Note: The effect of optional stopping on a proposed sequential statistic for the analysis of psi experiments

Chris Phillips
19 July 2021

Abstract. An new method of analysis using a sequential statistic was recently applied to data from online forced-choice psi experiments (Radin, 2019), and produced results that appeared to be highly statistically significant. Although the data showed evidence of significant optional stopping effects, it was argued on the basis of additional statistical checks that these were unlikely to explain the results. In this note it is demonstrated that in principle optional stopping can produce artefactual effects similar to those observed. It is also shown using a simple model that the additional statistical checks are not capable of excluding this possibility. Instead a straightforward test is recommended to determine whether or not optional stopping is responsible for the observations.

1. Introduction

In a recent paper, Dean Radin proposed an alternative method of analysis for forced-choice psi experiments (Radin, 2019). Instead of the conventional hit rate, he suggested using a new statistic depending on the sequence in which hits and misses occur. The motivation for this proposal was that even if the hit rate does not deviate significantly from its expected value (for example, owing to a "trickster" effect tending to conceal the presence of psi), more subtle deviations from chance expectation may be reflected in the structure of the sequence of hits and misses. Specifically, there may be a higher than expected rate of alternation between hits and misses.

Radin used this method to analyse a large set of data produced by two long-running online psi experiments. In each, participants had to choose between five alternatives, so the expected probability of success in the absence of psi was p=0.2. The data consisted of 83.9 million trials from a precognitive card test and 26.8 million from a remote viewing test. When analysed conventionally using the hit rate, the card test produced a non-significant result, while the hit rate in the remote viewing test was significantly above expectation (with Stouffer z=2.188). But the combined hit rate from both sets of data did not deviate significantly from chance.

However, when analysed using the new method, both sets of data produced results that appeared to be highly significant.

The sequential statistic used in this analysis was suggested by one used previously by other workers to analyse a two-choice, closed-deck card guessing experiment. The statistic, which will be called Y_s below, was defined as follows:
(1) The experimental data were divided into separate sessions, each session being defined as all the trials performed by a particular user on a given day.
(2) The trials in each session were further divided into up to five separate sub-sequences, according to which of the five alternatives the participant chose. So the first sub-sequence consisted of the trials (if any) for which the user chose the first of the five alternatives, and so on.
(3) Within each sub-sequence, a total was counted for the number of times a hit was directly followed by a miss, or a miss was directly followed by a hit. The statistic for each session was then defined by adding up the totals for all the sub-sequences in the session. Thus the statistic is the total number of alternations between hits and misses in the session, with each of the sub-sequences considered separately.

To analyse the calculated values of this statistic, Radin assumed that each pair of trials contributing to it (that is, each pair of consecutive trials within the sub-sequences) could be treated independently, and the probabilities of a hit being followed by a miss, and a miss being followed by a hit, could be calculated in the obvious way from the probability of a hit in a single trial. With these assumptions, if the probability of a hit in a single trial were p, then the probability of a hit being followed by a miss would be $p(1-p)$, and the probability of a miss being followed by a hit would be the same. Therefore, the value of the sequential statistic for the whole experiment would have a binomial distribution determined by the total number of contributing pairs of trials, with a probability parameter equal to $2p(1-p)$. (Note that, as discussed further below, the total number of pairs of trials contributing to the statistic - namely the number of pairs of consecutive trials within sub-sequences - was smaller than the total number of trials in the experiment. The difference between them was equal to the total number of the sub-sequences in the experiment, and in these experiments each session had up to five sub-sequences.)

On this basis, the deviation of the sequential statistic from the assumed value of $2p(1-p)$ per contributing pair appeared to be highly statistically significant for both experiments. The Stouffer z values were found to be 8.07 for the card test and 8.70 for the remote viewing test. When the two sets of data were combined a value of $z=11.28$ was obtained. This would constitute one of the statistically strongest pieces of experimental evidence for psi ever claimed.

Radin considered a number of possible conventional explanations for this finding. One possibility - inadequate randomisation of the choice of target - was ruled out on the basis of statistical checks on the data.

Another possibility considered was that the sequential statistic per contributing pair exceeded the assumed value of $2p(1-p)$ owing to the effects of optional stopping. In the online experiments, the participants were free to stop at any time, rather than completing a prespecified number of trials. They were also free to continue for as long as they liked, or to return several times on the same day, in which case all the trials would be analysed as one long session. As discussed further below, provided none of the experimental data are excluded from analysis, optional stopping cannot affect the overall hit rate in the experiment. But in principle it can introduce bias into the sequential statistic.

Nevertheless, Radin argued that optional stopping was unlikely to be responsible for the results he observed, on the basis of two further statistical checks:

(1) The average values of the hit rate per trial and the sequential statistic per contributing pair were calculated as functions of the length of the session. For sessions shorter than 20 trials (which was the commonest prespecified session length) the hit rate per trial was found to be well below the expected value of 0.2, which indicated a significant effect of optional stopping. But as session length varied, the average hit rate and the average sequential statistic were positively correlated. As the overall hit rate for the card test was found to be slightly (but not significantly) below expectation, it was argued that optional stopping was unlikely to explain the highly significant increase of the sequential statistic per contributing pair above the assumed value of $2p(1-p)$.

(2) The sequential statistic was recalculated from the experimental data, with the sequence of hits and misses randomly permuted within each session (but with the sequence of guesses unchanged). The results were non-significant for each experiment individually and also for both experiments combined. As the permutation left the overall hit rate for each session unchanged, it was argued that this also indicated that optional stopping was not responsible for the increase in the sequential statistic.

Radin's conclusion was that optional stopping was unlikely to explain the observed effect on the sequential statistic. The purpose of this note is to consider whether this conclusion is safe.

2. The effect of optional stopping

As is well known, when the total number of trials in an experiment is large, on the null hypothesis optional stopping has no effect on the overall expected hit rate, provided all the data are retained,[1] including the data from sessions terminated prematurely. This is because on the null hypothesis the probability of success in each trial is p and the results of all the trials are statistically independent. The division into individual sessions is irrelevant, and the data can be treated as a single long sequence of independent trials. Therefore, for a given number of trials, the total number of hits follows the binomial distribution.

However, if an attempt is made to apply the same reasoning to the sequential statistic, it fails. If the data are treated as a single long sequence of trials and a statistic Y_s' is defined by analogy to the sequential statistic Y_s above, it can indeed be argued that each pair of trials makes a contribution of 1 to the statistic with probability $2p(1\text{-}p)$ and that all these contributions are statistically independent. The distribution of the statistic Y_s' is therefore binomial.

But Y_s' differs from the sequential statistic Y_s defined above, because it includes contributions from pairs of trials from two different sessions. Specifically, for a given session, for each sub-sequence represented, Y_s' includes an additional contribution from the pair consisting of the last trial of the sub-sequence and the first trial of the corresponding sub-sequence in a later session. For prematurely terminated sessions, the last trials of the sub-sequences may not be typical of trials in general. For example, if participants tended to terminate sessions prematurely because the hit rate was low, then the probability of success in the last trials of the sub-sequences might tend to be smaller than p. That would mean that the contributions present in Y_s' but excluded from Y_s would tend to be smaller than the assumed value $2p(1\text{-}p)$. But Y_s' is unbiased, so in compensation the remaining contributions - the ones included in Y_s - would have to be larger than $2p(1\text{-}p)$. As a result, Y_s would have a positive bias relative to the assumed value.

To see this in quantitative terms, consider the general case in which there are G possible possible positions for the target and for the participant's guess. Then define the variable X_a as follows:

$$X_a = \begin{cases} 1 & \text{if trial } a \text{ is successful} \\ 0 & \text{if trial } a \text{ is unsuccessful} \end{cases} \tag{1}$$

Then consider a particular session of the experiment, and the sub-sequence of trials in which the participant chooses option g out of the G alternatives. Denote by D_g the contribution to the difference between Y_s' and Y_s arising from this sub-sequence. If there is no sub-sequence for g in the session, then D_g is zero. But when a sub-sequence is present at g, then it can be written as

$$D_g = X_b(1-X_c)+(1-X_b)X_c \tag{2}$$

1 But note that in Radin's analysis of the card test (and perhaps of the remote viewing test too), sessions with only a single trial were excluded. This is because the sequential statistic is defined in terms of pairs of trials, and is equal to zero for sessions containing only one trial. But this exclusion could potentially bias the overall hit rate. In an earlier version of the paper, for the card test (and perhaps for the remote viewing test too) sessions in which the participant chose only one of the five alternatives were excluded (Radin, 2018). These would include all sessions with only a single trial and other mainly very short sessions, so this exclusion could also bias the overall hit rate.

where *b* denotes the last trial of the sub-sequence at *g* in the current session, and and *c* denotes the next trial of the experiment after the current session in which option *g* is chosen.

This quantity can first be averaged over all realisations of the sessions following the current one. The probability of success in trial *c* is simply *p*. In other words, the expected value of X_c is *p*. Therefore when a sub-sequence is present at *g*, this first-stage average of D_g is

$$p + (1-2p)X_b \tag{3}$$

It remains to average over all realisations of the current session. In the presence of optional stopping, the outcome of each trial can influence the likelihood that the session is later prematurely terminated. That means there may be a statistical relationship between the probability of success in a given trial, and the probability that it is the last trial of its sub-sequence. Therefore, given that trial *b* is defined as the last trial of the sub-sequence corresponding to option *g*, the probability of success in trial *b* - equal to the expected value of X_b - may differ from *p*. Denote the average value of this probability by p_g, where the average is performed over all the realisations of the session in which a sub-sequence is present at *g*. (The average therefore covers all the possible positions of the last trial of the sub-sequence within the session.)

Recalling that D_g is zero when there is no sub-sequence at *g* in the session, it is also necessary to take into account this possibility. Denote by r_g the probability that sub-sequence is present at *g*. Then the average value of D_g over all sessions is given by

$$\overline{D_g} = \left(p + [1-2p]p_g\right)r_g \tag{4}$$

in which the overbar indicates an average over all realisations of a single session of the experiment, including averaging over different session lengths.

The total expected difference between Y_s' and Y_s per session is the sum of $\overline{D_g}$ over all *G* of the possible values of *g*. Using equation (4), this is

$$\sum_{g=1}^{G} \overline{D_g} = \left(p + [1-2p]p_e\right)\overline{S} \tag{5}$$

in which $\overline{S}$ is the average number of sub-sequences present in a session, given by

$$\overline{S} = \sum_{g=1}^{G} r_g \tag{6}$$

and p_e is the average probability of success in the last trial of any sub-sequence, given by

$$p_e = \frac{1}{\overline{S}} \sum_{g=1}^{G} r_g p_g \tag{7}$$

Note that p_e is an average weighted by the frequencies with which the different sub-sequences are present in the sessions. These frequencies may differ according to the value of *g*, because some options may be more popular with participants than others.

Finally, note that the expected contribution per session to the unbiased statistic Y_s' is equal to the average number of trials per session, $\overline{N}$, multiplied by the expected value per trial, $2p(1\text{-}p)$. Similarly, the expected contribution per session to the sequential statistic Y_s is equal to the average number of contributing pairs of trials per session, $\overline{N}-\overline{S}$, multiplied by the (unknown) expected contribution per pair, which will be denoted α.

Then from equation (5), after some rearrangement, the expected contribution per contributing pair is found to be

$$\alpha = 2p(1-p) + \frac{(1-2p)\overline{S}}{\overline{N}-\overline{S}}(p-p_e) \qquad (8)$$

This equation confirms that if, because of optional stopping, the average success rate in the last trials of the sub-sequences falls below the average success rate p, then the sequential statistic per contributing pair will be artefactually raised above the assumed value of $2p(1\text{-}p)$.

3. Arguments against optional stopping as an explanation of the experimental results

In Radin's paper, two arguments are made against optional stopping as an explanation for the increase of the value of the sequential statistic per pair of trials above the assumed value of $2p(1\text{-}p)$.

Firstly, it is pointed out that when averages are calculated for sessions of different lengths, there is a positive correlation between the hit rate and the sequential statistic. Both these quantities depend on session length because of optional stopping. Moreover a positive correlation is to be expected, because if the proportion of hits is reduced the opportunities for alternation between hits and misses will also decrease. But the overall hit rate in the card test was below the expected value of 0.2 (though the difference was not statistically significant). Therefore, it is argued, it might be expected that optional stopping would decrease the value of the sequential statistic, rather than raising it.

The problem with this argument is that if optional stopping introduces a small but systematic bias in the sequential statistic, the correlation between it and the hit rate could remain positive. When the hit rate is averaged over different session lengths, the effects of optional stopping cancel out, and the expected hit rate per trial is equal to exactly 0.2. But when the sequential statistic is averaged over different session lengths, any systematic bias will accumulate rather than cancelling out. Therefore, if bias is present, there is no inconsistency between a statistically non-significant decrease in the hit rate, and a statistically highly significant increase in the sequential statistic.

The second argument against optional stopping as an explanation is based on the observation that a permuted version of the sequential statistic was not significantly different from the assumed value of $2p(1\text{-}p)$ per contributing pair of trials.

The permuted version of the sequential statistic is defined as an average for each session over all possible permutations of the sequence of hits and misses within the session (but leaving unaltered the sequence of choices made by the participant, and therefore maintaining the same division of the trials into sub-sequences).

Suppose a session contains N trials, S different sub-sequences and a total of H hits. A simple expression for the permuted statistic can be obtained as follows. Consider first the contribution to

the statistic from each hit in the sequence. During the permutation operation, the hit is assigned to each of the N possible positions in the session with equal possibility. If it is placed at the end of one of the S sub-sequences, it does not contribute to the statistic. Therefore the number of positions for which it does contribute is $N-S$ out of a total of N.

When it is in one of these positions, the hit is followed by each of the other $N-1$ elements of the sequence with equal probability. When it is followed by another hit - that is in $H-1$ out of $N-1$ cases - this pair contributes 0 to the statistic. But when it is followed by a miss - that is in $N-H$ out of $N-1$ cases - the pair contributes 1.

Therefore the contribution of each hit to the permuted statistic is just the product of the fraction of cases where the hit contributes, that is $(N-S)/N$, and the fraction of cases where the contribution is 1, that is $(N-H)/(N-1)$. Multiplying by the number of hits H gives the total contribution of all the hits, namely

$$\frac{(N-S)H(N-H)}{N(N-1)} \tag{9}$$

The total contribution of all the misses in the sequence is obtained by exchanging the number of hits, H, with the number of misses, $N-H$. This gives exactly the same quantity. So the total permuted statistic is twice the quantity given by (9).

Dividing by $N-S$ gives the permuted sequential statistic per contributing pair, which will be denoted by $A^{(p)}$. This can be expressed in terms of the hit rate per trial, $\rho=H/N$, as

$$A^{(p)} = \frac{2N}{N-1}\rho(1-\rho) \tag{10}$$

To find the average value of the permuted sequential statistic per contributing pair, for all sessions of length N, it remains to average this expression over the different possible values of ρ, the hit rate. The answer will depend on the averages of both ρ and ρ^2. The expected contribution per contributing pair, $\alpha^{(p)}(N)$, is found to be

$$\alpha^{(p)}(N) = \frac{2N}{N-1}\left(\mathrm{E}(\rho)[1-\mathrm{E}(\rho)]-\mathrm{Var}(\rho)\right) \tag{11}$$

in which E indicates the expected value and Var the variance of ρ, calculated over all realisations of the session in which its length is N. In general both these quantities will vary with N.

Although, as expected, the permuted statistic has a dependence on ρ similar to the dependence of the unbiased value $2p(1-p)$ on p, this is modified both by the contribution of the variance of ρ, which will tend to decrease the statistic, and by the factor $N/(N-1)$, which will tend to increase it. Therefore even if optional stopping were influenced primarily by the overall hit rate in the session so far, the dependence of the permuted sequential statistic on the hit rate would not be a simple one.

4. A simple model

In order to check whether optional stopping is really capable of producing the results observed in the experiment, a simple theoretical model can be used.

Suppose the participant's choices are determined randomly and distributed uniformly among the possible options, with the choice in each trial being statistically independent of the others. Suppose also that after each trial a decision whether to continue the session is made randomly. Consider the simple case in which the probability of termination is determined by a prescribed stopping rule, that depends only on the trial number and the number of hits so far.

In this model, the probability of every possible outcome of a session can be computed exactly. To do this, the results of the trials preceding the *n*th trial are characterised in terms of three variables: (1) *h*, the total number of hits, (2) *s*, the number of sub-sequences present, and (3) *e*, the number of sub-sequences for which the last trial was successful. It is necessary to calculate the joint probability distribution of these three variables for each value of *n*, and also the mean value of the sequential statistic *a* as a function of the three variables and *n*.

With the assumption that the participant's choices are random, uniformly distributed and independent, the changes in the variables *h*, *s*, *e* and *a* at the *n*th trial can be described by a simple set of transition probabilities as follows:

(1) If the trial is successful, the number of hits increases by 1. Otherwise it is unchanged.

$$\text{Probability}\left(h \to h+1\right) = p \tag{12}$$

(2) If the participant's choice does not coincides with one of the *s* positions where there is already a sub-sequence (out of the *G* possible options) then the number of sub-sequences increases by 1. Otherwise it is unchanged.

$$\text{Probability}\left(s \to s+1\right) = \frac{G-s}{G} \tag{13}$$

(3) If the trial is successful and the participant's choice does not coincide with one of the *e* positions where there is already a sub-sequence ending in a hit (out of the *G* possible options) then the number of sub-sequences ending in hits increases by 1.

$$\text{Probability}\left(e \to e+1\right) = \frac{\left(G-e\right)p}{G} \tag{14}$$

But if the trial is unsuccessful and it does coincide with one of the *e* positions where there is already a sub-sequence ending in a hit, then the number of sub-sequences ending in hits decreases by 1.

$$\text{Probability}\left(e \to e-1\right) = \frac{e\left(1-p\right)}{G} \tag{15}$$

Otherwise it is unchanged.

(4) The sequential statistic increases by 1 in two cases. Firstly if the trial is successful and the

participant's choice coincides with the position of one of the $s-e$ sub-sequences ending in a miss (out of the G possible options). And secondly if the trial is unsuccessful and it coincides with the position of one of the e sub-sequences ending in a hit. Otherwise it is unchanged.

$$\text{Probability}\left(a \to a+1\right) = \frac{\left(s-e\right)p + e\left(1-p\right)}{G} \qquad (16)$$

From this, it is straightforward to compute the joint probability distribution of the variables h, s and e, and the mean value of a as a function of these variables, after the nth trial. The initial condition is that all of these variables are equal to zero before the first trial. Then equations (12-16) are repeatedly used to step forward by one trial. After each trial, the prescribed stopping rule is used to adjust the probability distribution, and to calculate the average characteristics of the sessions that terminate at that point.

To motivate the choice of a stopping rule that may be capable of producing an increase in the sequential statistic per contributing pair, α, without a corresponding increase in the permuted version, $\alpha^{(p)}$, consider equation (11) for the latter statistic. This shows that its value will tend to be smaller if the hit rate per trial, ρ, has a large variance over the set of sessions of a given length. That suggests that it would be worth examining a stopping rule that produces a large degree of inhomogeneity among the hit rates of different participants. Therefore consider a situation in which the session may be terminated prematurely in two very different circumstances. In the first case, it may be terminated if the hit rate is so discouragingly low that the participant gives up. In the second, it may be terminated if the hit rate is high and the participant wants to "quit while s/he is ahead". (In practice these two kinds of behaviour might represent distinct subgroups of participants with different psychological characteristics.)

To construct such a model, the following procedure is used. After each trial the probability distribution of hit rates is calculated, and the range of hit rates to be considered low is defined by specifying a percentile. (Because the hit rate is a discrete variable, it will be necessary to choose whether to include or exclude the value coinciding with the percentile. Here it will be included in the range, so that at each stage the percentage of hit rates defined as low will be somewhat larger than the specified percentile.) The range of hit rates to be considered high is defined similarly. Then, within each of the ranges, the session will be considered to terminate with a certain probability. At each stage this probability will be fixed by specifying the percentage of all the remaining sessions that are terminated because the hit rate is low, and the percentage that are terminated because the hit rate is high. The model therefore depends on four parameters - two percentiles and two stopping percentages for each of the low-hit-rate and high-hit-rate ranges.

It is not hard to find values of these parameters for which, in this model, optional stopping increases the sequential statistic but not the permuted version. For example, suppose the stopping rule is applied from the 10th trial onwards and that the maximum session length is fixed at 20 trials. Suppose also that the ranges of low and high hit rates are defined to include the lowest and highest 10% of the distribution, and that the percentages of the remaining sessions terminated after each trial are specified as 0.7% in the low range and 0.35% in the high range (for a total termination rate of 1.05% per trial). Then the resulting sequential statistic per contributing pair is found to be 0.32046 - similar to the value observed in the experiments - but the permuted version is 0.31996, which is actually below the assumed value of 0.32.

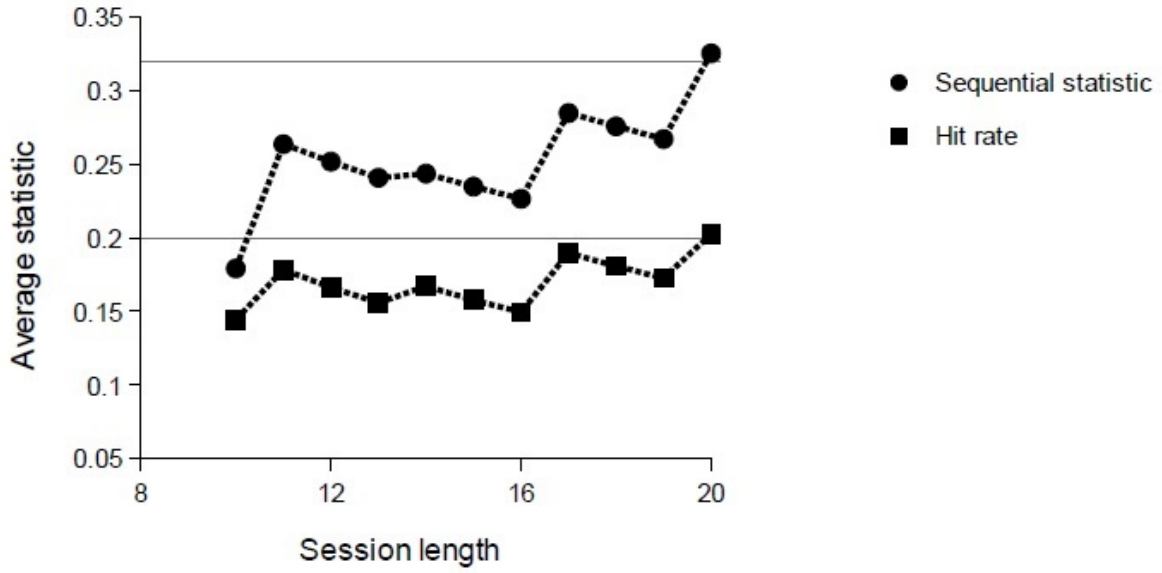

Figure. Average hit rate per trial (squares) and sequential statistic per contributing pair (circles), calculated as functions of session length.

With these parameters, about 10% of the sessions begun are terminated before the 20th trial, and the average hit rate per trial in all these terminated sessions is 0.16728 - below the unbiased value of 0.2, as observed in the experiments. If the average hit rate per trial and the average sequential statistic per contributing pair are plotted as functions of session length (see Figure), the behaviour is seen to be quantitatively similar to that observed in the experiments (as shown in Figure 7 of Radin, 2019). In particular, for all prematurely terminated trials the average hit rate per trial and sequential statistic per contributing pair are both lower than the unbiased values of 0.2 and 0.32 respectively, and as session length is varied, the values of the two statistics are strongly correlated.

This model outlined here is a very simple one, and in a real experiment the behaviour of human participants is bound to be much more complicated than it assumes, regarding both the distribution of guesses and decisions about terminating sessions. But it is sufficient to demonstrate that the reported observations could in principle be produced by optional stopping, and that the checks described in the report are insufficient to exclude that explanation. Indeed, the added complications of a real experiment will tend to make such checks less reliable, not more reliable.

5. Conclusion

The derivation in section 2 shows that optional stopping is capable of producing an artefactual increase in the sequential statistic per contributing pair of trials above the assumed value of $2p(1\text{-}p)$. The simple model in section 4 shows that under suitable assumptions this bias will not necessarily be accompanied by an increase in the permuted version of the sequential statistic, and that the hit rate per trial in prematurely terminated sessions will be smaller than p, as observed in the experiment. Therefore the arguments based on these features are not conclusive against optional stopping being responsible for the observed effect.

Of course, that is not to say that the effect necessarily is produced by optional stopping (and still less that the stopping rule used in the simple model accurately reflects the behaviour of the participants in the experiment). But clearly the checks performed on the experimental data so far are not sufficient to rule out optional stopping as an explanation.

Fortunately there is a straightforward way of settling the question. As discussed in section 2, if the individual sessions are first combined into one long sequence of trials, and a sequential statistic is then calculated for this sequence, it will remain unbiased by optional stopping, and its expected value will be $2p(1-p)$ per contributing pair of trials. The order in which the sessions are combined will not matter, provided it is not influenced by any information about their outcome.

(An alternative approach would be to use equation (8) above, which expresses the expected value of the sequential statistic per contributing pair, in terms of the average session length, the average number of sub-sequences present in each session and the average probability of success in the final trials of individual sub-sequences. This would allow the likely bias of the statistic owing to optional stopping, without the need to combine the individual sessions into a single long sequence.)

This test would introduce additional contributions from pairs of trials in different sessions. This might be expected to dilute any genuine effect that was present, but not to eliminate one as strong as was apparently observed. And if the result remained significant, this test would allow optional stopping to be definitely excluded as an explanation.

References

Dean Radin (2019). Tricking the Trickster: Evidence for Predicted Sequential Structure in a 19-Year Online Psi Experiment. *Journal of Scientific Exploration*, 33(4), 549-568.

Dean Radin (2018). Tricking the Trickster: Detecting Hidden Structure in Data from an 18-Year Online Psi Experiment. Paper presented at the *61st Annual Convention of the Parapsychological Association*. Retrieved from https://psyarxiv.com/9thae/ on 30 June 2021.