is 0.32. This result indicates that above-chance deviations were observed in 9 of 10 responses.

session length (typically 20 trials) when they were performing well. This was reflected in below-chance direct hit rates for users contributing fewer than 20 trials and above-chance hit rates for users who contributed exactly 20 trials. Could optional stopping have been responsible for the sequential hit rate? To study this question, further control tests were performed.

Note the positive correlation between the direct hit rate and hr_{seq} . This relationship is not surprising because when the direct hit rate is low compared to chance expectation, the imbalance between hits and misses permits fewer hit/miss alterations than would be expected by chance, and vice versa. Thus, given that the overall direct-hit rate for the Card test was negative, that might lead us to expect that the overall sequential hit rate would also be negative. But instead, as observed, it was highly positive (in statistical terms), suggesting that the sequential hit rate was not due to optional stopping.

Sequential Dependencies

In this test, the transitions from target T_n to target T_{n+1} were determined for all trials performed by each individual, and in the order that the trials were

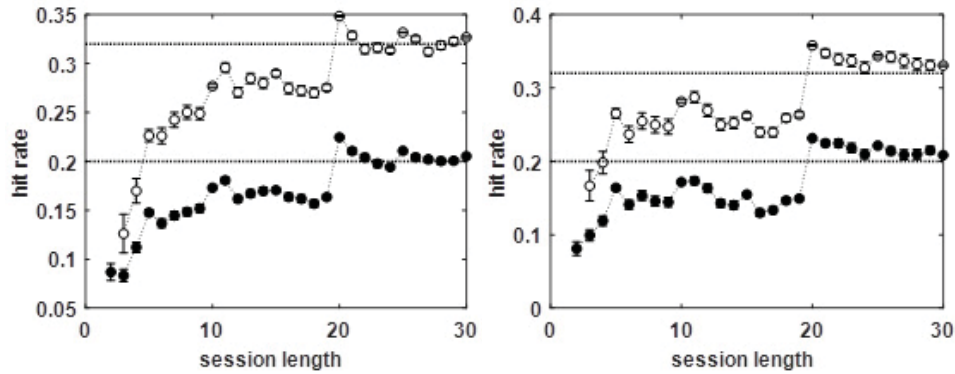


Figure 7. (Left) Card test direct hit rate (± 1 standard error, which is so small that the error bars are obscured by the black dots) and similarly for the sequential hit rate (white dots), for session lengths ranging from 2 to 30 trials. The negative hit rates prior to 20 trials are due to optional stopping behavior. Sessions of 20 trials were the most common predefined run-length and show a positive bias, also due to optional stopping. (Right) Same for QRV data.

contributed. The same transitions were also determined for the sequence of responses. A chi-square (χ^2) contingency test for uniform distribution was performed on these matrices. For the targets, the analysis resulted in $\chi^2 = 22.5$, $df = 16$, $p = 0.13$. For the responses, $\chi^2 = 10.9$ million, $p \approx 0$. The nonsignificant χ^2 for targets indicates that successive pairs of targets occurred in a random order, and the extremely large χ^2 for responses indicates, as expected, that people did not respond at random. Similar analyses, examining the transitions T_n to $T_{n+2,3,4}$, also resulted in nonsignificant effects: $\chi^2 = 15.7$ ($p = 0.47$), $\chi^2 = 9.7$ ($p = 0.88$), and $\chi^2 = 14.39$ ($p = 0.57$), respectively. Analyses of the responses all resulted in probabilities of essentially zero.

These analyses suggest that over the entire database there were no obvious dependencies in the target sequence that a participant might have exploited to produce an inflated hr_{seq} . But on a day-to-day basis, involving smaller numbers of trials, perhaps fluctuations in target dependencies did occur that provided clues. If that were the case, then perhaps daily variations in target sequence dependencies might have been correlated with the daily hr_{seq} .

To test this possibility, for each day's data we determined the χ^2 associated with target transitions T_n to T_{n+1} , as well as hr_{seq} , and then we examined the correlation between those two arrays over the total of 10,372 days of data

across the combined Card and QRV datasets. A positive correlation for a one-step dependency (i.e. r_{+1}), would suggest that deviant target sequences might have provided clues, resulting in higher hr_{seq} ; we refer to this idea as a “sequential clue hypothesis.”

No such relationship was found: $r_{+1} = 0.0099$, $p = 0.313$. The same analysis was then performed for dependencies T_n to $T_{n+2,3,4}$, resulting in $r_{+2} = -0.026$ ($p = 0.009$), $r_{+3} = -0.0040$ ($p = 0.681$), and $r_{+4} = 0.012$ ($p = 0.237$). The same correlations run for z scores associated with hr_{seq} , which took into account the different sample sizes obtained per day, resulted in $r_{+1} = 0.004$ ($p = 0.68$), $r_{+2} = -0.02$ ($p = 0.04$), $r_{+3} = -0.004$ ($p = 0.68$), $r_{+4} = 0.01$ ($p = 0.24$). Note that the r_{+2} correlation was significant in both tests, but *negative*, suggesting that as target dependencies were more deviant, hr_{seq} and their associated z scores were *lower*. This is opposite to the sequential clue hypothesis.

Following the Target

The above analysis found, as expected, that people do not respond at random, but it also found that those biases cannot account for the hr_{seq} result. This analysis examined in more detail why user responses are biased. Figure 8 shows that users tended to follow the targets, that is, R on trial N+1 tended to be the same as the randomly generated T on trial N. Could this following dependency have contributed to the inflated value of hr_{seq} ? The answer should be no, because no matter how the user responds, as long as the target sequences are adequately random, the user cannot generate an inflated score (barring genuine precognition).

To test this assumption, a simulation was implemented whereby each successive R was forced to be exactly the same as the previous T, i.e. a perfectly uniform following-the-target response bias. The simulation, run for 10 million trials, showed that $hr = 0.199874 \pm 0.000126$ ($z = -0.995$) and $hr_{seq} = 0.31979 \pm 0.000126$ ($z = -1.43$). In other words, even a highly exaggerated nonrandom response strategy that mimicked how people tended to respond did not produce inflated values for hr or for hr_{seq} .

Unconscious Learning

Perhaps subtle patterns were unconsciously noticed by participants who contributed many repeated trials. To explore this possibility, we determined the correlation between the number of trials contributed per person per day, versus the hr_{seq} calculated for that individual. A learning artifact predicts that this correlation should be positive.

To evaluate this hypothesis, we formed an array of trials per person per

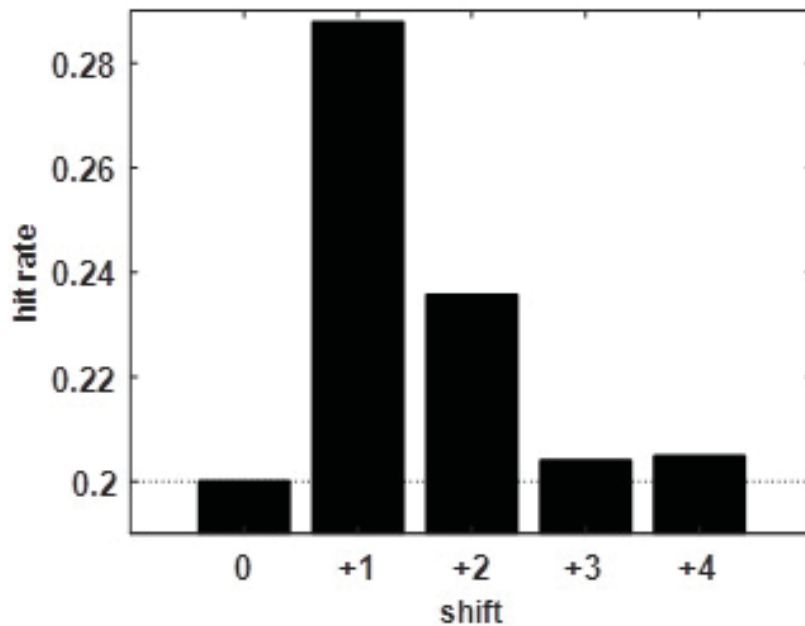


Figure 8. Direct hit rate for lags 0 to +4, indicating that people tended to respond on trial N_{+1} using the target that was presented on the previous trial, i.e. their responses followed the targets. This tendency was especially strong for N_{+1} and N_{+2} and to a limited extent for N_{+3} and N_{+4} .

day, another array of hr_{seq} per person per day, then calculated the correlation between those two arrays per day. This daily correlation was converted into a z score using a Fisher z transform, and this process was repeated for each day. Then the resulting mean z score was compared to 0 (the null hypothesis). The result was $\bar{z} = -0.0119$, $t = -2.53$, $p = 0.012$. This modestly significant *negative* outcome suggests that to a small extent the more trials an individual contributed, the *smaller* their resulting hr_{seq} . This is opposite to the prediction of a learning hypothesis.

Permutation Analysis

We return to the possibility that the sequential effect was an artifact of optional stopping. This is an important factor to consider because both optional stopping and hr_{seq} are positively correlated with the direct hit rate. Thus, to test if optional stopping may have been responsible for the observed hr_{seq} , we re-ran the original analysis but randomly permuted the

sequence of hits obtained by each person in each session. This maintained the same number of hits originally obtained per person, so the psychological motivations underlying optional stopping were also the same. Thus, if optional stopping were responsible for the inflated hr_{seq} results, then randomly permuting the hit sequence should not alter the original value of hr_{seq} .

The results showed that for the Card data, the randomly permuted hit analysis resulted in a nonsignificant $hr_{seq} = 0.319941 \pm 0.000051$, $z = -1.152$, $p = 0.249$; for the QRV data, again we obtained a nonsignificant $hr_{seq} = 0.320052 \pm 0.000092$, $z = 0.561$, $p = 0.57$. Overall the results were $hr_{seq} = 0.319967 \pm 0.000045$, $z = -0.737$, $p = 0.461$. In sum, optional stopping does not explain the observed sequential hit rate.

Another question that may arise is whether an alternating hit/miss analysis based on the original order of the trials, that is *without* regard to the type of response, also shows an anomalous positive bias. For the Card test, this analysis resulted in $p_1 = 0.320009 \pm 0.000050$, $z = 0.197$, $p = 0.84$. For the QRV test, $p_1 = 0.320302 \pm 0.000090$, $z = 3.336$, $p = 0.0009$. And combined, $p_1 = 0.320078 \pm 0.000044$, $z = 1.789$, $p = 0.074$. Thus, while the QRV test shows an intriguingly positive result, overall there is no evidence that an alternating hit/miss order was beyond chance (especially in light of the statistical power provided by 109 million trials). This again suggests that the sequential hit rate is indeed linked in some way with the participant's response on each trial, and it is not simply due to hits and misses alternating on successive trials.

Discussion

The sequential effect hr_{seq} does not appear to be due to nonrandomness in the frequency or sequence of targets, nonrandom transitions of targets from T_{i+1} to T_{i+4} , or optional stopping behavior. This suggests that hr_{seq} requires a very clever unconscious process, one that keeps track of the alternating sequence of hits and misses within each response type. That such a process might exist is not entirely unreasonable; the relevant literature has long suggested that psi effects bubble up out of the unconscious, and that they are modulated by a host of psychological filters and defense mechanisms (Carpenter, 2015; Carpenter, 1967, Crumbaugh, 1968; Eisenbud, 1983; Johnson & Haraldsson, 1984; Rogers & Carpenter, 1966). Analyzing psi data for sequential patterns is also not a new idea (Burdick & Kelly, 1977; Pratt & Soal, 1952; Timm, 1970; Zenker, Leslie, Port, & Kosloff, 1982) nor is the notion that psi effects are not simply difficult to detect but may be actively evasive (Kennedy, 2003).

What is new is a description of a way that a forced-choice psi test can result in a chance-expected direct hit rate, while at the same time producing a highly significant sequential hit rate. If the hr_{seq} effect had been discovered after extensive data snooping, it would still be somewhat surprising given the final $z > 11$. However, this effect was definitively not the result of data snooping. What was described here is the result of a *single, planned analysis* employed for the sole purpose of checking a prediction, and that prediction was significantly replicated to very high levels of confidence in two separate databases.

A Model

Imagine that the act of selecting a response R places a probabilistic bias on that response, such that the first time R is selected it “pulls” the subsequent target T so that it matches R. The next time the user selects R, that bias is reversed, “pushing” T away so it mismatches R. That is, on successive selections of R, the bias is reversed. This pull/push scheme suggests a tension between creation and destruction, an effect we can imagine as woven into the fabric of reality, or as a dynamic balance constantly attempting to sustain equilibrium between chaos and order. The tendency for psi effects to tweak probabilities of desired events, followed shortly afterwards by “anti-tweaks,” or by apparently influencing control series to produce results opposite to those observed in experimental series, has been noted in the parapsychological literature for many years (Cox, 1954; Jahn, Dunne, & Nelson, 1987; Palmer & Kramer, 1984; Radin, 1993; Schmeidler 1973).

To simulate this model, a five-item, forced-choice model was programmed with a hit bias of 20.00001% and a miss bias of 80.00001%. That is, the model simulated a user selecting response R, which resulted in a *hit* with a slightly greater probability than chance, and then after selecting the same R again, it resulted in a *miss* with a slightly greater probability than chance. After running a total of 10,000 repetitions of this scheme, with each repetition consisting of 1,000 trials, the resulting $hr = 0.199842$ (a *t*-test comparing this mean hit rate against the expected $hr = 0.20$ resulted in $t = -1.379$), and $hr_{seq} = 0.33013$ ($t = 57.178$). Thus, even an extremely small bias applied in a systematically alternating fashion (within each response type) can produce a null effect for the direct hit rate and a very significant positive sequential effect. It is not claimed that this model accurately reflects the unconscious processes actually used by participants in the online tests, only that it is possible to construct a rather simple model that mimics the observed results.

Conclusion

This study suggests two possibilities. The first is that the sequential pattern identified in this analysis is an illusion due to an undiscovered artifact. The various controls tests argued against this possibility, but with the statistical power afforded by more than 100 million trials, even tiny systematic artifacts could become visible above the noise. If further control tests do uncover such artifacts, they will be reported in a subsequent article.

The second possibility is that deeper analysis of simple psi experiments may reveal subtle but genuine patterns in data that previously were overlooked. Understanding those patterns, should they exist, may be useful in better understanding how unconscious strategies mask explicit psi effects, and that in turn may allow experimenters to figure out how to overcome difficulties in repeating psi effects. Finding such patterns may also reveal that the supposed capricious nature of psi may not be an intrinsic property that is impossible to overcome, but rather just a reminder of our ignorance about psi and the modulating effects of the human unconscious. As psi tests continue to generate increasingly larger datasets, we may eventually figure out how to detect subtle patterns that can reveal clever ways of out-tricking the trickster.

Acknowledgment

I am indebted to the late Richard Shoup, who co-founded the nonprofit Boundary Institute with me in 2000. I wrote several of the gotpsi.org psi tests while I was with the Boundary Institute, including the tests mentioned in this paper, and when I left the Boundary Institute in 2001 to join the staff of the Institute of Noetic Sciences, Shoup continued to develop and maintain that website for many years. When Shoup passed away in 2015, the Institute of Noetic Sciences adopted gotpsi.org and continues to administer and develop the site.

References

- Beloff, J. (1994). Lessons of history. *Journal of the American Society for Psychical Research*, 88, 7–22.
- Brier, R. M. (1969). A mass school test of precognition. *Journal of Parapsychology*, 33(2), 125–135.
- Burdick, D. S., & Kelly, E. F. (1977). Statistical methods in parapsychological research. In B. B. Wolman (Ed.), *Handbook of Parapsychology* (pp. 81–130). New York: Van Nostrand Reinhold.
- Carpenter, J. C. (1967). Decline of variability of ESP scoring across a period of effort. *Journal of Parapsychology*, 31(3), 179–191.
- Carpenter, J. C. (2015). *First Sight: ESP and Parapsychology in Everyday Life*. Lanham, MD: Rowman & Littlefield.

- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), 1304–1312. <https://doi.org/10.1037/0003-066X.45.12.1304>
- Cox, W. E. (1954). A comparison of spheres and cubes in placement PK tests. *Journal of Parapsychology*, 18(4), 234–239.
- Crumbaugh, J. C. (1968). Variance declines as indicators of a stimulator–suppressor mechanism in ESP. *Journal of the American Society for Psychical Research*, 62(4), 356–365.
- Eisenbud, J. (1983). *Parapsychology and the Unconscious*. Berkeley, CA: North Atlantic Books.
- Ferrenberg, A. M., Landau, D. P., & Wong, J. Y. (1992). Monte Carlo simulations: Hidden errors from “good” random number generators. *Physical Review Letters*, 69(23), 3382–3384.
- Hansen, G. P. (2001). *The Trickster and the paranormal* (1st ed.). Philadelphia, PA: Xlibris.
- Honorton, C., & Ferrari, D. C. (1989). “Future telling”: A meta-analysis of forced-choice precognition experiments, 1935–1987. *Journal of Parapsychology*, 53(4), 281–308.
- Jahn, R. G. (1982). The persistent paradox of psychic phenomena: An engineering perspective. *Proceedings of the IEEE*, 70(2), 136–170.
- Jahn, R. G., Dunne, B. J., & Nelson, R. D. (1987). Engineering anomalies research. *Journal of Scientific Exploration*, 1(1), 21–50.
- Johnson, M., & Haraldsson, H. E. (1984). The Defense Mechanism Test as a predictor of ESP scores. *Journal of Parapsychology*, 48(3), 185–200.
- Kaplan, R. M., Chambers, D. A., & Glasgow, R. E. (2014). Big data and large sample size: A cautionary note on the potential for bias. *Clinical and Translational Science*, 7(4), 342–346. <https://doi.org/10.1111/cts.12178>
- Kennedy, J. E. (2003). The capricious, actively evasive, unsustainable nature of psi: A summary and hypotheses. *Journal of Parapsychology*, 67(1), 53–74.
- Palmer, J., & Kramer, W. (1984). Internal state and temporal factors in psychokinesis. *Journal of Parapsychology*, 48(1), 1–25.
- Pratt, J. G., & Rhine, J. B. (1967). *Extra-sensory perception after sixty years; a critical appraisal of the research in extra-sensory perception*. Boston, MA: Bruce Humphries.
- Pratt, J. G., & Soal, S. G. (1952). Some relations between call sequence and ESP performance. *Journal of Parapsychology*, 16(3), 165–186.
- Radin, P. (1969). *The trickster, a study in American Indian mythology*. New York: Philosophical Library.
- Radin, D. I. (1993). Environmental modulation and statistical equilibrium in mind–matter interaction. *Subtle Energies & Energy Medicine Journal Archives*, 4(1), 1–30.
- Radin, D., & Cross, G. (1990). *Sequential ordering effects in a perceptual experiment: Testing the nature of prayer and psi*. GTE Laboratories.
- Rhine, J. B. (1962). The precognition of computer numbers in a public test. *Journal of Parapsychology*, 26(4), 244–251.
- Rogers, D. P., & Carpenter, J. C. (1966). The decline of variance of ESP scores within a testing session. *Journal of Parapsychology*, 30(3), 141–150.
- Schmeidler, G. R. (1973). PK effects upon continuously recorded temperature. *Journal of the American Society for Psychical Research*, 67(4), 325–340.
- Storm, L., Tressoldi, P. E., & Di Risio, L. (2010). Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin*, 136(4), 471–485. <https://doi.org/10.1037/a0019457>
- Sullivan, G. M., & Feinn, R. (2012). Using effect size—why the p value is not enough. *Journal of Graduate Medical Education*, 4(3), 279–282. <https://doi.org/10.4300/JGME-D-12-00156.1>
- Sweet, B. (2007). *A Journey Into prayer: Pioneers of prayer in the laboratory: Agents of science or Satan?* United States: Xlibris.
- Timm, U. (1970). The statistical analysis of hit sequences in quantitative ESP experiments. *Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie*, 12(3), 163–185.

- von Lucadou, W. (1995). The model of pragmatic information (MPI). *European Journal of Parapsychology, 11*, 58–75.
- Zenker, S., Leslie, R. C., Port, E., & Kosloff, J. (1982). The sequence of outcomes and ESP: More evidence from a primacy effect. *Personality and Social Psychology Bulletin, 8*(2), 233–238.