

# TRICKING THE TRICKSTER: DETECTING HIDDEN STRUCTURE IN DATA FROM AN 18-YEAR ONLINE PSI EXPERIMENT

Dean Radin

*Institute of Noetic Sciences, CA, USA*

## ABSTRACT

From August 2000 through October 2017, two online psi experiments based on a five-target, forced-choice protocol collected over 100 million trials from an estimated 200,000 individuals around the world. The direct hit rate combined across both experiments was consistent with a null effect; where  $p_o = 0.20$ ,  $p_I = 0.19996 \pm 0.00004$ ,  $z = -0.94$ ,  $p = 0.35$  (two-tail). A planned secondary analysis, designed to detect a subtle but predicted pattern in the data, resulted in a significant deviation; where  $p_o = 0.32$ ,  $p_I = 0.32051 \pm 0.00005$ ,  $z = 10.6$ ,  $p < 10^{-25}$ . Control tests found no evidence that this small magnitude but highly significant positive deviation was due to optional stopping, response biases, target sequence dependencies, learning of subtle cues, or other potential artifacts.

## INTRODUCTION

One of the most puzzling aspects of psi is its apparently capricious nature (Beloff, 1994; Hanson, 2001; Kennedy, 2003). This refers to the oft-reported difficulty of repeating highly successful pilot studies in formal replications, or worse, finding that strong effects in one study significantly reverse in follow-up attempts. These fickle “trickster” effects are sometimes named after mischievous mythological characters found in many cultures. The trickster is Pan and Loki in Greek and Norse myths (Campbell, 2008), Coyote in Native American folklore (Radin, 1956), and Murphy’s Law in modern technological contexts (Bloch, 1978).<sup>1</sup>

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 (Kennedy, 2003; von Lucadou, 2015). 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 we may be able to trick the trickster and reveal what was hidden. The present paper explores this theme in a simple psi task.

### *Forced-Choice Tasks*

One of the earliest designs for a psi experiment is the forced-choice task, such as the ESP card test popularized by J. B Rhine in the 1930s (Pratt et al, 1940). Meta-analysis of 145 reports of ESP card tests published from 1882 to 1939 suggests that those studies produced a small but overall significant and repeatable effect (Bösch, 2004), but because selective reporting was only beginning to be recognized as a

---

<sup>1</sup> Murphy’s Law: If anything can go wrong, it will, and at the worst possible time.

problem during that early era, it is difficult to provide an accurate estimate of the effect size obtained in those studies.

A meta-analysis of 309 forced-choice studies after Rhine's heyday, published from 1935 to 1987, again showed a small, repeatable, and significant effect size ( $es$  (per study) = 0.02,  $z = 6.02$ ,  $p = 1.10 \times 10^{-9}$ , Honorton and Ferrari, 1989). A meta-analysis of 72 more recent forced-choice tests, published from 1987 to 2010, confirmed that the forced-choice design continues to be a simple and effective way to study psi effects ( $es$  (per study) = 0.01,  $z = 4.86$ ,  $p = 5.90 \times 10^{-7}$ , Storm, Tressoldi, Di Risio, 2012). In sum, over 500 published forced-choice psi experiments indicate that the technique works, but it is also highly inefficient because the effect size is so small.

### *Statistical Power*

Experimental protocols that yield very small effect sizes require substantial statistical power to reliably detect effects that deviate from chance expectation. Historically, the two ways to achieve large sample sizes have either involved long-term, single-lab efforts (e.g., Jahn et al, 1997), or by combining studies with meta-analysis. Then, since the rise of the Internet, a third approach has become increasingly popular: online psi experiments.

While helping to solving the statistical power problem, publicly accessible online experiments are not immune to a host of 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 theory, to obtain a p-value as small as one wishes (Sullivan & Feinn, 2012; Kaplan, Chambers & Glasgow, 2014). As Cohen (1990) 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. (Emphasis in the original, p. 1306.)

Another problem with unsupervised online experiments is optional stopping, which occurs when a participant receives trial-by-trial feedback and is performing poorly. Online attention is typically measured in seconds, so those who become dissatisfied with their ongoing scores are likely to quit the experiment before the pre-defined run- or session-length. Others who perform well may be motivated to continue to the end of the planned session. These biases can substantially 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.

### *Goal*

The purpose of the present analysis was to study a subtle pattern predicted to arise in the data of forced-choice psi experiments. This was tested in data from two online psi tests that were launched online by the author (DR) in August 2000. As of October 2017, these two tests had together accumulated over 100 million trials, contributed by an estimated 200,000 individuals. For the first 18 years of these experiments, neither DR or the other investigator had planned to use these data to search for the effect described here.<sup>2</sup> Then, in August 2017, while reviewing some old files, DR ran across an unpublished paper that described an experiment he had conducted decades earlier (Radin & Cross, 1990), which was designed to explore the

---

<sup>2</sup> I am indebted to the late Dr. Richard Shoup for refining and adding to the GotPsi.org suite of tests.

effect described here. The present study was thus sparked by re-reading that paper and realizing that the predicted effect could be tested with “big data.”

## METHODS

### Tasks

The two forced-choice tests are part of a suite of psi tests located at [GotPsi.org](http://GotPsi.org). The first, referred to as the *Card* test, consists of five card images displayed on the 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 coded in Perl,<sup>3</sup> hosted on several different web servers over the years, and the Perl *rand* and *srand* functions were used to randomly generate the targets.

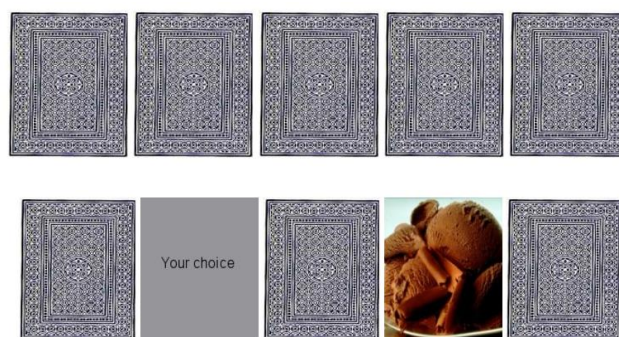


Figure 1. (Left) Five “cards” displayed in the browser. (Right) Participant’s choice indicated along with the randomly selected target. This trial would be recorded as a miss.

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 selected at random from a large pool of photos 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  $p_o = 0.20$ .

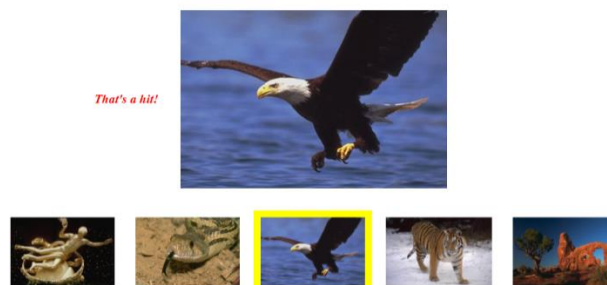


Figure 2. Quick remote viewing test, showing the participant’s choice below (in a yellow outline) and the randomly selected target above. This trial would be recorded as a hit.

<sup>3</sup> <https://www.perl.org/about.html>

## Analyses

Data produced in these tests consisted of one line of information per trial per person, and all trials contributed per day were stored in a single file in chronological order.<sup>4</sup> The information of interest was the response (R) and target (T) in each trial, along with the user name of the individual who contributed the trial. Direct *hits* referred to the number of trials where  $R = T$ , and the resulting *hit rate* was simply  $p_I = \text{hits}/N$ , 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  $z = (p_I - p_o)/\sqrt{p_o q_o/N}$ , and where  $p_o = 0.2$  and  $q_o = 1 - p_o$ . This analysis reflected performance on the “surface” or explicit task, and it is the conventional way of analyzing the hit rate in a forced-choice psi experiment.

The second analysis, which is of principal interest here, examined the sequence of hits. It must be emphasized that the sequence of interest *is not based on the usual meaning of a series of hits and misses*. Reviewers of early drafts of this paper assumed that they understood what “sequence of hits” meant, and that led them to mistaken interpretations about the results of this analysis. Figure 3 will be used to illustrate the nature of this analysis for one person’s data in a simple, three-target, forced-choice design. That simplified explanation will then be expanded to the actual five-target forced-choice design.

The first three columns of Figure 3 show 14 trials, with response R, target T, and hit H; in the H column X refers to a hit and O to a miss. A conventional hit/miss analysis with the null hypothesis  $p_o = 1/3$  would compare  $p_o$  against the observed 7 hits in 14 trials, or  $p_I = 1/2$ . This is the usual analysis for a forced-choice test, but it is not what we are interested in here.

What *is of interest* is column A, which is formed by extracting those trials where the response  $R = 1$ ; there are 5 such cases in the example. The values in 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 create 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. To create columns D and F, we follow the same procedure. Now we count the total number of pairs and 1s. This example has 11 pairs and 9 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 response type*.

---

<sup>4</sup> A “day” was defined according to local time in the time zone of the server hosting the tests.

| R    | T | H | A   | B | C   | D | E   | F |
|------|---|---|-----|---|-----|---|-----|---|
| 1    | 1 | X | X   |   | 0   |   |     |   |
| 2    | 3 | 0 |     |   |     |   |     |   |
| 3    | 3 | X |     |   |     |   |     |   |
| 1    | 1 | X | X   |   |     |   | X   |   |
| 2    | 2 | X |     |   |     |   |     |   |
| 2    | 3 | 0 |     |   |     |   |     |   |
| 1    | 2 | 0 | 0   |   |     |   |     |   |
| 3    | 1 | 0 |     |   |     |   | 0   |   |
| 1    | 1 | X | X   |   |     |   |     |   |
| 2    | 3 | 0 |     |   |     |   |     |   |
| 2    | 2 | X |     |   |     |   |     |   |
| 1    | 3 | 0 | 0   |   |     |   |     |   |
| 2    | 1 | 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.

To determine how much  $p_1$  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 is 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 run is  $[0.8 \times 0.2] + [0.2 \times 0.8] = 0.32$ . The appropriate statistical test is  $z = (p_1 - p_o) / \sqrt{p_o q_o / N}$ , where  $p_o = 0.32$  and  $N$  is the number of hit-pairs examined. In these tests,  $R$  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 October 2017, a total of 85.9 million trials were contributed in the Card test on 6,013 days by 234,105 unique usernames (of which an estimated 90% were different individuals). Among all trials, 76.9 million were contributed under conditions where the participant's responses were distributed between two or more of the five target possibilities. That is, these trials excluded sessions where a participant repeatedly selected just one target, where a "session" refers to all trials contributed by a unique individual over the course of a day.

Some of the excluded sessions were due to a webbot programmed by the author to assess the Card test (discussed later). In other cases, they may have been due to participants trying to mentally "will" the computer to conform to their intention rather than predict the target, and still other cases may have been due to webbots programmed by hackers in an attempt to gain control of the web server. In any case, the analysis of interest assumed that the participant was actively deciding which target to select. Thus, data from the excluded sessions are not considered here (but are considered later).

Of the 76.9 million trials of interest, the direct hit rate was  $hr = 0.199892 \pm 0.000046$ , which is associated with  $z = -2.35$ ,  $p = 0.02$  (two-tailed). While significantly below chance in conventional terms, this small magnitude psi-missing effect is not especially remarkable given the available statistical power. By contrast, the hit rate for the sequential analysis was  $hr_{seq} = 0.320400 \pm 0.000054$ . This too is a small deviation from chance expectation, but it is associated with a deviation of 7.4 standard errors above the null (associated with  $p = 2 \times 10^{-13}$ ).

### *QRV Test*

From April 2005 through September 2017, a total of 24 million QRV trials were contributed on 4,311 days by 62,856 unique usernames (again, roughly 90% were likely to be unique individuals). The direct hit rate was  $hr = 0.200183 \pm 0.000082$ , associated with  $z = 2.23$ ,  $p = 0.03$ . Again, while “significant” this deviation is unremarkable, but by contrast the hit rate for the sequential analysis was  $hr_{seq} = 0.320832 \pm 0.000098$ , associated with  $z = 8.49$  ( $p = 2.1 \times 10^{-17}$ ).

### *Combined Results*

Because the Card and QRV tests both used the same five-target, forced-choice design, their databases can be combined. With a total of nearly 101 million trials, the direct hit rate was  $hr = 0.19996 \pm 0.00004$ ,  $z = -0.94$ ,  $p = 0.35$ . A conventional direct hit psi test would end at this point with a resounding failure. However, the combined  $hr_{seq} = 0.32050 \pm 0.00005$ ,  $z = 10.6$ ,  $p < 10^{-25}$ , based on over 96 million paired-trials. Even with the substantial power afforded by millions of trials, a 10 *sigma* (i.e. standard error) deviation is most unlikely to be attributable to chance. This indicates either the presence of a genuine sequential pattern or one or more artifacts or mistaken assumptions. To explore the latter possibilities, a variety of control tests were conducted.

### *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 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  $p = 0.94$ . These tests detected no obvious non-random structures in the targets. But perhaps there were less obvious biases.

### *Response Biases*

Could the results of the sequential analysis have been 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.  $R = 1, 2, 3, \dots$ ) across the two databases. Out of 10 tests, we see that in 9 of 10 cases the results were substantially above chance-expectation of  $p_o = 0.32$ . The consistency across responses suggests that the sequential analysis results were not influenced by response biases.

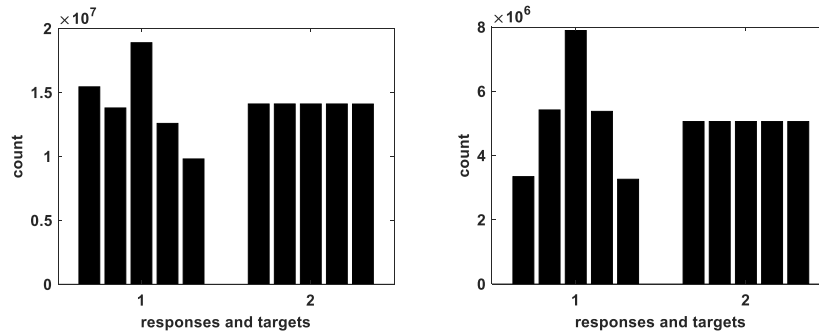


Figure 5. Distribution of responses and targets for the (left) Card test and (right) QRV tests.

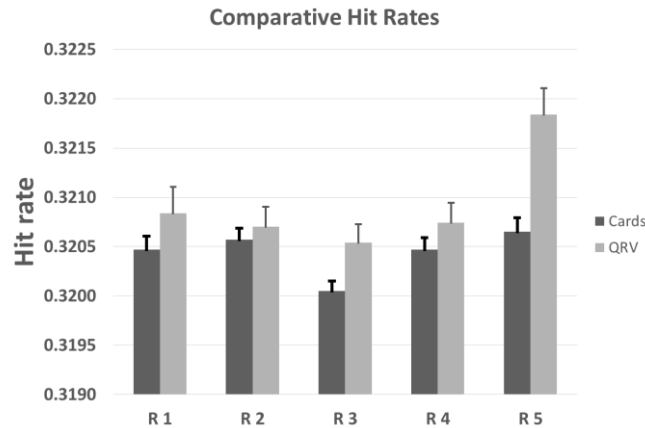


Figure 6. Mean sequential hit rates in the Card and QRV databases, with one standard-error bars; the chance-expected hit rate is 0.32. This result indicates that above-chance deviations were observed – with the exception of the middle card in the Card test – regardless of the user's responses.

### Optional Stopping

Optional stopping was clearly evident in each of the two databases (see Figure 7). If users quit the test when they were performing poorly, but continued to contribute trials up to the predefined session length (typically 20 trials) when they were performing well, then we should expect to see below-chance direct hit rates (black circles in Figure 7) for users who contributed fewer than 20 trials and above-chance hit rates for users who contributed exactly 20 trials. And this is what we see. This optional-stopping behavior raises the question of whether this effect might have been responsible for the large deviation observed in  $hr_{seq}$ . To study this question, further control tests were performed.

It is important to note the *positive correlation* between the direct hit rate and  $hr_{seq}$ . This relationship should not be surprising because in cases where the direct hit rate is low as compared to chance expectation, there will be too many misses. That in turn will lead to fewer hit/miss alternations than expected by chance, and vice versa. Thus, given that the overall direct hit rate was slightly negative ( $hr = 0.19996$ ), this would lead us to expect that the sequential hit rate would also be slightly negative. But instead it is *highly positive* (in statistical terms), suggesting that the sequential hit rate is not due to optional stopping.

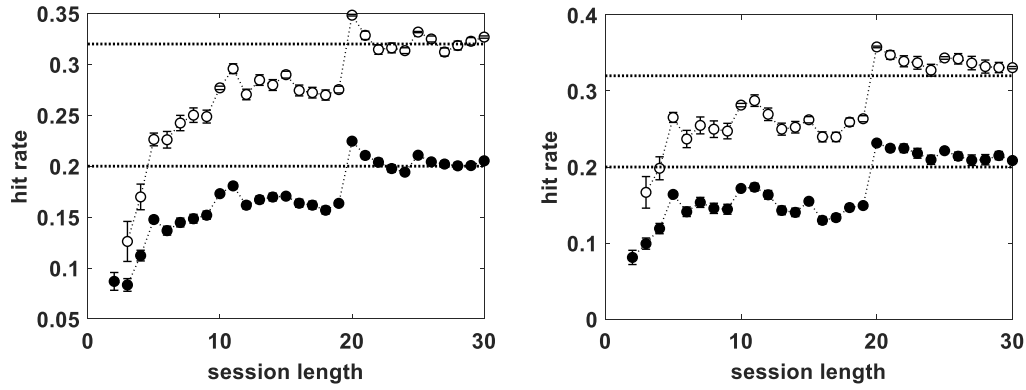


Figure 7. (Left). Card test direct hit rate ( $\pm 1$  se, which is so small that the error bars are difficult to see) (black dots) and sequential hit rate (white dots), for session lengths from exactly 2 to 30 trials. The negative hit rates prior to 20 trials is due to optional stopping behavior. Sessions of 20 trials are the most common predefined run-length. (Right). Same for QRV data.

### 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 contributed. The same was separately determined for the sequence of responses. A chi-squared ( $\chi^2$ ) 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,  $\chi^2 = 10,974,545$ ,  $p \approx 0$ . The nonsignificant  $\chi^2$  for targets indicates that successive pairs of targets occurred in a random order, and the extremely large  $\chi^2$  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 was 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  $\chi^2$  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 provided people with 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 did not respond at random, but that could not account for the  $hr_{seq}$  result. This analysis examined one factor responsible for a portion of that non-random behavior. Figure 8 shows that users' responses tended to follow the targets, that is R on trial  $N_{+1}$  was the same as the randomly generated T on trial N. Could this 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.

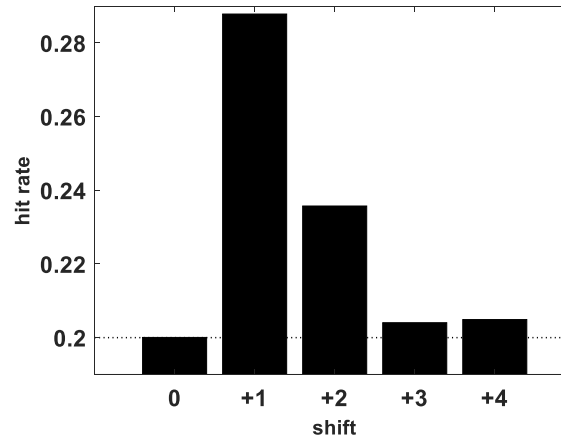


Figure 8. Direct hit rate for circular shift lag 0 to +4, indicating that people tended to respond on trial  $N_{+1}$  using the target that was presented on trial N, i.e. they followed the trials. This tendency continued for  $N_{+2}$  and to a limited extent to +3 and +4.

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 actually responded did not generate inflated values for  $hr$  or for  $hr_{seq}$ .

## Unconscious Learning

Perhaps subtle patterns were unconsciously noticed by individuals who had 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. The learning hypothesis predicts that this correlation should be positive.

To evaluate this hypothesis, we formed one array of trials per person per 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 was repeated for all 10,372 days. Then the resulting mean z score was compared to 0 (the null hypothesis) using a two-tailed t-test. 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 each individual contributed, the *smaller* their resulting  $hr_{seq}$ . This is opposite to the prediction of the learning hypothesis.

## Permutation and Circular Shift Analyses

The anomaly we are studying is summarized as the left-most points in the top and bottom graphs of Figure 9. These two means and error bars indicate that the overall mean direct  $hr$  across all trials was close

to chance (top graph), whereas the sequential  $hr_{seq}$  was 10 sigma above chance (bottom graph). Note that if the  $hr_{seq}$  results were due to optional stopping, then we would expect  $hr_{seq}$  to be *negative* given that  $hr$  (the conventional meaning of hit rate) and  $hr_{seq}$  are correlated, as shown in Figure 7. But that is not what we see, arguing against optional stopping as the explanation for the  $hr_{seq}$  effect. The next point in Figure 9, labeled “random,” checked if the analytical method might have contained an inherent bias. This was performed by generating an entirely new randomized set of targets, then recalculating  $hr$  and  $hr_{seq}$ . The resulting hit rates are both closely in alignment with chance expectation.

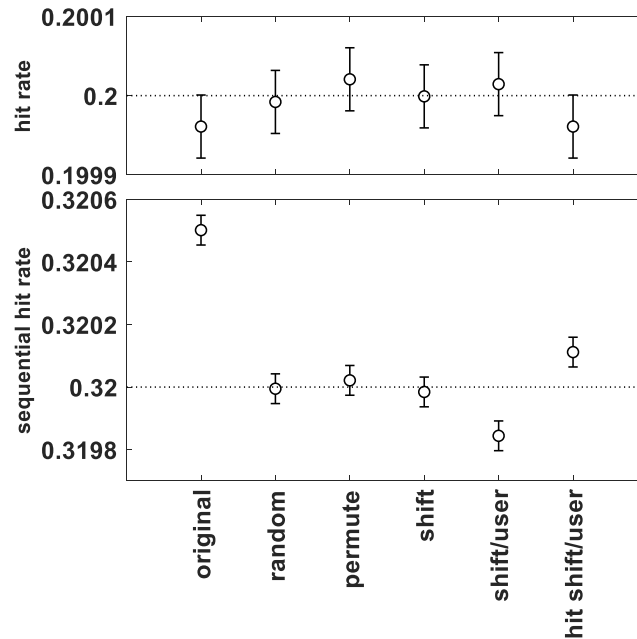


Figure 9. (Top) Mean direct hit rates and one standard error bars for (1) original data combined across Card and QRV tests, (2) newly generated random targets, (3) randomly permuted targets, (4) targets shifted left lag -1 per day (5) targets shifted lag -1 per person, and (6) hits shifted lag -1 per person. (Bottom). Same analyses for sequential hit rates.

The next point in Figure 9, labeled “permute,” maintained the same frequency of targets (i.e., numbers of 1s, 2s, ...) as originally recorded on each day, except that their sequence was randomly permuted (i.e., scrambled). This too resulted in chance outcomes. The third test, labeled “shift,” maintained both the same frequency and sequential structure of the original targets by circular-shifting the targets backwards one step. In this way each response  $R_n$  was matched not to target  $T_n$ , but to  $T_{n+1}$  (i.e., the target after the current target). The results were again in accord with chance expectation.

The fourth test, labeled “shift/user,” was like the previous test except that it maintained the same target frequency and sequence observed by *each participant*, rather than shifting the chronological order of targets as they were recorded over the course of each day. That is, many people could have been taking this test at the same time, so the chronological sequence of targets generated was not necessarily the same as the sequence that each individual actually obtained. This test resulted in a nonsignificant direct  $hr$  and a  $hr_{seq}$  about 3 sigma below chance.

The fifth point, labeled “hit shift/user,” was like the previous test except it maintained the same number and sequence of *hits* generated by each participant, where the hits were shifted one step backwards. This test therefore maintained the same overall hit rate as originally obtained per user, *so it mimicked any optional stopping behavior*. It also closely matched the analytical method used to determine  $hr_{seq}$ , because it relied only on the sequential order of hits and misses. The result was a mean  $hr_{seq}$  about 2 sigma above chance.

Notice that by design the direct *hr* for this test was *exactly* the same as the original *hr*. In sum, none of the control tests produced deviations approaching 10 sigma.

## DISCUSSION

The sequential hit rate  $hr_{seq}$  (a) does not appear to be due to analytical or computational artifacts, (b) the frequency and sequence of targets used in these online experiments were adequately random, (c) transitions of targets from  $T_{i+1}$  to  $T_{i+4}$  were in alignment with chance expectation, (d) circular shifts of targets and hits did not result in remarkable deviations, (e) the possibility of learning subtle cues from dependencies in target sequences showed results consistent with chance or opposite from what one would expect if those dependencies were the cause of the  $hr_{seq}$  effect, and (f) the results are not consistent with optional stopping behavior.

If  $hr_{seq}$  is indeed a genuine effect, then it would appear to require a 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 unreasonable; the literature suggests that psi effects arise from the unconscious, and that it is modulated by a host of psychological filters and defense mechanisms (e.g., Crumbaugh, 1968; Carpenter, 1966, 1977, 2012; Eisenbud, 1993; Johnson & Kanthamani, 1967). Analyzing psi data for sequential patterns is also not a new idea (Burdick & Kelly, 1977), 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 prediction of a particular way that a forced-choice psi test can result in a null effect via a direct hit measure while at the same time produce a highly significant effect in a rather complex sequential measure. If this  $hr_{seq}$  deviation had been discovered after extensive data snooping, it would not be especially remarkable. But this was not the case. What was presented here is the result of a single analysis employed for the sole purpose of checking a prediction, and that prediction was ultimately confirmed in two different forced-choice databases. What was the origin of this prediction?

### “Holy Thought”

The sequential hit rate effect was proposed by an empirically-oriented group of Christian Scientists, known as *Spindrift*, which was active from the 1970s to 1990s (Klingbeil, nd; Sweet, 2007). The two primary Spindrift researchers envisioned that perceptual psi effects are ubiquitous and robust, but that the unconscious mind is trickster-like and highly adept at hiding these abilities. A similar notion, based on clinical and empirical observations instead of Christian Science concepts, has been proposed by Carpenter (2012).

Spindrift’s underlying belief was that the mind has three primary components: *an ordering force*, a *perceptive ability*, and a *defense mechanism*. The ordering force was related to what they called “holy thought,” an ego-less, or “thy will be done” mode of consciousness. They proposed that this force induces *order* – synonymous terms would be equilibrium, coherence, balancing, and negentropy – into any system that it decides to focus upon (Pallikari-Viras, 1997; Radin, 1993; Radin, Taft & Yount, 2004). Spindrift’s concept of a perceptive ability was what a parapsychologist would call psi, and their notion of a defense mechanism was an unconscious mental effect that served to actively hide psi effects.

Spindrift researchers tested their ideas in many ways, often reporting highly positive effects. One such experiment, which they called VIUR for “visual image, unconscious response,” was a simple binary card test. The test involved a deck consisting of 12 copies each of two images. To begin the test, a participant would start by selecting an image that they liked from a pool of images, and then a second image that they did not like. Cards containing these images were then placed into opaque envelopes, shuffled, and then the participant guessed what they thought each envelope contained. Later, they or an independent experimenter would record the resulting sequence of hits and misses, and then 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 the “holy” attention focused on the task would 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 would defeat the masking action of the defense mechanism. And thus the “best” order would be to alternate hits and misses and to hide that sequence within each response type. Figure 3 illustrated the analytical method they used to detect this hypothetical three-component process.

In 1990, the author and a colleague attempted to replicate the VIUR 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  $p_l = 0.52$ , which was virtually identical to what Spindrift had been reporting in their tests. A 2.5% effect over chance is not all that impressive, but after being amplified with more than 10,000 trials, as was common in that group, they were able to report large positive  $z$  scores.

In the analyses reported here, the sequential hit rate 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 Card and QRV data were contributed in an online, 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 claims due to suspicions about Spindrift’s religious motives, and because the reported results seemed too good to be true. It was also a problem that the method of analysis was unorthodox and not clearly described. These discomforts contributed to an assumption that the results might have been due to sloppy methods or to one or more analytical artifacts. After we obtained essentially the same results that Spindrift was reporting, we were less sure that such dismissals were valid. But we did continue to worry about analytical artifacts because, among other things, the VIUR test used a closed deck design, and that complicates the determination of the chance-expected hit rate. Based on our concerns, we placed this study in the filedrawer, where it patiently sat for nearly three decades.

### *Uniform Responses*

The Perl pseudorandom number generator (PRNG) is the source of randomness in the GotPsi.org suite of experiments. PRNGs are deterministic algorithms, and as such the sequence generated is not susceptible to external influences. However, the PRNG in these tests was reseeded on each trial based on the server system timestamp, so if the user interacted with the computer at a “fortuitous time” (a euphemism for precognition), then in principle it would be possible to select a portion of the PRNG output that provided a result that achieved the desired goal. Assuming the targets generated were sufficiently random, which all the above control tests suggest, then any robustly non-chance result, including the sequential hit rate  $hr_{seq}$ , would presumably require precognition. If this explanation is sound, then trials in the Card and QRV tests where the user repeatedly selected only one response (we will call these data “uniform-responses”), but the *timing* of each response was still up to the user’s discretion (as it was), should also show positive results for  $hr_{seq}$ .

To test this prediction, the uniform-response trials were extracted from the Card and QRV databases and  $hr_{seq}$  was determined. For the Card test, 6,948,377 suitable trials were found; they resulted in  $hr_{seq} = 0.32060 \pm 0.00018$  ( $z = 3.41$ ). For the QRV data 769,082 suitable trials were found; they resulted in  $hr_{seq} = 0.320605 \pm 0.000533$  ( $z = 1.14$ ). Note that in both cases  $hr_{seq}$  was about the same magnitude as  $hr_{seq}$  measured when the response R had been newly selected trial-by-trial. This means that if the uniform-response hit rate had been evaluated based on 100 million trials, the resulting  $z$  score from this subset would be over 10 sigma.

To check on this outcome, a control test was performed by having a simulated user repeatedly select the same response on the GotPsi.org website. A web-based program (iMacro, from www.ipswitch.com) was

used to generate a total of 150,000 Card test trials by repeatedly selecting  $R = 1$  for 50,000 trials, then repeating this for  $R = 3$  and  $R = 5$ . The resulting direct hit rate was  $hr = 0.199367 \pm 0.001033$  ( $z = -0.613$ ) and  $hr_{seq} = 0.319649 \pm 0.001204$  ( $z = -0.292$ ). Thus, this control indicated that a machine that selected the same trial repeatedly did not result in above-chance biases. Note that it also indicated that a negative  $hr$  was associated with a negative  $hr_{seq}$ , which is the positive correlation that we would expect, again arguing against an optional-stopping artifact.

### *A Trimūrti Model*

Finally, we explored a model for the sequential effect, whereby the act of selecting a response  $R$  the first time places a probabilistic goal-oriented bias on  $R$ , such that 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 goal-oriented bias is reversed. This pull/push scheme suggests a tension between creation and destruction, an effect proposed to be part of the fabric of reality, a dynamic balance that is constantly attempting to sustain order. We call the model *Trimūrti*, named after the Hindu trinity of Brahma the Creator, Vishnu the Preserver, and Shiva the Destroyer. The tendency for psi effects to tweak probabilities of desired events, followed shortly afterwards by “anti-tweaks,” has been repeatedly noted in the parapsychological literature (Pallikari-Viras, 1997; Palmer & Kramer, 1984; Radin, 1993; Stanford & Fox, 1975; Williams, 2008).

To simulate the effect of small, alternating hit and miss biases, 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 can produce a null effect for the direct hit rate and a very significant positive sequential hit rate. It is not claimed that this model accurately reflects the unconscious processes followed by the users in the online test, only that it is possible to construct a straightforward model that mimicks the observed results.

## CONCLUSION

This study suggests that deeper analysis of the results of simple psi experiments may reveal subtle patterns in data that were previously overlooked. Understanding those patterns may allow experimenters to overcome difficulties in repeating psi effects. It may also reveal that interference by the so-called trickster may not be due to mythical creatures determined to hide psi, but rather to our own ignorance about the unconscious mental mechanisms underlying performance in psi effects.

## REFERENCES

- Beloff, J. (1994). Lessons of history. *Journal of the American Society for Psychical Research*, 88, 7–22.
- Bloch, A. (1978). *Murphy's Law and other reasons why things go wrong*. Price/Stern/Sloan Publishers, Inc.; Third Printing edition (March 1, 1978)
- Bösch, H. (2004). Reanalyzing a meta-analysis on extra-sensory perception dating from 1940, the first comprehensive meta-analysis in the history of science. *Proceedings of Presented Papers*, 47th Annual Convention, August 5-8, 2004. The Parapsychological Association, Inc.
- Burdick, D. S., & Kelly, E. F. (1977). Statistical methods in parapsychological research. In B. B. Wolman (Ed.), *Handbook of Parapsychology*. New York: Van Nostrand Reinhold Company, pp. 81-130.

- Campbell, J. (2008) *The hero with a thousand faces*. New World Library; Third edition.
- Carpenter, J. C. (1966). Scoring effects within the run. *Journal of Parapsychology*, 30, 73-83.
- Carpenter, J. C. (1977). Intrasubject and subject-agent effects in ESP experiments. In B. B. Wolman (Ed.), *Handbook of Parapsychology*. New York: Van Nostrand Reinhold Company, pp. 202-272.
- Carpenter, J. C. (2012). *First Sight: ESP and Parapsychology in Everyday Life*. Rowman & Littlefield Publishers.
- Cohen, J. 1990. Things I have learned (so far). *American Psychologist*. 45(12) 1304-1312.
- Crumbaugh, J. C. (1968). Variance declines as indicators of a stimulator-suppressor mechanism in ESP. *Journal of American Society for Psychical Research*, 62, 356-365.
- Eisenbud, J. (1993). *Parapsychology and the Unconscious*. North Atlantic Books; Revised.
- Hanson, G. (2001). *The Trickster and the Paranormal*. Xlibris, Corp.
- Honorton, C., & Ferrari, D. C. (1989). "Future telling": A meta-analysis of forced-choice precognition experiments, 1935–1987. *Journal of Parapsychology*, 53, 281–308.
- Jahn, R.G., Dunne, B. J., Nelson, R. D., Dobyns, Y. H., and Bradish, G. J. (1997). Correlations of Random Binary Sequences with Pre-Stated Operator Intention: A Review of a 12-Year Program. *Journal of Scientific Exploration*, 11 (3), 345–367.
- Johnson, M., & Kanthamani, B. K. (1967). The Defense Mechanism Test as a predictor of ESP scoring direction. *Journal of Parapsychology*, 31(2), 99-110.
- 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.
- Kennedy, J. E. (2003). The capricious, actively evasive, unsustainable nature of psi: A summary and hypotheses. *Journal of Parapsychology*, 67, 53-74.
- Klingbeil, J. & Klingbeil, B. (no date). *The Spindrift Papers: Exploring Prayer and Healing Through the Experimental Test*. Available from <http://www.spindriftresearch.org/> and <http://www.bloomingtononline.net/directory/docs/857.pdf>, as of October 22, 2017.
- Lucadou, v. W. (2015). The Model of Pragmatic Information (MPI). In: Edwin C. May & Sonali Marwaha (eds.) *Extrasensory Perception: Support, Skepticism, and Science: Vol. 2: Theories and the Future of the Field*. Praeger publications, SantaBarbara, USA, Ca., pp.221-242.
- Pallikari-Viras F., (1997). Further evidence for a statistical balancing in probabilistic systems influenced by the anomalous effect of conscious intention. *Journal of the Society for Psychical Research*. 62, 114-137
- Palmer, J. & Kramer, W. (1984). Internal state and temporal factors in psychokinesis. *Journal of Parapsychology*, 48, 1—25.
- Pratt, J. G., Rhine, J. B., Smith, B., Stuart, C. & Greenwood, J. (1940). *Extrasensory Perception After Sixty Years*. Boston: Bruce Humphries Publishers.
- Radin, D. I. & Cross, G. R. (1990). Sequential ordering effects in a perceptual experiment: Testing the nature of prayer and psi. Unpublished report, GTE Laboratories.
- Radin, D. I. (1993). Environmental modulation and statistical equilibrium in mind-matter interaction. *Subtle Energies and Energy Medicine*, 4 (1), 1-30.
- Radin, D., Taft, R., & Yount, G. (2004). Effects of healing intention on cultured cells and truly random events. *Journal of Alternative and Complementary Medicine*, 10 (1), 103–112.
- Radin, P. (1956). *The Trickster: A study in American Indian mythology*. New York : Philosophical Library.
- Stanford, R. G. (2015). Psychological concepts of psi function. A review and constructive critique. In Etzel Cardena, John Palmer, Dvaid Marcusson-Clavertz (Eds). *Parapsychology: A Handbook for the 21st Century*. McFarland & Company, Jefferson, NC, Pp. 94-109.
- Stanford, R.G. & Fox, C. (1975). An effect of release of effort in a psychokinetic task [abstract]. In J. D. Morris, W. G. Roll & R. L. Morris (Eds.), *Research in Parapsychology 1974* (pp. 61-63). Metuchen, NJ: Scarecrow Press.

- Storm, L., Tressoldi, P. Di Risio, L. (2012). Meta-analysis of ESP studies, 1987-2010: Assessing the success of the forced-choice design in parapsychology. *Journal of Parapsychology*, 76 (2), 243-273.
- Sullivan, G. M., Feinn, R. (2012, September). Using Effect Size—or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, 279-282
- Sweet, B. (2007). *A Journey into Prayer: Pioneers of Prayer in the Laboratory: Agents of Science or Satan?* Xlibris. Kindle Edition.
- Williams, C. (2008). Conceptual metaphor:: A meaning-oriented approach for parapsychology. *Journal of the Society for Psychical Research*, 72.3, No. 892, p.142.