Notes on some of the studies registered with the Koestler Parapsychology Unit

Chris Phillips
21 February 2023

[The discussion of Study 8 has been revised. The original version of these notes can be seen here.]

[I am grateful to Patrizio Tressoldi and Julia Mossbridge for providing information about their registered studies, and to Jim Kennedy for comments on these notes.]

(1) Study 1

Registration:
Pupil dilation accuracy in the prediction of random events
Patrizio Tressoldi
Initial submission 26 November 2012
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_Registry_1001.pdf
Publication (Experiment 1):
Pupil dilation prediction of random events
P. E. Tressoldi, M. Martinelli and L. Semenzato
F1000Research 2:262 (2014)
https://doi.org/10.12688/f1000research.2-262.v2

In the first of two experiments of this study, two different stimuli are administered to a participant while a physiological variable is measured. The purpose is to test whether the physiological measurements can predict the type of a future stimulus. In order to define the prediction criterion, measurements from a series of trials are used, and an average z-value is calculated for each of the two stimulus types. For each experimental trial, a z-value is calculated and the prediction of stimulus type is based on comparison with the two average z-values.

In the pre-registered protocol, the average z-values are calculated in an initial "individual reactivity recording phase", separate from the series of experimental trials. But when the experiments were carried out, this phase was omitted, and the average z-values were calculated from the experimental trials themselves. As pointed out by a reviewer (Chris Baker), because the same trials are being used both to formulate the prediction criterion and to test its accuracy, this procedure is circular. (This can be seen most clearly for the case where there are only two trials, one of each type, and 100% accuracy is guaranteed.)

The published results showed both stimulus types were 'predicted' at levels well above chance (though the pre-registered expectation had been the opposite for the neutral stimulus). But for the reason explained above, this cannot be taken as evidence of a real effect.

[Marks excluded this experiment as confirmation of the hypothesis because of concerns expressed about the statistical analysis by the paper's reviewers (partly relating to the problem discussed above).]

(2) Study 2

Registration:
Pupil dilation prediction of random negative events. Can they be avoided?

Patrizio Tressoldi
Initial submission 1 February 2013
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_Registry_1002.pdf
Publication:
Does Psychophysiological Predictive Anticipatory Activity Predict Real or Future Probable Events?
Patrizio E. Tressoldi, Massimiliano Martinelli, Luca Semenzato and Alessandro Gonella
EXPLORE 11(2):109-117 (2015)
https://doi.org/10.1016/j.explore.2014.12.003
(Preprint: https://dx.doi.org/10.2139/ssrn.2371577)

This study is similar to the first experiment of Study 1, but only a single confirmatory hypothesis was pre-registered: that the prediction of negative stimuli would be better than chance. The pre-registered protocol, as in Study 1, specified an initial "individual reactivity recording phase".

The published results do show 'predictions' at levels well above chance but, as for Study 1, when the experiment was carried out the initial phase was omitted. Therefore this cannot be taken as evidence of a real effect.

[Marks excluded this experiment because it was presented as confirmation of a hypothesis claimed to have been supported by the results of Study 1 above, which he had excluded. He argued that if the previous positive results were excluded, further positive results for the same hypothesis could not be considered confirmatory. (However, in itself this argument is not sufficient, because the Registry's guidance indicates that even a test of a purely theoretical hypothesis can be accepted as confirmatory provided the analysis is well-designed.)]

(3) Study 4

Registration:
Does characteristic alpha EEG activity predict upcoming sensory events?
Julia Mossbridge
Initial submission 7 May 2013
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_Registry_1004.pdf
Report:
Results Report For the Koestler Parapsychology Unit Study Registry
Julia Mossbridge
Received 18 July 2013
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/Study_Results_1004.pdf
Subsequent publication (in which the results of the pre-registered experiment were combined with those of a previous experiment):
Characteristic Alpha Reflects Predictive Anticipatory Activity (PAA) in an Auditory-Visual Task
Julia A. Mossbridge
In D. Schmorrow and C. Fidopiastis (editors), Augmented Cognition. Neurocognition and Machine Learning. Proceedings of 11th International Conference, AC 2017. Springer, Cham.
https://doi.org/10.1007/978-3-319-58628-1_7

This study investigates whether future events can be predicted from EEG data. Two methods of analysis were pre-registered. A report of the results was submitted to the registry. This states that the second analysis method produced a non-significant result, but the discussion of the first is not as straightforward. Two data sets are referred to. The first relates to the previous experiment, while the second - the "replication data set" - is for the registered study. The second is is stated to have produced a p-value between 0.05 and 0.1. There is a suggestion that a significance level of 0.1 could be considered appropriate given that the results of the previous experiment were significant.

But a significance level of 0.05 is specified in the registration document, and accordingly in the summary of results it is acknowledged that the use of a significance level of 0.1 would be "an exploratory move". Therefore the results of both confirmatory tests must be considered negative.

It can also be noted that according to the verbal statement of the registered hypothesis, it is the left parietal electrodes that are expected to differentiate between future stimuli. In contrast, the summary of results states that it is the right frontal electrodes that were found to be important.

[N.B. A subsequent publication presented an analysis of the data from the pre-registered experiment, combined with that from the experiment that preceded the registration. The results of the pre-registered experiment are not considered separately, and the concluding section acknowledges that there is an exploratory step in the procedure and that the results need to be replicated independently. There is also a note that the pre-registered prediction was not confirmed.]

[Marks described the outcome of this study as unclear.]

(4) Study 6

Registration:
Evaluation of alterations of consciousness and the Model of Pragmatic Information in a ganzfeld protocol
Etzel Cardeña
Initial submission 23 August 2013
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_Registry_1006.pdf
Publication:
Changes in State of Consciousness and Psi in Ganzfeld and Hypnosis Conditions
E. Cardeña and D. Marcusson-Clavertz
Journal of Parapsychology 84(1):66-84 (2020)
https://doi.org/10.30891/jopar.2020.01.07

This study involved telepathy trials using either a Ganzfeld protocol or hypnosis. The pre-registration includes three confirmatory hypotheses. One is that psi scores will correlate with a measure of altered states of consciousness.

In the published results a correlation between psi scores and a measure of altered states of consciousness is found, and achieves a p-value of 0.009, but only for the Ganzfeld condition. The p-value for the hypnosis condition is p=0.75. No combined result is given, but from these p-values it seems unlikely that there would have been a significant correlation for the two conditions considered together. As the pre-registration did not specify that the conditions should be considered separately, the result of this confirmatory test must be considered negative.

[Marks did not evaluate this study because the link to the published paper on the KPU website was broken.]

(5) Study 8

Registration:
Brain-to-brain (mind-to-mind) interaction at distance: a proof of concept of mental telecommunication
Patrizio Tressoldi
Initial submission 23 April 2014

https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_Registry_1008.pdf
Publication:
Brain-to-Brain (mind-to-mind) interaction at distance: a confirmatory study
P. E. Tressoldi, L. Pederzoli, M. Bilucaglia, P. Caini, P. Fedele, A. Ferrini, S. Melloni, D. Richeldi, F. Richeldi and A. Accardo
F1000Research 3:182 (2014)
https://doi.org/10.12688/f1000research.4336.3

In this study, a sender is exposed to a random series of stimuli and the EEG of a remote receiver is recorded. For each session, a classification algorithm is trained by comparing 50% of the receiver's EEG signal to the stimulus sequence. The same classification algorithm is then applied to the whole of the receiver's EEG signal and is compared with the stimulus sequence to produce an accuracy score.

There are three problems with the analysis technique:
(1) The null hypothesis is that the expected percentage of agreements between the classifications produced by the algorithm and the measurements will be 50%, which does not appear to be correct given the protocol.
(2) The comparison includes the data that have been used to train the algorithm, which will tend to inflate the accuracy score.
(3) Even for the comparison with the data not used to train the algorithm, the accuracy score will tend to be inflated because of correlations between the EEG signals in adjacent time periods.

The first and third of these points were raised by a reviewer (Sam Schwarzkopf) when the results were submitted for publication. In response, the authors also applied a different method of analysis, in which the classification algorithm is trained by comparing the sender's EEG signal to the stimulus sequence, and then applied to the receiver's EEG signal to produce a predicted stimulus sequence. This was found to produce much less accurate results. In the concluding paragraph of the registration document, the lead author had acknowledged that the experimental design "does not preclude alternative explanations" and that further work would be needed.

[This discussion has been revised. Initially I had thought that the comparison excluded the data that had been used to train the classifier, and had discussed only the third of the problems described above. I am grateful to Jim Kennedy for correcting this misunderstanding, and to Patrizio Tressoldi for confirming what was done.]

[Marks excluded this study because of concerns expressed about the protocol and statistical analysis by one of the paper's reviewers (relating to the first and third problems discussed above).]

(6) Study 9

Registration:
Mind-matter interaction at distance on a random events generator (REG): a confirmatory study
Patrizio Tressoldi
Initial submission 15 May 2014
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_registry_1009.pdf
Publication:
Mind-Matter Interaction at a Distance of 190 km: Effects on a Random Event Generator Using a Cutoff Method
Patrizio Tressoldi, Luciano Pederzoli, Patrizio Caini, Alessandro Ferrini, Simone Melloni, Diana Richeldi, Florentina Richeldi and Gian Marco Duma

NeuroQuantology 12(3):337-343 (2014)
https://dx.doi.org/10.14704/nq.2014.12.3.767

In this study, participants attempt to influence the output of a random number generator. A session is described as successful if at any time the sum of the random numbers generated so far deviates sufficiently from chance expectation. The criterion for deviation is a z-score whose absolute value is greater than 1.65, corresponding to a two-tailed test with significance level 0.1, based on a normal approximation for the sum.

The published results show a large difference between the percentage of successful sessions with participants, and the percentage of successful control sessions. However, if the expected percentage of successful sessions in the absence of any psi effect is calculated numerically, using the specified operating parameters for the random number generator, it is found to be significantly larger than was measured in the control sessions. Specifically, in the pilot study, with 2 samples per second of 10 bits each, and with an average duration of 87 seconds (therefore 174 samples), the calculated expected percentage of sessions in which the significance criterion should be met is 64.4%, but in the experiment it was met only in 48% of the 50 control sessions performed. In the confirmatory experiment, with one sample per second of 200 bits, and with an average duration of 62 seconds, the calculated expected percentage is 54.2%, but the experimental result was only 13.7% of 102 control sessions (note that in theory the figure of 13.7% would have been expected to be achieved after only 2 seconds).

For some reason, the operation of the random number generator seems to have been seriously inconsistent with its specification. As the reason for this is not known, the findings must be considered doubtful.

[Marks excluded this study because almost one third of the data had already been collected at the time of registration, and also because he had concerns about the speed with which the published paper was reviewed and the quality of the description of the results. (Though it should be noted that for a number of other registered studies, the results were presented as a brief report to the Registry rather than in a peer-reviewed paper.)]

(7) Study 10

Registration:
Biophotons as physical correlates of mental interaction at distance: a confirmatory study
Patrizio Tressoldi and John G. Kruth
Initial submission 8 October 2014
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_registry_1010.pdf
Publication:
Can our Mind emit light? Mental entanglement at distance with a photomultiplier
Patrizio Tressoldi, Luciano Pederzoli, Alessandro Ferrini, Marzio Matteoli, Simone Melloni and John G. Kruth
Preprint posted 2 July 2015; revised 8 August 2015
https://dx.doi.org/10.2139/ssrn.2625527

[This study did not produce any positive confirmatory results (though in an early version of Marks's evaluation it was incorrectly indicated that it did).]

(8) Study 11

Registration:
CardioAlert: a portable assistant for the choice between negative or positive random events
Patrizio Tressoldi
Initial submission 17 March 2015
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_Registry_1011.pdf
Publication:
CardioAlert: A Heart Rate Based Decision Support System for Improving Choices Related to Negative or Positive Future Events
Patrizio Tressoldi, Massimiliano Martinelli, Jacopo Torre, Sara Zanette and Gian Marco Duma
Preprint posted 10 May 2015; revised 1 August 2015
https://dx.doi.org/10.2139/ssrn.2604206

In this study the participant has to guess the location of a target, in one series of trials using only intuition, and in a further two series with the assistance of a device that makes predictions based on heart rate. The confirmatory hypothesis is that the success rate will be higher in the second of the two series using the device than in the initial series using only intuition.

According to the published results, it was found that if the two series using the device were combined, the success rate was significantly higher than chance. But the pre-registered confirmatory hypothesis, based on a comparison between the second series using the device and the initial series using only intuition, was not confrrmed. The success rate using intuition was 52.3% and the success rate in the second series using the device was 52.4% (Figure 4), and the difference between these success rates is not statistically significant.

[Marks excluded this study because one fifth of the data had already been collected at the time of registration, and also because the results were published "in a journal with questionable peer-reviewing procedures" (In fact the publication was a preprint rather than a peer-reviewed paper. But, as above, it should be noted that for a number of other registered studies, the results were presented as a brief report to the Registry rather than in a peer-reviewed paper.)]

(9) Study 12

Registration:
Biophotons as physical correlates of mental interaction at distance: a new confirmatory study
Patrizio Tressoldi and John G. Kruth
Initial submission 29 April 2015
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_registry_1012.pdf
Publication:
Can our Mind emit light? Mental entanglement at distance with a photomultiplier
Patrizio Tressoldi, Luciano Pederzoli, Alessandro Ferrini, Marzio Matteoli, Simone Melloni and John G. Kruth
Preprint posted 2 July 2015; revised 8 August 2015
https://dx.doi.org/10.2139/ssrn.2625527

[This study did not produce any positive confirmatory results (though in an early version of Marks's evaluation it was incorrectly indicated that it did).]

(10) Study 13

Registration:
Can our Mind emit light? A confirmatory experiment of Mental interaction at distance on a photomultiplier
Patrizio Tressoldi and John G. Kruth
Initial submission 6 July 2015
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_registry_1013.pdf
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_1013_Errata.pdf
Publication:
Can Our Minds Emit Light at 7300 km Distance? A Pre-Registered Confirmatory Experiment of Mental Entanglement with a Photomultiplier
Patrizio Tressoldi, Luciano Pederzoli, Marzio Matteoli, Elena Prati and John G. Kruth
NeuroQuantology 14(3):447-455 (2016)
https://dx.doi.org/10.14704/nq.2016.14.3.906

In this experiment a photomultiplier is used to measure photons during three 40-minute periods, designated as (1) MI (mental intention) and post-MI, (2) pre-MI and (3) control. The number of photons detected in each half-second period is counted, and if the number is 11 or more it is classified as a burst. One of two pre-registered confirmatory hypotheses concerns the numbers of photons that are detected in such bursts, which for convenience can be called "burst photons". The hypothesis is that, considering the number of burst photons in each period as a percentage of the total number over all three periods, the percentage in period 1 will be higher than the percentages in periods 2 and 3. The registration document specifies that these differences will be analysed (i) by using "the z-test of differences between dependent-samples proportions" to calculate confidence intervals and (ii) by calculating Bayes Factors using an 'applet' due to Morey (2014).

In the publication of results, a confirmation of this hypothesis is claimed, based on confidence intervals calculated "using *probit* method estimation", and Bayes Factors calculated using an updated version of Morey's applet.

Apparently, both these calculation methods would require the variables being compared to be independent and binomially distributed. As the variables from the three periods are percentages constrained to add up to 100, they are clearly not independent and binomially distributed. There seems to be no reason to expect even the total numbers of burst photons in the three periods to be binomially distributed. As the percentages must add up to 100, each pair of variables will tend to be negatively correlated, and an analysis method that assumes they are independent will tend to overestimate the statistical significance of any differences between them. Therefore it is not safe to accept the finding of significance based on these methods of analysis.

[Marks excluded this study on the basis that the hypothesis tested had been changed after registration. (Although the presentation could be clearer, the data relating to the pre-registered hypotheses are given in the second column of Table 3 and the second row of Table 4 of the published paper. But, as argued above, the method of analysis used to test the first hypothesis is not valid.)]

(13) Study 26

Registration:
A test of reward contingent precall
David Vernon
Initial submission 19 October 2016

https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_Registry_1026.pdf
Publication:
A Test of Reward Contingent Recall
David J. Vernon
Journal of Parapsychology 82(1):8-23 (2018)
http://doi.org/10.30891/jopar.2018.01.02
(Report: https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/Study_Results_1026.pdf)

One of two confirmatory tests registered was positive. A two-tailed test produced a p-value of 0.021 (and the effect was in the expected direction).

[Marks accepted this as a partial confirmation, as one of the two tests was positive.]

(14) Study 46

Registration:
An implicit and explicit assessment of morphic resonance theory using Chinese characters
David Vernon
Initial submission 23 May 2018
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_Registry_1046.pdf
Publication:
An Implicit and Explicit Assessment of Morphic Resonance Theory Using Chinese Characters
David Vernon, Glenn Hitchman and Chris Roe
Journal of the Society for Psychical Research 85(3):129-144 (2021)
(Report: https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_1046_Summary.pdf)

In this study, participants without knowledge of Chinese are shown pairs of apparent Chinese characters, one real and one a decoy, and asked either which they prefer or to identify which is real. Two confirmatory hypotheses were pre-registered - that real characters would be preferred to decoys, and that real characters would be identified as real more often than decoys. Despite being directional hypotheses, the pre-registered tests are two-tailed.

In the statement of results, one of the tests (for preference) produced a significant result, though the effect was in the opposite direction to that expected. In terms of the specified two-tailed test, this must be considered as a positive result.

Unfortunately there is a serious difficulty with the protocol of this study, because it is impossible to guarantee that the decoy characters are a match for the real characters in all the respects which may influence participants' preferences. This means that a positive result may reflect only a failure to match the aesthetic appeal of the real characters when constructing the decoys (in this case the reported result could be explained by the decoys having been more, rather than less, aesthetically appealing than the real characters). Consequently, the result cannot be considered an indication of a psi effect, as it is capable of being explained in conventional terms.

There is a similar difficulty with the test for the other confirmatory hypothesis in this study - that real characters would be identified as real more often than decoys - because it is impossible to guarantee that the decoy characters are a match for the real characters in all the respects which may influence this choice.

[Marks excluded the positive result in this study because the verbal statement of the first confirmatory hypothesis was

directional: "Participants will prefer (i.e., select) real Chinese characters at a level significantly greater than chance (i.e., 50%)." In common with a number of other registered studies in which the verbal statement of the hypothesis is directional but a two-tailed test is specified, this does present a problematic inconsistency. In the context of a pre-registered study, there is no doubt that the most appropriate tests for directional effects are one-tailed. But as an important purpose of pre-registration is to minimise the freedom of the researcher (or others) to make decisions about analysis after the results are known, it seems preferable to use the pre-registered statistical test.]

(15) Study 49

Registration:
Mind-matter interaction at distance on a standalone device
Patrizio Tressoldi and Luciano Pederzoli
Initial submission 21 February 2019
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_Registry_1049.pdf
Publication:
Mind Control of a Distant Electronic Device: A Proof-of-Concept Pre-Registered Study
Patrizio Tressoldi, Luciano Pederzoli, Elena Prati and Luca Semenzato
Journal of Scientific Exploration 34(2):233–245 (2020)
https://doi.org/10.31275/20201573

In this study, participants attempt to influence a random number generator for a 15-minute period, and its output during this period is compared with that during earlier and later periods. The data from each one-minute period are analysed using two tests of randomness. The confirmatory hypothesis is that the percentage of minutes for which at least one of the tests indicates a departure from randomness, will be larger during the period of influence than in both the other periods.

In the published paper the method of analysis is different. Instead of a comparison being made between the numbers of minutes in the different periods for which at least one of the criteria is met, three separate comparisons are made, between the numbers of minutes in the different periods for which (1) the first criterion is met, (2) the second criterion is met and (3) both criteria are met.

It is not clear whether the pre-registered confirmatory analysis would have produced significant results, so in assessing the overall success rate this study should be excluded from consideration.

(16) Study 56

Registration:
Three confirmatory analyses of precognition and micro-pk data gathered using online methods
Julia Mossbridge and Dean Radin
Initial submission 29 April 2020
https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_Registry_1056.pdf
Publication (Experiment 1):
Psi Performance as a Function of Demographic and Personality Factors in Smartphone-Based Tests: Using a “SEARCH” Approach
Julia Mossbridge and Dean Radin
Journal of Anomalous Experience and Cognition 1(1-2):78-113 (2021)
https://doi.org/10.31156/jaex.23419
Publication (Experiment 2):
State, trait, and target parameters associated with accuracy in two online tests of precognitive remote viewing
Julia Mossbridge, Kirsten Cameron and Mark Boccuzzi

Preprint posted 28 October 2022
https://doi.org/10.31234/osf.io/9zjus

This study contains two experiments using games played on a smartphone.

Experiment 1:

The first experiment concerns psychokinesis. Participants play a series of games called Heart Quest. In each game, two reference bits are randomly generated and then there is a series of trials in which further trial bits are generated and compared with the reference bits. (A complete game has ten trials, though this does not seem to be stated in the registration document.) The pre-registered predictions stated in words are that the reference bits will contain more zeroes than expected by chance and that this effect will be stronger in women than in men. Although the predictions are directional, two-tailed tests are specified.

In the published results, the first analysis presented does show more zeroes than expected by chance for each of the two reference bits and for both men and women, though contrary to expectation the overall percentage of zeroes is larger for men than for women. The excess of zeroes is described as statistically significant in every case, but there is a serious problem with the analysis. The percentage of zeroes is compared with a binomial distribution, but it is a distribution based on the number of trials, not the number of games. Because the reference bits are generated only once per game, not once per trial, this is not appropriate. Since the number of trials is much larger than the number of games, this means that the likelihood of the deviation occurring by chance is severely underestimated, so that statistical significance tends to be overestimated.

The authors acknowledge that this analysis method "might provide a false impression of reference bit consistency" but state that they did not recognise the problem until the results had been analysed, when it was too late to pre-register an alternative method. (It is not explicitly stated in the registration document that the binomial distribution will be based on the number of trials rather than games, though there is a reference in the power estimation section to the number of trials required. The analysis of the preliminary study had not been published at the time of registration.) But for this reason they also present the results of an analysis based on the number of games, and this shows that none of the deviations from chance relevant to the confirmatory hypotheses is statistically significant. This is clearly the correct way to analyse the reference bits, so it must be concluded that the data do not confirm the hypothesis.

Experiment 2:

This is a precognitive remote viewing study in which participants are shown graphs describing features of two unseen images, and have to choose which they think will be randomly selected as the target. The hypothesis concerns a correlation between the success rate and how interesting the target image is. For each image used, trials are considered in which the image in question is presented in the second of the two positions and is randomly selected as the target, and the overall success rate of participants' choices is calculated. Two sets of images, one with high success rates and the other with low success rates, are then independently rated for interestingness. A highly significant difference between the ratings for the two sets is reported.

There are two serious problems with this experiment:

(1) If the images were visible to the participants, and if there were a tendency for them to choose the more interesting image, then because the only trials considered are those in which the image being

assessed is the target, interesting images would tend to be associated with a higher success rate. The images are not visible to the participants, but graphs representing their properties are. If these graphs include features that are positively correlated with interestingness, and that are attractive to participants, that will tend to produce a false positive result. The authors are aware of this but argue that it is not a problem, on the basis of additional post hoc analysis of particular subsets of the image set whose graphs share common features. But such analysis is not capable of excluding the possibility of such an effect (and in fact that possibility is supported by the relative numbers of graphs with features deemed to be attractive and unattractive that are found in the sets with high and low success rates).

(2) The results of two different methods of statistical analysis are reported, both using t-tests to compare the interestingness of the images in the sets with high and low success rates. Both employ the same ratings of interestingness, which are defined by comparisons between pairs of images, one from each set. Because the interestingness of images in each set is defined by comparison with those in the other set, the ratings in the two sets are not statistically independent. This means that the first method, in which t-test is used to compare the average ratings of the two sets, is not valid because the required assumption of independence is not satisfied (in fact, by definition the average ratings of the two sets are equal and opposite). In the second method, there is an even more serious problem of non-independence, because each pairwise comparison is treated as an independent variable, despite the fact that the number of pairwise comparisons is much larger than the total number of images that are being compared. Both these tests will tend to overestimate the significance of any differences between the two sets, and for the second method the overestimate may be very large.

Because of these problems, the results reported cannot be taken as confirmation of the hypothesis.

[Marks was evidently unaware of the published results of Experiment 1. The preprint containing the results of Experiment 2 appeared after Marks posted his evaluation.]