Short title: Using Online Tasks to Test the Robustness of Intuitive Abilities

Full title: Testing the robustness of accurate intuitive abilities and assessment of reproducibility with a group of potentially talented individuals

Authors: Arnaud Delorme[1,2,*], Helané Wahbeh[1,3,4], Dean Radin[1,3]

Affiliations:

1. Institute of Noetic Sciences, 101 San Antonio Rd, Petaluma CA 94952
2. SCCN, INC, University of California, San Diego, 9500 Gilman Drive, La Jolla CA92093-0559
3. California Institute of Integral Studies, 1453 Mission St, San Francisco, CA 94103
4. Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97239

* corresponding author

Authors contribution

- Arnaud Delorme, Conceptualization, Data curation, Formal Analysis, Funding Acquisition, Methodology, Software, Writing – Original Draft Preparation
- Dean Radin, Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Writing – Review & Editing
- Helané Wahbeh, Conceptualization, Funding Acquisition, Methodology, Writing – Review & Editing

Abstract

In this study, we tested whether it was possible to use an online psi task to identify psi-talented individuals who could produce reliably significant results in repeated tests. The study aimed to develop psi tasks on an online platform, analyze data to identify talented individuals among a large sample (n=1014); 3) select the top 50 performing individuals (talents), assess their consistent performance, and evaluate potential personality predictors. We found that retest performance on two tasks was above chance expectations, but this finding did not hold up after correcting for multiple comparisons. In selected analyses, we also found that conscientiousness had a negative correlation with higher performance, while agreeableness had a positive correlation. We suggest that increasing the number of participants could yield more robust results.

Keywords: intuition, psi, reproducibility, talent, online experience.

Introduction

A key practical challenge faced by scientific research on psychic (psi) phenomena includes the small effect sizes typically observed in laboratory studies, as well as reliable ways of replicating these effects. This, in turn, requires large sample sizes to study the purported effects. Psi talent, if real, appears relatively rare in the general population; thus, testing unselected individuals is often not successful in reliably detecting psi effects. Rhine's early card guessing studies focused on identifying individuals who consistently exhibited high levels of extrasensory perception (ESP) abilities, often referred to as "star" subjects, as these individuals were believed to possess extraordinary psychic capabilities (Rhine, 1934). More recently, the US government's formerly classified program of psychic espionage, which ran from 1972 to 1995, was designed to use psychics to collect information of use to military, law enforcement, and intelligence applications. That program relied on a small group of psi talents (May and Marwaha, 2018; Targ and Puthoff, 1974), individuals such as Joseph McMoneagle, Ingo Swann, Sean Harribance, Matthew Manning, Pavel Stepanek, and several others are examples of apparently talented individuals who have been extensively studied in laboratory settings. These individuals produced results indicating that their skills were reliable enough to be useful for practical purposes (May and Marwaha, 2018; Targ & Puthoff, 1974). This suggests that rather than relying on large sample sizes of unselected participants, another approach would be to identify small samples of psi-talented individuals (talents) who may be more likely to produce statistically robust results.

Efforts to locate talents in the past have relied on in-person testing groups of unselected people, selecting subsets who appeared to perform well on psi psychophysics tasks, and then testing and retesting those groups to find a handful who appeared to repeatedly perform beyond chance (May and Marwaha, 2018). Those methods worked to some degree but were inefficient because of the time and effort involved in dealing with large numbers of candidates. Today, the Internet can be used to locate such individuals. Web-based studies offer a number of significant advantages over more traditional laboratory tests. Candidates can be recruited and easily tested at a low cost, and the tests can be provided automatically to practically anyone at any time. In addition, participants can take these tests from the comfort of their own homes. Although not as controlled as in laboratory settings, some types of online tests have been shown to be a viable method for conducting psychological and psychophysics experiments (Kim et al., 2019).

The first known use of computer networks to conduct psi experiments was in 1972 and published more than a decade later by Vallee (1988), an internet pioneer who conducted a remote viewing study using a forerunner of the Internet (the ARPAnet). In 1977, Radin used the PLATO computer-based education network at the University of Illinois to provide public-access precognition tests (Johnson, 1977). Some twenty years later, other public-access Web-based psi experiments were reported by Bierman (1995) and Rebman et al. (1996). Starting in 2000, Radin developed and launched a suite of web-based psi experiments dubbed "GotPsi," the name of which was based on a popular milk commercial at the time, which featured the phrase, "Got milk?" By 2020, over 250 million individual trials were collected from more than 300,000 people around the world, yielding significant results for some of the tests (Radin, 2019). The GotPsi website and the other online experiments could have potentially been used to identify psi talents. However, those tests were not designed for that purpose, so contact information was not systematically collected, making it impossible to contact individuals who appeared to be performing well. The current project used modern online testing techniques to identify people with potential psi talent and to evaluate predictor variables associated with such individuals (Cardeña et al., 2015).

Most of the psychophysics tasks used in previous online experiments as well as the current manuscript pertain to remote viewing, a modern euphemism for clairvoyance, which refers to the ability to perceive or describe details about target object distant in space or time. In addition to Rhine's early card-guessing studies (Rhine, 1934), investigations into remote viewing gained momentum in the 1970s and 1980s, with the involvement of U.S. government agencies, which sought to explore its potential applications for intelligence and military purposes as mentioned previously (Targ and Puthoff, 1974; Utts, 2019). Despite claims of methodological flaws and biases by skeptics (Hyman, 1988; Galak et al., 2012), more recent controlled experiments have reported statistically significant results (Roe et al., 2020; Schwartz, 2020).

The objectives of this study were to 1) develop and integrate various psi psychophysics tasks into a single online platform; 2) develop data analysis methods to identify talented individuals in a large sample of volunteers; 3) Select the 50 top-performing talents; 4) assess consistent above-chance performance in test-retest scores of talents; and 5) evaluate potential predictors for talented performance.

Methods

Participants

We intended to recruit between 500 and 2,000 volunteers to perform psi psychophysics tasks over an 8-month period. The 50 participants with the highest scores would be identified after 8 months or after data from 2,000 participants had been collected (whichever came first). If data from fewer than 500 individuals had not been collected by 8 months, the experiment would be extended until a minimum of 500 participants was reached. The 50 selected potentially talented people were compensated $20 each to perform the test again. Among those 50 selected individuals, if some did not respond or declined to participate, the next top performers were considered for inclusion until 50 participants completed all the tasks and were included in a second round of tests. After eight months, 1,014 participants performed all the tasks and qualified to be included in the first phase (8,112 participants performed some of the tasks). Then, 50 of the 200 top performers were tested again – we first recontacted the top 50 participants and increased that number (up to 200) until 50 participants redid all the tasks. This study was approved by the Institutional Review Board at the Institute of Noetic Sciences (IONS; IORG#0003743) and preregistered on the University of Edinburgh's KPU website (https://koestlerunit.wordpress.com; KPU Registry 1070).

Volunteers were recruited through the IONS mailing list (60,000+ individuals), private mailing lists (10,000+ individuals), various social media accounts (100,000+ followers), and *viva voce* by those who had participated. The IONS marketing teams assisted in posting announcements on various external media and recruiting members of the 300+ IONS community groups worldwide.

Psychometric measures

Participants completed the IONS Discovery Lab (IDL) questionnaire (Wahbeh et al., 2022) on SurveyMonkey.com, a 15-minute questionnaire that includes items about demographics, interconnectedness, perceived extended human capacities, creativity, transformation, and well-being. These questions were taken from validated instruments, including the Arizona Integrative Outcomes Scale (Bell et al., 2004), the Single General Self-Rated Health Question (DeSalvo et al., 2006), the Sleep Quality Scale (Cappelleri et al., 2009), the Numeric Pain Rating Scale (Gurefe et al., 1998), a two-question questionnaire on physical activity (Johansson et al., 2008), the Brief Five Inventory 10 item personality questionnaire (Rammstedt et al., 2007), the Positive and Negative Affect Scale (Watson et al., 1988), the Cloninger Self-transcendence subscale (Cloninger et al., 1997), Interconnectedness with Nature and Others (Aron et al., 1992; Schultz et al., 2004), and the Noetic Beliefs and Experiences scale (Wahbeh et al., 2020).

We tested the following predictors of psi performance: gender, meditation status (binary), paranormal belief and experience (2 items), personality (5 items), and self-transcendence (1 item). The other personality traits were not included in the pre-registration document, so they were not included in our analyses. However, the data from these questionnaires may be used in future publications.

Psychophysics tasks

Task 1 is a photo guessing task - also known as a "quick remote viewing" task. This is a forced-choice photo guessing task where the participant attempts to guess a randomly selected future image from five possible choices. For this task, a blank frame is displayed in the center of the screen, and five photos out of a pool of 130 images (taken from the Corel picture database), some depicting individuals and other nature or urban scenes, are displayed below it. The participant chooses which one of the five images they think will appear in the blank frame. After they select an image, the target picture is randomly selected by the web server that hosts these tasks using a Javascript-based pseudorandom number generator and is shown in the blank frame. Thus, the participant is able to see if their choice was correct or not, and the following feedback is displayed: "That was a hit" or "That was a miss." The participant then presses a button to continue to the next trial. The participant completes 20 such trials. At the end of the task, the participant is shown the percentage of "hits" achieved. A *z*-score is assigned to each 20-trial experiment. The z-score is calculated as follows:

$$Z = \frac{(h-np)}{\sqrt{np(1-p)}} \qquad (1)$$

where n is the number of trials (20 in this case), p is the probability of hitting a target (0.2), h is the number of hits or correct responses, np is the mean hit rate, and $\sqrt{np(1-p)}$ is the standard deviation of a binomial distribution.

Task 2 is a free-response photo guessing task - also known as the "long remote viewing" task. Participants are shown a blank frame and are asked to imagine what image will appear there. For the image they have in mind, they select the presence or absence (i.e., binary choice) of the following 6 shapes (Figure 1, task 2): (1) rounded, arcs, (2) angular, squares, (3) linear, flat, (4) points, edges, (5) soft, cloudy, liquid, (6) repeated patterns. They also rate the relative presence of the following 9 descriptors on a 5-point Likert scale: (1) water, (2) people, (3) plants, (4) food, (5) animals, (6) temperature (cold to hot), (7) scene type (artificial to natural), (8) movement (from still to dynamic) and (9) location (indoor to outdoor). Then, they are allowed to enter several keywords to describe the image, after which they press a "submit" button. The web server then randomly selects a target image (out of 217 possibilities) and shows it, along with a score from 0 to 100.

The score is calculated by comparing the participant's responses to the average responses of three judges, each of whom was experienced in evaluating similar binary and descriptor-based remote viewing trials and who had earlier examined each of the 217 images in the target pool. For each image, the judges filled out the 15 descriptors while they were visually inspecting each image. For the six binary measures, the participant's response (0 or 1) was subtracted from the response average that the judges provided for the same image, multiplied by 2 (so it is similar in scale compared to the other Likert scale measures), and then the absolute value was taken. For the Likert scale measures, the difference between the responses (0 to 4) was subtracted from the response average provided by the judges, and then the absolute value was taken. If the result was larger than 2, it was capped at 2. The sum of all these values was added to obtain a raw score for the target image.

To assess if this raw score was above chance expectation, we calculated the raw score the participant would have obtained on each of the remaining 216 non-target images. The position of the target-image raw score in the distribution of non-target image raw scores was then calculated. The percentage of images where the target-image raw score exceeded chance expectations constituted the final participant score. For example, if the participant's descriptors for their image resulted in a raw score higher than all of the other images, then the final score was 100. If the raw score for the target image was lower than all other images, the final score was 0. Other resulting final scores were arranged proportionally, so the final score was limited to the range from 0 to 100.

Any keywords that the participant entered were not considered for the present study but may be used for subsequent analyses. The final descriptor score was shown to the participant along with the number of correctly guessed words. $1 - score/100$ is a number from 0 to 1 that represents the likelihood of producing an answer that was close to the target image. After each trial, a *z*-score was then calculated by taking the inverse of the normal cumulative distribution (Abramowitz and Stegun, 1964).

$$Z = normInv(1 - score/100) \qquad (2)$$

Tasks 3, 4, and 5 are forced-choice card guessing tasks. These include three variations of card-guessing tasks, where participants attempt to guess which card will be selected by the webserver. In all of these tasks, five cards are initially shown "face down."

- The first card test was to attempt to locate the position of one of five cards that, when "turned over," would contain an image. After the participant guessed, the position of the actual target was displayed. Participants were shown the hit rate after each trial, and at the end of a series of 10 trials, their overall hit rate was shown along with the odds against chance (1 out of *n*) of obtaining such a result. Note that showing the hit rate after each trial did not affect performance unless the participant could guess the outcome of the next trial (precognition). A *z*-score was also calculated using equation (1) with $n = 10$ and $p = 0.2$.
- The second card test was the same as the first but performed sequentially, such that the participants could keep selecting cards until they found the target. After finding the target, the participant was shown the average number of steps to hit a target. For

example, if they always picked the target on their first attempt, then the average number of steps was 1. At the end of a series of 10 trials, their average number of steps was shown along with the odds against chance (1 out of *n*) of obtaining such a result. The likelihood of each step was equiprobable and equal to 0.2: the likelihood of 1 step was equal to the likelihood of picking the target in 5 images and equal to 0.2; the likelihood of 2 steps was equal to the likelihood of choosing a non-target in 5 images (0.8 probability) and choosing a target in 4 images (0.25 probability) so 0.8*0.25=0.2, etc. Hence, the number of average steps was 3. The standard deviation was constant and equal to

$$\sigma = \sqrt{\frac{\sum_{1}^{n} (x_i - 3)^2}{n}} = 2\sqrt{\frac{\frac{n}{5}2^2}{n}} + 2\sqrt{\frac{\frac{n}{5}1^2}{n}} = \sqrt{2}$$

allowing the computation of a z-score as follows:

$$z = \frac{3 - s/n}{\sqrt{2}} \qquad (3)$$

where n is the number of trials (10 here), 3 is the average number of chance-expected steps and s is the total number of steps across all trials.

- For the third card test, a target image was shown above 5 face-down cards, and participants were asked to try to *force* that image to appear on each of the five cards, which they selected one at a time. The probability of the target image showing up on any individual card was 50 percent. They performed 10 trials. For each trial, the number of hits (out of 5 possible) was shown. The total hit rate for all trials was also displayed. A *z*-score was calculated for each trial as in equation 1 with n as the number of images (5 in this case), p as the probability of an image appearing (0.5), and h as the actual number of images appearing. The *z*-score was combined across trials using Stouffer Z formula ($k = 10$ trials).

$$Z = \frac{\sum_{i=1}^{k} z_i}{\sqrt{k}} \qquad (4)$$

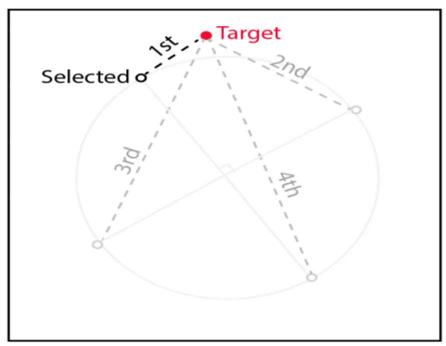

Figure 1. The distance of the selected point with the

Task 6 is a location guessing task - similar to a dowsing task. This task explored participants' ability to locate a randomly selected future target. They were asked to select a location in a 250 x 250-pixel blank square where they thought the target was hidden. After they selected a spot, the location where they clicked was shown as a small blue square, and a randomly selected target location was shown as a small red dot. A line was drawn between the two locations. To score the results, first, the Euclidean distance between the user's *x, y* selection and the computer's *x, y* selection in the 250 x 250 target area was calculated. For example, if the user selected the point [22, 20] and then the computer randomly chose a target at [30, 35], the distance between those two points was calculated as follows $D = \sqrt{(22-30)^2 + (20-35)^2} = 17$. We used two statistics to calculate the *z*-score, one for Phase 1 and an improved one for Phase 2. For the first statistic (Phase 1), the distance between the target and all of the other possible places that the user could have chosen was calculated. Because of the 250 x 250 target area, there were 90,000 possible places where the user could select or a 1/90,000 chance of getting the target area exactly correct. The user's answer was thus evaluated in relation to all of the other possible distances. The web server then calculated the exact probability of selecting a point that was that close to the target and converted it to odds against chance, which was shown to participants after each trial. The probability was also converted to a *z*-score using the *normInv* method (equation 2). A Stouffer's *Z* (equation 4) was used to combine the individual *z*-score over multiple trials. The Stouffer Z was then converted back to odds against chance and shown as feedback to participants. Participants performed 20 trials of this task.

We realized after Phase 1 of the experiment that this method might be biased because people do not select locations at random. Clicks in the center would generate a higher *z*-score than clicks at the periphery. This is because clicks in the center are closer to all possible target locations than clicks at the periphery. People tended to select the center of the space, therefore inflating their z-scores. In the result for section 1, the *z*-scores are shown but were not used to calculate individual performance. In Phase 2, we recomputed *z*-scores using a second method. We compared the distance between the selected point and the target point with three other possible locations, which correspond to the selected point rotated by 90, 180, and 270 degrees with respect to the square (Figure 1). The rank of the selected point (which may be 1, 2, 3, or 4) was then converted to a z-score using the formula:

$$z = (rank - 2.5)/\sigma \text{ where } \sigma = \sqrt{\frac{(1-2.5)^2+(2-2.5)^2+(3-2.5)^2+(4-2.5)^2}{4}} \quad (4)$$

This method ensures that the location selected by participants is compared to locations at the same distance compared to the center. A Stouffer's *Z* (equation 4) was then used to combine the individual z-score over multiple trials.

Task 7 is a lottery task. This task simulates the California Super Lotto Plus. Participants selected five numbers in the range 1 to 47 and 1 "mega" number between 1 and 25. The web server performed a random drawing, and the number of matching numbers (5 numbers and mega numbers combined) was then shown to participants. Following Füredi et al. (1996) and verifying the formulas with numerical simulations, the probability of obtaining at least a given number of hits *h* with or without the mega number was estimated as follows:

$$p(hit \geq h \,\&\, mega\ = 0) = \sum_{k=h}^{5} \frac{(5\ k\ )(42\ 5-k\ )}{(47\ 5\ )}$$

$$p(hit \geq h \,\&\, mega\ = 1) = \frac{1}{25}\sum_{k=h}^{5} \frac{(5\ k\ )(42\ 5-k\ )}{(47\ 5\ )}$$

For example, if the participant had one number correct, the probability of obtaining at least 1 number correct was 0.44 (top formula above) if it was not the mega, and 1/25 = 0.04 if it was the mega (bottom formula).

Task 8 is a psychokinetic task, dubbed the bubble task. The participant saw small bubbles randomly moving around a bounded area of the screen. The task was to focus attention on the bubbles while holding the intention that the bubbles would coalesce into a circle. While this method allows for the collection of a large psychokinetic dataset, which can provide valuable insights, we acknowledge that it may not fully replicate the conditions of more personalized or hands-on testing.

The movement of the bubbles was linked to numbers being produced by a Javascript-based pseudorandom number generator (PRNG) based on the KISS07 algorithm (Marsaglia and Zaman, 1993). That algorithm was re-seeded at each call with the Javascript *performance.now()* window function, which returns the time it takes to display the page in milliseconds (and depends on the local computer load, so the time was quasi-random). Two random numbers were drawn every second and converted to a Gaussian random number using the Box and Muller equation (Box and Muller, 1958). The random number was then used to influence bubbles on the screen. A smoothing function was used to avoid sudden transitions in the bubbles' positions.

If the participant could mentally affect the random number generator (or possibly the random number seed) to cause the average output to deviate from chance, the bubbles would coalesce into a circle shape more often than expected by chance. The larger the deviation, the larger the presumed psychokinetic "influence." The task began with 8 seconds of baseline recordings where no intention was applied (and no bubbles were shown), then the participant concentrated their intention on the bubbles for 16 seconds, and then they relaxed again for 8 seconds (with no bubbles shown). During the concentration period, a relaxing sound was played, where the volume of the sound was controlled by the degree to which the bubbles were coalescing. More coalescing increased the sound volume. During the baseline period, a countdown clock and blank screen were shown. The scoring was calculated by applying an unpaired t-test (with unpooled variance estimates) comparing the collection of pseudorandom numbers generated during the baseline period (16 numbers – twice 8 numbers – at a rate of 1 number per second) versus the values generated during the focused attention period (16 numbers as well). This *p-value* (two-tailed) was converted to a z-score by calculating the normal inverse (*normInv)* of that probability.

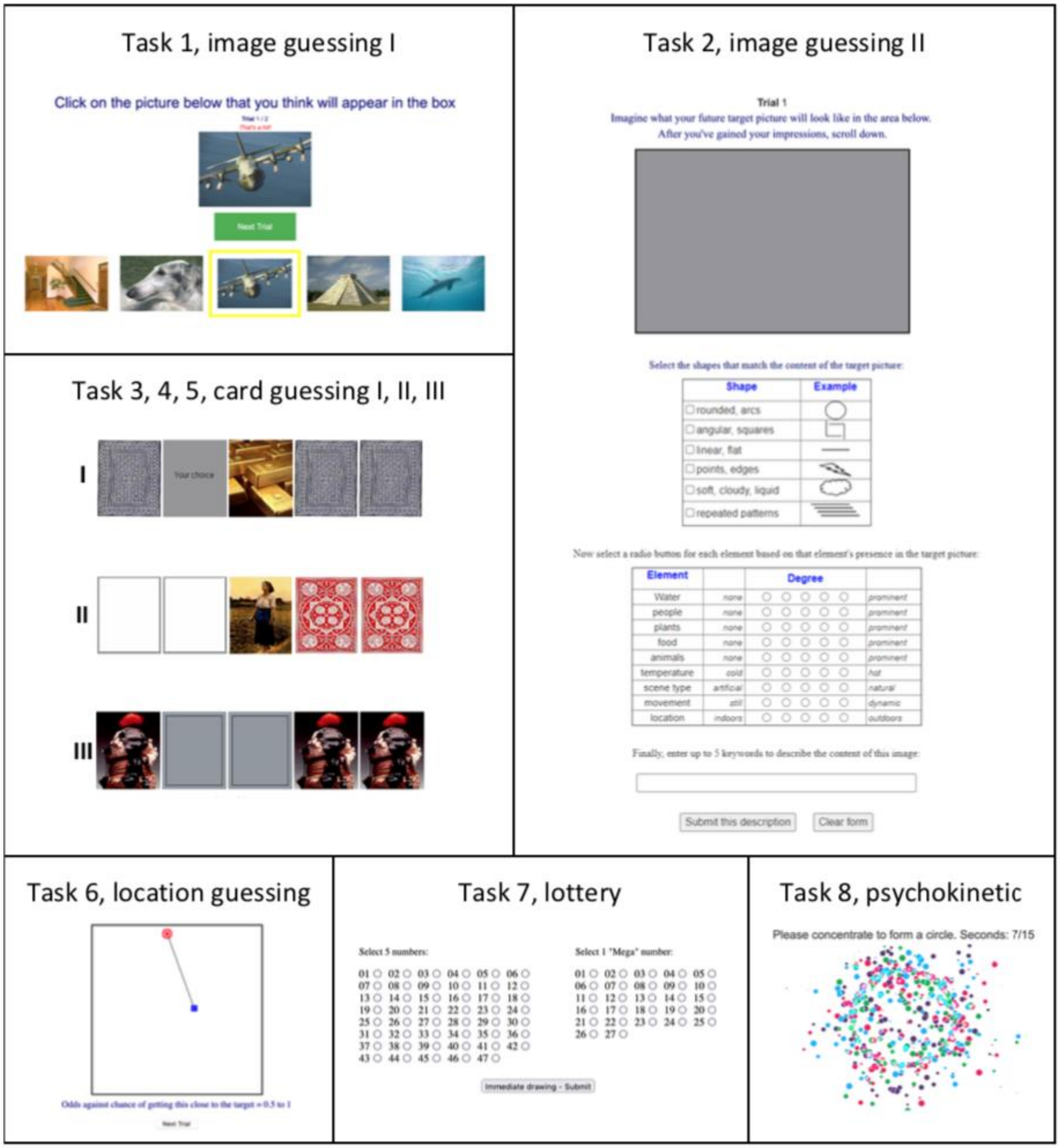

Figure 2. *Depiction of tasks 1 to 8. Tasks are organized on a website "dashboard." Participants can select tasks and perform them in any order.*

Aggregated score. Participants could perform each task several times (Table 1), although only their first attempt was considered in calculating their aggregated score. Scoring for each task was described above. We used the Stouffer Z method (equation 4) to obtain the aggregated score. Participants who did not complete each task at least once were not considered for inclusion in this study.

Participant feedback. A dashboard page indicated participants' progress, and a "Hall of Fame" showed their performance on each task compared to that of all other Gotpsi participants on a daily, monthly, and yearly basis. *Z*-scores across tasks were pooled using the Stouffer Z method (equation 4). Raw z-scores were not shown to participants. Instead, a score from 0 to 100 was calculated using the logistic formula, so $score = \frac{100}{1+e^{-(Z-1)}}$. Note that we offset the *z*-score by 1 to shift most score values above 50, providing, on average, more above 50 scores and a more enjoyable user experience.

Table 1. List of tasks and the minimum number of trials required to be considered for the calculation of an aggregated score and inclusion of a participant in the study.

Task	Description	Trials
1	Photo guessing I	20
2	Photo guessing II	2
3	Card guessing I	10
4	Card guessing II	10
5	Card guessing III	10
6	Location guessing	20
7	Lottery	1
8	Psychokinetic	1

Test/retest

1014 individuals completed all tasks and questionnaires. We assessed the test-retest scores of the top performers. The 50 individuals (potential talents) with the highest aggregated scores across all tasks excluding the lottery one (see Methods) on the suite of tasks were invited (and compensated for their efforts) to repeat all 8 tests to evaluate test-retest reliability. If any of the 50 participants declined to participate in the (compensated) retest, then the next person with the highest score was invited. In total, we contacted the top 200 performers (out of 1014 individuals), and 50 of them agreed to participate.

Sample size calculation

As an exploratory study, the sample size for this study was determined by assuming a small percentage of talent in the general population (May and Marwaha, 2018) for these kinds of online tasks. The exact percentage is difficult to assess but estimated on the order of 1 in 40 to 1 in 1000 (May and Marwaha, 2018); 1 in 40 corresponds to about 50 talents in 2,000 individuals. Once these individuals were identified, we ran statistical tests on each of these individuals independently.

Statistical analysis

All statistical analyses were preregistered on the KPU Registry (reference 1070).

Hypothesis 1. Do participants (both pre-test and retest) perform above chance expectations in at least one of the eight tasks? We will report the 95% confidence interval of the score for each task (across participants) using a bootstrap technique based on all participants' scores, assess the overlap with chance expectation, and derive an overall *p*-value. We will assess if all participants (not just the talented ones) performed each task above chance expectation, and we will consider trials of participants who did not complete all the tasks. Because there are eight tasks, this represents eight statistical tests, so appropriate correction for comparison will be performed using the Bonferroni method.

Hypothesis 2. We hypothesize that some percentage of top performers will have scores above chance expectations at their retest. Group analysis of the performance on the eight tasks will be assessed in the same way as Hypothesis 1 for the 50 top performers. In addition, we will test for correlation of aggregate z-scores between sessions 1 and 2 for these participants (regression analysis).

Hypothesis 3. Did personality traits or questionnaire results influence performance? Predictor (see Methods) evaluation will be performed using general linear models and *post hoc* analyses (all statistical tests will be corrected for multiple comparisons). This will be performed using data from all participants, including talented participants.

Technology. The questionnaires were implemented using the SurveyMonkey.com platform. The tasks on our website were implemented using the Laravel framework, which is a content management system (CMS) using PHP. All of the psychophysics task databases shared the same anonymized user IDs. Our website passed an anonymized user ID to SurveyMonkey, and that user ID was later used to uniquely identify the individual (to provide a way to contact top performers). Real-time statistics were implemented in client-side Javascript and/or server-side PHP functions and validated against functions of the Matlab statistics toolbox (The Mathworks, Inc.).

Resistance to tampering. Ensuring absolute online data integrity is challenging because of the ever-present threat of motivated computer hackers altering the data or the underlying programs. Authentication to the website was handled using Google Oauth 2.0. The advantage of using the Laravel CMS is its improved security and routing. We further protected our website from client-side (Javascript) and server-side (SQL database) code injection. In the first phase, we realized that most tasks had been programmed in the Javascript language and that it was possible, with some web engineering, for users to cheat. However, it was impossible to assess which users may have cheated with certainty. Also, cheating would likely happen after users have completed the tasks several times. In the analyses above, we were only interested in the first-time users who completed the task. Users could, for example, use some of the browser debug features to assess information about the next trials. Because we realized it was possible to tamper with the results in Phase 1, we changed the code for some of the tasks for Phase 2. To prevent cheating, we reprogrammed most of the tasks using server-side PHP code execution instead of Javascript client-side code execution. All random choices and score calculations were performed on the server and hidden from the client browser, therefore preventing cheating. Following these changes, there was no indication of cheating in the second version of the tasks. To cheat in the second version of the platform, participants would have needed to log into or hack our server. We used a standard Amazon Web Services solution, with only three ports open (SSH and HTTP, HTTPS) and only allowed authorized IP address to log in. Another avenue for possible hacking would have been to take advantage of a Laravel platform vulnerability. However, there are no known vulnerabilities in the version of Laravel we used (version 8.0). It is, therefore, unlikely that any participants tampered with the server data.

The implementation of the task statistics between the server (Phase 2) and client (Phase 1) implementation was compared to make sure they were identical. Overall, the tasks were identical between Phase 1 and Phase 2. There was nevertheless an issue with the implementation of the card draw task (although the statistics were correct, the task sometimes dropped data trials). This issue was unpredictable and due to the latency of the event in the client browser. This bug prevented the interpretation of the results of the card draw task for Phase 2, so this task was not considered for Phase 2. The lottery task was also ignored for both Phase 1 and Phase 2 because the z-score distribution for this task was not normal – due to the low likelihood of "winning" – and could not be pooled with the z-score of other tasks (see Results).

Public data release. Anonymized data collected in this study and offline MATLAB analysis scripts are publicly accessible on GitHub at https://github.com/InstNoeticSciences/gotpsi_offline_analysis.

Results

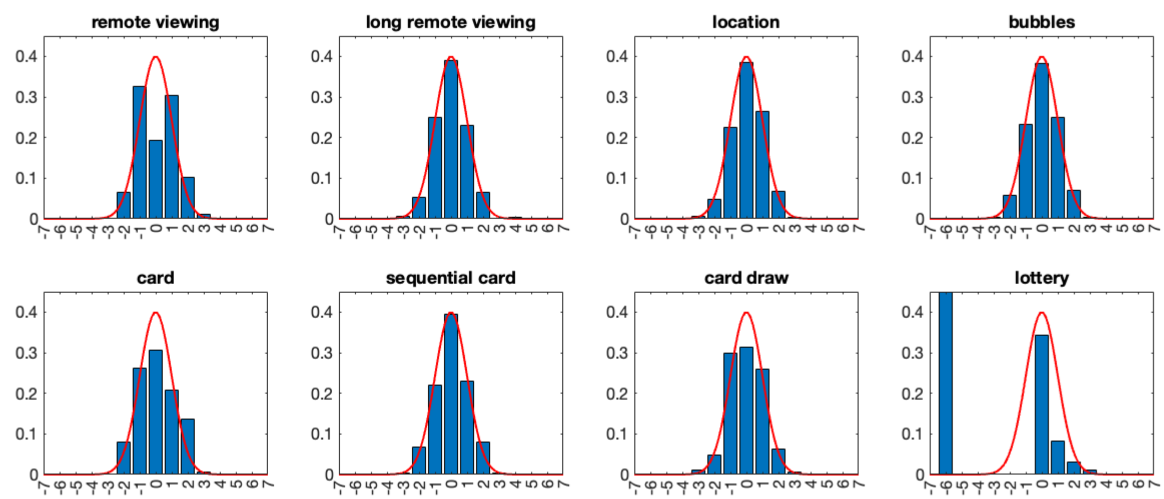

Figure 3. Z-score distribution for each task implemented in Phase 1 for the 1014 participants who finished all the required tasks.

A total of 8,112 persons participated in the tasks. Among them, 1,014 finished all the required number of trials in each task (see Method) to qualify for Phase 2. Results are shown in table 2 below. Although results for the lottery task are shown, we did not use them in the calculation of the *z*-score. The reason is that most users failed to find even one number, leading to a very negative *z*-score. In Figure 3, we can see that the distribution of the z-score was highly skewed, so when combined with another task score, the lottery score tends to be the only score that is taken into account. In retrospect, we should not have used the lottery task as there is no solution to integrate the score of this task with other tasks.

Table 2. Performance during Phase 1. The first column indicates the mean z-score for that task (see Methods). Note that in this case, z-scores are considered a performance measure, not a statistical measure. The second indicates the 95% confidence interval for the z-score using surrogate statistics (see Methods). The third indicates the p-value computed using surrogate statistics, and the last one indicates the Bonferroni-corrected p-value.

task	***mean z***	**Conf. Interval**	***p*-value**	**p-value Bonf.**
remote_viewing	0.11	0.04 to 0.18	0.0012	0.0096
long_remote_viewing	0.03	-0.03 to 0.10	0.1566	1
location	0.07	0.01 to 0.13	0.0154	0.1232
bubbles	0.04	-0.03 to 0.10	0.1264	1
card	0.1	0.04 to 0.17	0.0008	0.0064
sequential_card	0.05	-0.01 to 0.12	0.0648	0.5184
card_draw	-0.01	-0.07 to 0.06	0.5995	1
Stouffer Z global (excluding lottery)	0.15	0.08 to 0.21	0.0001	

As indicated in the Method section, we selected 50 top participants and asked them to complete the tasks again. Their results in the first and second phases for these participants are shown in Table 3.

Table 3. Performance during Phase 2 of the top performing participants in Phase 1. Both Phase 1 and Phase 2 are shown. See Table 2 for more information.

	Phase 1				Phase 2			
	mean z	Conf. Interval	p-value	p-value Bonf.	mean z	Conf. Interval	p-value	p-value Bonf.
remote_viewing	0.56	0.24 to 0.88	0.0002	0.0014	0.28	-0.01 to 0.58	0.0284	0.1704
long_remote_viewing	0.58	0.34 to 0.81	0.0001	0.0007	-0.28	-0.52 to -0.03	0.9854	1
location	0.64	0.40 to 0.86	0.0001	0.0007	0.16	-0.12 to 0.44	0.1272	0.7632
bubbles	0.47	0.24 to 0.70	0.0001	0.0007	0.24	-0.02 to 0.51	0.0347	0.2082
card	0.65	0.36 to 0.93	0.0001	0.0007	-0.03	-0.25 to 0.19	0.6439	1
sequential_card	0.37	0.11 to 0.64	0.0039	0.0273	-0.07	-0.41 to 0.24	0.6542	1
card_draw	0.54	0.24 to 0.83	0.0002	0.0014				
Stouffer Z selected	1.44	1.27 to 1.62	0.0001		0.13	-0.18 to 0.42	0.2028	

The 50 selected top performers performed above chance expectations in all the tasks of Phase 1 (Table 3, Phase 1) – which is not a surprise given that this is how we selected these participants. They did not perform consistently above chance at the retest (Table 3, Phase 2), except for the remote viewing and bubble tasks. However, that outcome did not resist correction for multiple comparisons.

Finally, we tested hypothesis 3 for a relationship between questionnaire items and overall performance (Stouffer Z across tasks for individual participants). We tested all participants, the selected participants in Phase 1, and the same participants in Phase 2 using a general linear model (MATLAB function *glmfit*), taking into account IDL questionnaire results (with the item *meditation* as a categorical variable and all other questionnaire results as continuous variables). The rationale for this analysis is that some individuals might have performed better than others at these tasks, and it would be important to know if these individuals shared some personality traits. For example, does regular meditation influence participants' performance?

This corresponds to hypothesis 3. The results are shown in Table 4. We first checked that the residuals were normal using the Lilliefors goodness-of-fit test of composite normality (*lillitest* MATLAB function). Participants who did not respond to some questions or whose data was corrupted were not included. The intercept, which is the global offset, is significant for Phase 1 for the 50 selected participants. This is expected because we selected these participants based on their performance in the tasks. Some associations were observed between the Big Five category of "conscientiousness" versus performance in Phase 2, where lower conscientiousness led to increased performance. However, the opposite trend was observed when all participants were considered together. Other significant trends above 0.07 are indicated in Table 4 for Phase 2. Note that the p-values for each trait are part of a single model, so they do not need to be corrected for multiple comparisons using Bonferroni. We also checked specifically for the remote viewing task and the bubble task, which showed a significant trend for Phase 2 (Table 3). We found that higher agreeableness was significantly correlated with higher performance (p=0.03) for remote viewing. No personality trait was associated with higher bubble task performance.

Table 4. Association between Stouffer Z across tasks and questionnaire scores.

	Phase 1 all	**Phase 1 (selected)**	**Phase 2 (selected)**
(Intercept)	ns	0.02 (+)	ns
meditate	ns	ns	ns
paranormal belief	ns	ns	ns
paranormal experience	ns	ns	ns
extraversion	ns	ns	ns
agreeableness	ns	ns	ns
conscientiousness	0.05 (+)	ns	0.02 (-)
neuroticism	ns	ns	0.09 (+)
openness	ns	ns	0.07 (-)
Cloninger score	0.07 (-)	ns	ns

Discussion

The first aim of this study was to develop and integrate various psychophysics tasks into a single online platform, which we were able to accomplish and deploy successfully. With the second and third aim, we developed data analysis methods to identify potentially talented individuals among a large sample (n=1014) and selected the top 50 performing individuals to assess the reliability and reproducibility of their abilities. With the fourth aim, we assessed whether there was any consistent above-chance performance in test-retest scores of talents. The performance of two of the eight tasks (remote viewing and psychokinesis) was above chance expectation in the retest phase, although it fell below the significance threshold after correction for multiple comparisons, possibly because of low statistical power. The final aim evaluated the potential personality predictors (if any) of high performance. In selected tests, we observed that the personality trait of conscientiousness was negatively correlated, and agreeableness was positively correlated with higher performance.

Limitations

There were minor differences compared to the pre-registration. We decided not to include the lottery task in the final analysis due to the skewed nature of the results, which was not foreseen in the pre-registration. We also decided to compute the statistics differently for the location task due to the limitation of the original calculation method. In addition, the card draw task was excluded from the Phase 2 analysis due to technical issues.

In any study where participants are selected based on particularly high or low scores and subsequently re-tested, regression to the mean is an important statistical artifact that must be acknowledged. This phenomenon occurs when extreme initial scores tend to shift closer to the population average upon re-testing, potentially skewing the study’s results. In our research, participants were identified based on exceptional initial performance, which may make the study vulnerable to this effect. While this could reduce the observed magnitude of effects in subsequent testing, any positive findings would not necessarily be incorrect. Nevertheless, we recognize the importance of addressing this limitation by including a control group in future phases of the study to better account for regression to the mean.

Nevertheless, the significance of this study in advancing the field of psi research cannot be overstated. Our research brings novel methodology to the field, notably the use of an online platform, which enabled us to collect data from a substantial number of participants. Our strategy is to select individuals with the highest initial performance in psi psychophysics tasks. To the best of our knowledge, this approach is unique and innovative. Nonetheless, it is important to acknowledge a potential limitation in our methodology. Selecting top performers based solely on their initial results may have introduced a selection bias: this approach risks excluding individuals who might demonstrate more consistent performance over time or those who are particularly adept at these specific tests but were not identified in the initial phase. Considering these factors in future research is crucial to ensure a more comprehensive understanding of purported psi abilities.

In our research, we considered the relatively rare prevalence of psi talents in the general population, as estimated by May and Marwaha (2018), to be approximately between 1 in 40 to 1 in 1000 individuals. This estimation provided a crucial benchmark for our study's sample size. However, in light of this, our sample of 1014 individuals can be regarded as relatively low, especially when aiming to capture a statistically significant representation of psi abilities. This limitation is evident when we analyze the trends observed in Table 2 of our study. The data suggests a more comprehensive understanding of these psi phenomena could be achieved with a larger sample size.

Future directions

We recommend that this study be replicated, incorporating a much higher number of initial candidates. Such an approach would not only enhance the available statistical power but also provide a more robust and generalizable understanding of the nature and prevalence of psi abilities in the general population.

The quick remote viewing task stands out as particularly promising. According to Cardeña (2018; Table 1), remote viewing is among the most solidly established psi psychophysics tasks, with some six large-scale studies demonstrating significant positive outcomes. For future replication efforts, it may be beneficial to concentrate on this specific task.

We invite the broader scientific community for further exploration and validation. Recognizing the importance of open science and collaborative research, we have made our anonymized dataset publicly available for download. Interested researchers may access and reanalyze these data, providing an opportunity for independent verification of the results and potentially offering new insights.

Abbreviations

CMS - content management system
PRNG - pseudorandom number generator

Ethics approval and consent to participate

This study was approved by the IONS Institutional Review Board (reference WAHH_2018_01).

Images depicted in figure 1

These images are our own or distributed under creative common licenses (according to Google image).

Consent for publication

Not applicable.

Availability of data and material

The material and data will be made available for download.

Competing interests

The authors declare no competing interests.

Funding

Bial Foundation, grant number 20/360

Authors' contributions

AD, HW, and DR designed the experiment. AD drafted the manuscript. HW and DR edited the manuscript.

Acknowledgments

This study was supported by a research grant from the Bial Foundation, grant number 20/360.

The authors thank Maria Wojakowski for her helpful comments.

References

Abramowitz, M. and Stegun, I.A. (1964) Handbook of Mathematical Functions, Dover, New York, 1046pp., sections 7.1, 26.2

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. Journal of Personality and Social Psychology, 63(4), 596-612

Bell, I. R., et al. (2004). "Development and validation of a new global well-being outcomes rating scale for integrative medicine research." BMC Complementary and Alternative Medicine 4: 1.

Bierman, D. J. (1995). A free response precognition experiment via the World Wide Web. Proceedings of Presented Papers, Parapsychological Association Annual Convention, August 5-8, 1995, Omni Durham Hotel and Civic Center, Durham, North Carolina, USA, p. 38 - 42.

Bierman DJ, Spottiswoode JP, Bijl A. Testing for Questionable Research Practices in a Meta-Analysis: An Example from Experimental Parapsychology. PLoS One. 2016 May 4;11(5):e0153049. doi: 10.1371/journal.pone.0153049. PMID: 27144889; PMCID: PMC4856278.

Box, G.E.P. and Muller, M.E. (1958). A Note on the Generation of Random Normal Deviates. The Annals of Mathematical Statistics. 29 (2): 610–611.

Braud, W. G. & Schlitz, M. J. (1991). Consciousness interactions with remote biological systems: Anomalous intentionality effects. Subtle Energies, 2 (1), 1-46.

Cappelleri, J. C., et al. (2009). "Psychometric properties of a single-item scale to assess sleep quality among individuals with fibromyalgia." Health Qual Life Outcomes 7: 54.

Cardeña E. The experimental evidence for parapsychological phenomena: A review. Am Psychol. 2018 Jul-Aug;73(5):663-677. doi: 10.1037/amp0000236. Epub 2018 May 24. PMID: 29792448.

Cardeña, E., & Marcusson-Clavertz, D. (2015). States, Traits, Cognitive Variables and Psi. Parapsychology: A Handbook for the 21st Century, 110.

Cloninger, C. R., et al. (1994). The Temperament and Character Inventory (TCI): A guide to its development and use. St. Louis, MO, Center for Psychobiology of Personality, Washington University.

Cloninger, C., Svrakic, N., & Svrakic, D. (1997). Role of personality self-organization in development of mental order and disorder. Development and Psychopathology, 9(4), 881-906.

DeSalvo et al., (2006) Mortality Prediction with a Single General Self-Rated Health Question.

J gen. Intern. med. 2005; 20:267–275.

Füredi, F. Székely, G.J., Zubor, Z. (1996). On the lottery problem. Journal of Combinatorial Designs. 4 (1): 5–10.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: failures to replicate psi.. Journal of Personality and Social Psychology, 103(6), 933-948. https://doi.org/10.1037/a0029709

Greenwald, A. G., D. E. McGhee and J. L. Schwartz (1998). "Measuring individual differences in implicit cognition: the implicit association test." J Pers Soc Psychol 74(6): 1464-1480.

Guilford, J.P., Christensen, P.R., Merrifield, P.R., and Wilson, R.C., 1960. Alternative Uses Manual. Sheridan Supply Co.

Gureje O, Von Korff M, Simon GE, Gater R. Persistent pain and well-being: a World Health Organization study in primary care. JAMA. 1998;280(2):147-151

Hyman, R. (1988). Psi experiments: do the best parapsychological experiments justify the claims for psi?. Experientia, 44(4), 315-322. https://doi.org/10.1007/bf01961269

Jahn, R. G., & Dunne, B. J. (1987). Margins of Reality. Harcourt Brace.

Jebb, 2017 Happiness, income, satiation and turning points around the world. validated in 1.7 million people

Johansson, Gunnar, and KR Westerterp. “Assessment of the Physical Activity Level with Two Questions: Validation with Doubly Labeled Water.” International Journal of Obesity 32, no. 6 (2008): 1031–33.

Johnson, D. (1977). PLATOnic parapsychology. The Daily Illini, November 19, 1977, p. 2.

Karen B. DeSalvo MD, Nicole Bloser, Kristi Reynolds, Jiang He, Paul Muntner (2006) Mortality Prediction with a Single General SelfRated Health Question: A MetaAnalysis, Journal of General Internal Medicine banner, Volume 21, Issue 3, Pages 267-275

Kim, J., Gabriel, U., & Gygax, P. (2019). Testing the effectiveness of the Internet-based instrument PsyToolkit: A comparison between web-based (PsyToolkit) and lab-based (E-Prime 3.0) measurements of response choice and response time in a complex psycholinguistic task. PloS one, 14(9), e0221802. https://doi.org/10.1371/journal.pone.0221802

Marsaglia, G., Zaman, A. (1993). The KISS generator. Technical Report, Department of Statistics, Florida State University, Tallahassee, FL, USA.

May, E. and Marwaha, S.B. (2018) The Star Gate Archives: Reports of the United States Government Sponsored Psi Program, 1972-1995: Volume 1: Remote Viewing, 1972-1984. McFarland & Company.

Milton, J. (1994). Mass-ESP: A Meta-Analysis of Mass-Media Recruitment ESP Studies. The Parapsychological Association 37th Annual Convention: Proceedings of Presented Papers, pp 284-334.

Nelson, R. D. (2015). Implicit Physical Psi: The Global Consciousness Project. Parapsychology: A Handbook for the 21st Century. E. Cardena, J. Palmer and D. Marcusson-Clavertz. Jefferson, North Carolina, McFarland & Company: 282-292.

Pratt, J. G., Rhine, J. B., Smith, B. M., Stuart, C. E., & Greenwood, J A. (1940/1966). Extra-sensory perception after sixty years. Boston: Bruce Humphries.

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

Rammstedt, Beatrice, & John, Oliver P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality, 41(1), 203-212.

Rebman, J. M., Radin, D. I. & Stevens, P. R. (1996). A precognition experiment on the World Wide Web. Proceedings of Presented Papers, Parapsychological Association Annual Convention, August 17-20, 1996, Handlery Hotel, San Diego, CA, p. 207 - 218.

Rhine, J. B. (1934). Extrasensory perception. Boston Society for Psychical Research.

Roe, C. A., Cooper, C. E., Hickinbotham, L., Hodrien, A., Kirkwood, L., & Martin, H. (2020). Performance at a precognitive remote viewing task, with and without ganzfeld stimulation: three experiments. Journal of Parapsychology, 84(1), 38-65. https://doi.org/10.30891/jopar.2020.01.06

Shiota, M. N., Keltner, D., & John, O. P. (2006). Positive emotion dispositions differentially associated with Big Five personality and attachment style. The Journal of Positive Psychology, 1(2), 61-71

Schwartz, S. A. (2020). The discovery of an american brig: fieldwork involving applied remote viewing including a comparison with electronic remote sensing. Journal of Scientific Exploration, 34(1), 62-92. https://doi.org/10.31275/20201481

Schultz, P.W., C. Shriver, J.J. Tabanico, and A.M. Khazian. 2004. “Implicit connections with nature.” Journal of Environmental Psychology 24 (1): 31–42.

Targ, R., & Puthoff, H. (1974). Information transmission under conditions of sensory shielding. Nature, 251(5476), 602–607. https://doi.org/10.1038/251602a0

Utts, J. (2019). An assessment of the evidence for psychic functioning. Journal of Parapsychology, 2(Suppl), 118-146. https://doi.org/10.30891/jopar.2018.03.10

Vallee, J. (1988) Remote viewing and computer communications - An experiment. JSE 2(1) p 13-27.

Wahbeh, H., Yount, G., Vieten, C., Radin, D., Delorme, A. (2022) Exploring Personal Development Workshops' Effect on Well-Being and Interconnectedness. Journal of Integrative and Complementary Medicine, vol. 28, no. 1, https://doi.org/10.1089/jicm.2021.0043.

Wahbeh, H., Yount, G., Vieten, C. et al. (2020) Measuring extraordinary experiences and beliefs: A validation and reliability study. F1000Research 2020, 8:1741. https://doi.org/10.12688/f1000research.20409.3.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of personality and social psychology, 54(6), 1063.