Journal of Anomalistics
Zeitschrift für Anomalistik

www.anomalistik.de www.igpp.de

Probing Top Performers in a Forced-Choice Clairvoyant Task

Helané Wahbeh [a,*], Michael Kriegsman [a], Beth Glick [b], Arnaud Delorme [a], Dean Radin [a]

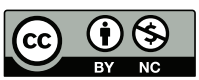
Abstract – This preregistered study analyzed more than 25 million trials from a web-based forced-choice remote viewing task to examine patterns of clairvoyant performance across all participants and among a subset of top performers. At the aggregate level, runs ending at the four planned lengths (5, 10, 25, or 100 trials) conformed to chance expectations. In contrast, optionally stopped runs displayed systematic fluctuations: short runs (1–3 trials) were above chance before declining, runs of 11–19 fell below chance, and runs beginning at 20 showed recurring above-chance spikes at every fifth run length (e.g., 30, 35, 40, 45, 50), which diminished beyond 80 trials. A Monte Carlo simulation matched to the empirical stopping distribution clarified the extent to which these patterns could be reproduced by optional-stopping behavior alone, with much of the run-length pattern, including the 11–19 trough and round-number variability, falling within the simulated null envelope. Exploratory analyses of top performers, defined post-hoc as the 1,235 users (2.64%) who exceeded chance uncorrected, after no users survived the preregistered FDR-corrected criterion, examined belief in psi, prior precognitive experience, meditation, total trials, and optional stopping as predictors. Optional stopping was the predictor most consistently associated with hits, both on first trials and across all trials, where it also interacted with belief, prior precognitive experience, and meditation jointly with cumulative task experience. Effect sizes were small (Δp and Cohen's d near zero for most predictors), and findings are interpreted as exploratory. The results suggest that group-level outcomes primarily reflect optional-stopping and related behavioral dynamics, whereas top-performer analyses surface more nuanced,

(a) Research Department, Institute of Noetic Sciences, Novato, California

(b) California Institute for Human Science, Encinitas, California

* Corresponding author: Helané Wahbeh

hwahbeh@noetic.org



but small, context-dependent associations between belief, experience, and behavior. These findings highlight the methodological challenges of large-scale, open online testing and the value of preregistered, participant-level approaches, combined with benchmark simulations, for distinguishing behavioral artifacts from potential psi signals.

Keywords: clairvoyance, forced-choice task, optional stopping, top performers, individual differences, preregistration

Introduction

Forced-choice tasks are a central experimental paradigm in parapsychology, requiring participants to choose among fixed alternatives to determine whether correct guesses exceed chance expectations. These designs afford precise control and have produced statistically significant, though small, effects in meta-analyses (Baptista et al., 2015; Cardeña, 2018; Storm et al., 2010; Storm & Tressoldi, 2020; Tressoldi & Storm, 2024a). However, their reliance on aggregating trials across participants has attracted methodological scrutiny.

Traditionally, parapsychological researchers have assumed that all trials are independent, enabling the use of straightforward statistical models. Under the null hypothesis, this assumption holds, and pooled trials increase statistical power and the apparent robustness of findings (Palmer, 1985). However, critics have noted that pooling can obscure important individual differences and inflate significance in the presence of statistical dependencies (Hyman, 1995; Utts, 1996). Because forced-choice tasks are among the most frequently cited evidence for psi, unexamined dependencies in trial data have implications for both interpretation and the credibility of the field[1].

Empirical and theoretical work suggests that the independence assumption may not always be justified. Participant performance can be influenced by factors such as learning, fatigue, feedback, and expectation bias, introducing potential dependencies between sequential trials (Dalkvist et al., 2014; Kennedy, 2013, 2014; Varvoglis et al., 2019). Failure to account for statistical dependence may lead to spurious results, exaggerated effect sizes, and invalid conclusions (Kennedy, 2016).

Moreover, there is substantial individual variability in psi performance. Selected participants, or those with traits, beliefs, or backgrounds thought to favor psi, tend to achieve higher

1 While free-response tests typically have stronger effect sizes, forced-choice tests have been commonly used in parapsychological research (see Storm et al., 2010, 2012; Storm & Tressoldi, 2020 for meta-analyses of forced choice and free-response studies).

hit rates than unselected samples (Cardeña, 2018; Storm et al., 2010; Tressoldi & Storm, 2024a; Utts, 1996). For example, Tressoldi and Storm (2024a) found that individuals selected for characteristics or experiences thought to be linked to psi performance, such as prior experience with extrasensory perception (ESP) experiments, belief in psi, psi training, or long-term meditation practice, showed effect sizes up to three times higher than unselected participants. These differences underscore the need for participant-level analyses that explicitly model within-participant variability and incorporate psychological and contextual moderators. To confront these challenges, parapsychological researchers have begun advocating for participant-level analyses, calculating individual hit rates, and explicitly modeling within-subject variation (Kennedy, 2013, 2016). Such methods offer a more nuanced and accurate understanding of psi effects, mitigating the limitations of aggregate analyses.

The Current Study

Beginning in 2000, Radin et al. (2019) launched the web-based GotPsi experiment suite, which by 2025 had collected hundreds of millions of trials from over 300,000 participants worldwide and produced significant results for some tests. In 2005, a new task called "Quick Remote Viewing" (QRV) was introduced, a misnomer, as it is best described as a simple forced-choice clairvoyance task. Building on previous research, the present study conducts preregistered analyses addressing two questions: (1) how hit rate varies as a function of run length across all participants, and (2) among participants who perform significantly above chance, which psychological and behavioral factors best predict performance, both for first-trial success and across all trials.

By combining a large archival database (over 25 million trials) with participant-level analyses of top performers, this study aims to clarify the extent of statistical dependence in forced-choice psi tasks and identify individual differences linked to exceptional performance.

Specifically, we address the following research questions that are approached with exploratory analyses:

Research Question 1. How does hit rate vary as a function of run length across all participants?

Research Question 2. Among participants who perform significantly above chance on the QRV task:

- 2a. Which, if any, of the following predictors are associated with first-trial success: belief in psi, past precognitive experience, meditation/meditative movement, total number of trials completed, and optional stopping ratio (including all main effects and interactions)?
- 2b. Which of these factors, if any, best explain success across all trials completed?

By simultaneously examining group-level trends and individual variability, this study aims to provide a more detailed understanding of the predictors and mechanisms that may support above-chance performance in forced-choice psi tasks. These analyses were preregistered with the Koestler Parapsychology Unit (KPU Registry 1094). Our preregistration specified these analyses, including the complete set of predictors and interactions. Here we present them as broader research questions to aid readability, while adhering to the pre-registered plan.

Methods

Study Procedures

Participants registered on the GotPsi website, www.gotpsi.org. During registration, they entered information about themselves, including handedness, beliefs, and experiences of psi, creativity, and remote viewing training and experience. They could then choose to engage in one of eight tasks. This study specifically examines the Quick Remote Viewing (QRV) task, which was initiated within the task suite on April 8, 2005. All human study activities were approved and overseen by the Institutional Review Board at the Institute of Noetic Sciences (IORG#0003743).

Task Description

The QRV task is a forced-choice photo guessing task where the participant attempts to identify a target image selected by the web server from five possible choices. For this task, a blank frame is displayed in the center of the screen, and five photos, drawn from a pool of 130 images, are displayed below it. The images were selected from the Corel Professional Photo image database and depict various scenes, including individuals, nature, and urban scenes. When the participant begins a trial, the web server selects the target image using a Perl-based linear congruential pseudorandom number generator function. This target exists in server-side state before the participant views the five response options and makes their selection. The participant then selects which of the five images they believe matches the hidden target, after which the target is revealed in the blank frame. Because the target is determined at the start of each trial rather than after the participant's choice, the QRV task is best characterized as a real-time clairvoyance paradigm (the target exists at the time of response and the participant attempts to identify it), rather than a precognition paradigm (in which the target would not yet be determined at the time of response). Although the study was preregistered under the label "precognition,"

subsequent clarification of the task code established that clairvoyance is the more accurate characterization. When the target picture is displayed, the participant sees 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 subsequent trial. The participant can select the number of trials they would like to perform, 5, 10, 25, or 100, which we define as the run length (run_length) for this study. Ten trials was the default option; users who did not actively select a different run length would begin a 10-trial run. Despite selecting the number of trials the participant intends to provide for each run, they could discontinue at any time. Because the task offered only four planned run lengths (5, 10, 25, and 100), any run ending at other lengths greater than 25 necessarily reflects within-run stopping from a planned 100-trial attempt rather than a chosen planned length. However, the dataset records only the number of trials actually completed per run, not the participant's chosen planned length. Runs ending at 5, 10, or 25 trials therefore cannot be distinguished between completed planned runs and optionally stopped attempts from a longer planned length. Only runs ending at 100 trials can be unambiguously identified as completed planned runs. If the participant completes the run, the participant is shown the percentage of "hits" achieved. A *z*-score is assigned to each run. The *z*-score is calculated as follows,

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

where n is the number of trials in the run, 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. For the present dataset, participants were able to take and repeat the task at their leisure. Thus, data can be viewed at multiple levels: as one omnibus hit rate, or as nested hit rates: within participant, within year/date, or within observed run length.

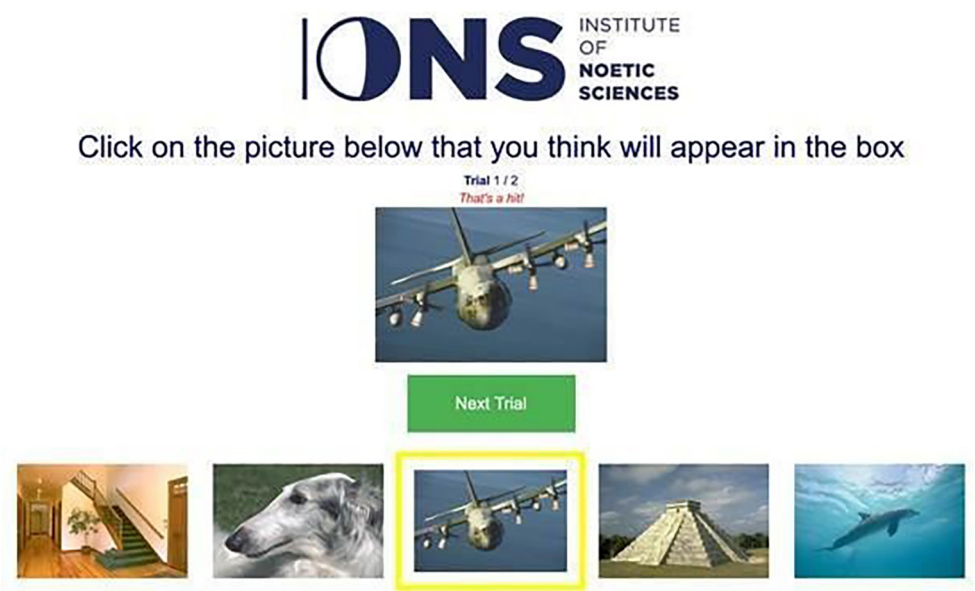
Subjective Measures

Basic demographics such as age, gender, race, and socioeconomic status were not collected for any participants. Handedness, creativity, and remote viewing experience were completed and not included in these analyses (see Supplemental Data for these methods and results).[2] Participant Beliefs and Experiences were part of the current analyses, and the measure is described below.

2 The Supplemental Data are available at: https://www.anomalistik.de/images/pdf/zfa/supp_mat/JAnom26-1_QRV-Supplemental-Materials.pdf

Figure 1

QRV Participant View



Note. This is the participant's view of the task after they have made their selection and the correct image has been revealed (highlighted in yellow). In this case, the trial was a hit.

Beliefs and Experiences (Table 3) – Participants were asked, "Please answer the following questions using a 5-point scale. For example, your answer to the question, "The degree to which you practice meditation," can range from "none" (the leftmost radio button) to "extensive" (the rightmost button)." These items could be considered ordinal. However, we considered them continuous because they were displayed to the user as equal intervals (i.e., the steps or gaps between each point on the scale are equal, and the perceived difference in agreement or frequency between each level is consistent across the scale). Items 1, 2, 3, 4, 12, and 13 were included in the regression models as potential predictors. Item 1 is about "*Belief*." Items 2, 12, and 13 are averaged to create a variable called "*Precog*." Items 3 and 4 are averaged to create a variable called "*Meditation*."

Data Cleaning

The present data set is extraordinarily vast because participants were able to repeat the task *ad infinitum*. Raw data were collected from April 8, 2005, to the present (October 2025). However, the data analyzed for this study included trials contributed only through December 29, 2018, because at that point, the underlying code, database, and host server were updated using more modern web-based programming methods. The QRV task was modified with the 2018 system update, and although data collection has continued with the updated platform through the present, the task changes made merging the pre- and post-update datasets inappropriate for the present analyses. The post-2018 data are retained for future analyses and could serve as an independent dataset for confirming the exploratory findings reported here. The raw data were imported into R (R Development Core Team, 2023) and merged into a single file for cleaning and analysis. Rows, each corresponding to one trial, were reordered by Username and Date to organize consecutive trials within their corresponding runs. A cleaning algorithm was then applied to flag, triage, and remove malformed data records and potentially nefarious attempts to game the system, such as utilizing multiple windows or bots. Specifically, the following types of errors were addressed: NAs (no discernable value due to unknown cause), disk access glitches causing duplicate consecutive trial numbers, skipped trial numbers, trials submitted with identical timestamps, or trial or hit count carryover into subsequent runs, and trials submitted "too fast" (1 second or less elapsed), suggesting that those trials were performed by bots attempting to hack into the web server. Among the raw data (i.e., before other cleaning steps), 0.51% of trials had a recorded duration of 0 seconds, and 16.36% had a duration of 1 second or less; inter-trial intervals are recorded as integer timestamps and therefore lack finer-grained resolution. Our rationale for this 1 second threshold is that participants must (1) view five target images, (2) click one, (3) see the feedback/result, and (4) proceed to the next trial, which we posit cannot be completed legitimately in under 1 second.

NAs were found only in the trial number variable, and only between runs. More specifically, NAs seemed to appear between runs that were submitted on different days, suggesting an encoding issue when the browser window was left open overnight. Because NAs occurred only on trials between runs, the implemented cleaning strategy was to remove these rows. Duplicated consecutive trial numbers were first segmented into two types. Consecutive first trials could occur when participants answered one trial, then quit and restarted a new run. This amounts to optional stopping, not an explicit error. However, consecutive trial numbers larger than 1 did indicate some kind of error. Trial skips were defined as instances where one or more trial numbers were missing in the consecutive count of trial numbers within a run. Carryover hits corresponded to the impossible situation where the hit count exceeded the trial number.

Carryover hits also included instances where the first trial was a miss, and yet the hit count was 1. These checks were applied to every trial in the dataset. In addition to the two patterns described above, runs were also flagged if a trial with trial number = 1 had hit = 1 but hit count > 1 (i.e., the first trial was a hit but the count had already carried over from a prior run). All retained trials satisfy hit count ≤ trial number.

These errors could have various causes, including users attempting to game the system (e.g., running the task with multiple windows open simultaneously, programming bots, or rapidly submitting responses), or errors in the back-end encoding of the data collection software. Because we could not discern the actual cause of these errors, we opted to remove runs that contained flagged trials with these error types (i.e., all trials for the suspect run).

Statistical Plan

For trial-level analyses, a random subset of 25% of users across all years was withheld for unplanned future analyses. The purpose of this withholding was to preserve a portion of the dataset for future work, rather than to support a specific planned analysis. In particular, because the top-performer analyses reported here are exploratory, the held-out subset could serve as an independent sample for confirming, or failing to confirm, the effects observed in the present analyses. Statistical analyses included three sections: (a) descriptives, (b) run-level analyses exploring Research Question 1, and (c) trial-level analyses exploring Research Question 2.

Descriptive statistics (means, *SDs*) were computed and two-tailed Welch's *t*-test was used to compare the 15 belief and experience variables across Top Performers versus Not Top Performers (see Table 3). Means and standard deviations for continuous variables, as well as counts and percentages for categorical variables, were calculated and presented for all included participants, including the random subset and the subset of top performers (see below). Participant characteristics were compared between top performers and non–top performers using Welch's two-sample *t*-tests, with effect sizes estimated by Cohen's *d* and accompanied by 95% confidence intervals.

Research Question 1:

How does the hit rate vary as a function of run length in all participants? (exploratory hypotheses with no *a priori* prediction)

The effect of optional stopping and patterns of performance by run length were assessed using a ribbon plot (see Figure 2), which shows the distribution of hit rates on the y-axis and the

number of trials in each given run on the x-axis. *Post-hoc* binomial tests were also conducted at each of the 100 run lengths to test if the average hit rate differed from chance. In addition to hit rates, we plotted the log of the number of runs submitted at each length. This was not preregistered but was included to aid interpretation, as run counts varied greatly across lengths. The log transform allowed both common and rare run lengths to be visualized on the same scale, providing context for evaluating which deviations from chance occurred at highly frequent versus rare run lengths. Descriptive binomial tests of hit rate were conducted for six nested trial subsets (1st trial, 1st 5, 1st 10, 1st 25, 1st 100, and all trials), separately for the full 75% retention sample and for the Top Performer subset. This nested-subset framing was consistent with the preregistered Stats Plan, which specified that hit rates could be examined at multiple levels. Finally, following the reviewer's suggestion, we conducted a Monte Carlo simulation to evaluate whether the observed run-length patterns could be explained by optional stopping alone. Null data were generated by drawing hits from a binomial distribution with $p = 0.20$, using run lengths sampled from the empirical run-length distribution of the 75% retention sample. The simulation was repeated 1000 times to construct a 95% null envelope against which the observed mean hit-rate curve (Figure 3) could be compared. Deviations of the observed curve beyond the envelope indicate patterns not reducible to optional-stopping behavior under the null.

Research Question 2:

Identifying Top Performers

In the preregistration, we intended to define top performers as those users who performed significantly above chance after FDR correction (Benjamini & Hochberg, 1995). No users survived FDR correction, so we post-hoc modified the definition of top performers to correspond to the 1235 users (2.64% of the total 46,722 users) who performed above chance, uncorrected. The full distribution of user-level Z scores across the sample is shown in Supplemental Figure 1, with corresponding numerical comparisons against the standard normal distribution in Supplemental Table 9; the observed upper tail of the distribution exceeded normal-theory expectations at every threshold examined (e.g., 9.06% of users exceeded $z = 1.645$ versus 5.00% expected).

Research Question 2a:

Which, if any, of the five predictors, belief in psi, past precognitive experience, meditation/meditative movement, total number of trials completed, and optional stopping ratio (including all main effects and interactions), are associated with top performers' success on their first

trial (i.e., first trial ever completed by each user)? The five predictors were chosen based on what the archival registration data actually measured and on prior literature. Belief in psi, prior precognitive experience, and meditation were included as trait-level predictors because each has been linked to psi performance in prior work (Cardeña, 2018; Storm et al., 2010). Total trials and optional stopping ratio were included as behavioral predictors of task engagement and stopping dynamics, with optional stopping, in particular, being repeatedly implicated as a driver of apparent above-chance effects. Remaining registration items (handedness, creativity, and remote-viewing training) did not have the same theoretical grounding and are reported descriptively in the Supplement. The first trial of each user's first run was used because it represents a unique instance where every user had no prior experience with the task, providing a homogenized dataset for cross-user comparison. This choice also minimizes contamination from optional stopping (since no stopping decisions have yet been made), yields one independent observation per user, and is consistent with prior reports of front-loaded or early-trial psi effects (Kennedy, 2003; Mossbridge & Radin, 2018). Hit outcomes were modeled with logistic regression, with a logit offset for chance ($p = 0.2$). Two models were estimated: a main-effects-only model (Supplemental Table 5) and the preregistered interaction model (Supplemental Table 6), which includes pairwise and three-way interactions among the three Psi-type predictors and Total Trials × Optional Stopping (see Supplement for full list of terms).

Research Question 2b:

Which of these factors, if any, best explain success across all of their trials? (Exploratory hypotheses with no *a priori* prediction.) Four regression-based models were assessed, all with the same five continuous predictors: (a) belief in psi (Belief), (b) past precognitive experience (Precog), (c) meditation and meditative movement (Meditation), (d) total number of trials completed (TotalTrials), and (e) optional stopping ratio (OptStop): the number of trials from runs optionally stopped divided by total number of trials. For each outcome (first trial; all trials), two models were fit: a main-effects-only model and the preregistered interaction model that adds pairwise and three-way interactions among the predictors (see Supplement for full list of terms). The first pair of models uses logistic regression to predict each top performer's first trial (hit or miss). The second pair predicts all trials from all top performers. In addition to reporting statistical significance, we calculated delta probabilities (Δp), which represent the average marginal change in predicted probability of a hit for a one-unit increase in each predictor. Because our dataset is very large, p-values alone could indicate significance for substantively trivial effects; Δp provides a more interpretable measure of effect size on the probability scale.

To control Type I error, FDR correction (Benjamini & Hochberg, 1995) was applied to the non-intercept predictors within each logistic regression model. This conservative approach exceeds standard practice but ensures that reported effects remain reliable even under a broad family-wise correction. All models were estimated as fixed-effect logistic regressions, as preregistered.

Results

Data Collected and Cleaning Procedures

There were 26,708,074 raw data trials collected from 64,775 participants (i. e., each unique username was treated as an independent participant, although a single person may have created multiple usernames). There were 25,457,187 trials after data cleaning. Please see Supplemental Data Table 2 for detailed trials per year and Supplemental Data Table 3 for the number of trials and users by year after cleaning procedures were implemented over all participants. The number and percent of trials removed during cleaning were as follows (see Methods): NAs = 8,736 (0.033%); duplicated consecutive trial numbers = 125,233 (0.469%); skipped trial numbers = 9,397 (0.035%); carryover hits = 16,269 (0.061%); and too fast = 211,271 (0.791%).

All top-performer analyses used N = 1,235 users (defined post-hoc as those exceeding chance uncorrected; see Methods) contributing 414,951 runs and 4,432,173 trials.

Table 1

Number of Trials, Runs, and Users Included in Analyses

	Trials	Runs	Users
All participants			
Raw data	26,708,074	not computed	64,775
Cleaned data	25,457,187	2,479,188	62,296
25% Users withheld	18,122,863	1,655,523	46,722
Top Performers	4,432,173	414,951	1,235
Not Top Performers	13,690,690	1,240,572	45,487

Table 2.

Descriptive hit rates by trial subset for the full 75% retention sample and Top Performers

Trial subset	Users (*N*)	Trials	Hits	Hit rate	Binomial *p*	Weighted mean	95% CI
Panel A. Full 75% retention sample							
1st trial	46,722	46,722	9,658	0.207	< .001	0.207	[0.203, 0.210]
1st 5 trials	36,615	183,075	39,179	0.214	< .001	0.214	[0.212, 0.216]
1st 10 trials	22,004	220,040	47,612	0.216	< .001	0.216	[0.215, 0.218]
1st 25 trials	11,053	276,325	58,307	0.211	< .001	0.211	[0.210, 0.213]
1st 100 trials	3,069	306,900	63,082	0.206	< .001	0.206	[0.204, 0.207]
All trials	46,722	18,122,863	3,633,161	0.201	< .001	0.201	[0.200, 0.201]
Panel B. Top Performers (exploratory, post-hoc; *N* = 1,235)							
1st trial	1,235	1,235	615	0.498	< .001	0.498	[0.470, 0.526]
1st 5 trials	1,128	5,640	2,453	0.435	< .001	0.435	[0.422, 0.448]
1st 10 trials	870	8,700	3,279	0.377	< .001	0.377	[0.367, 0.387]
1st 25 trials	497	12,425	3,937	0.317	< .001	0.317	[0.309, 0.325]
1st 100 trials	151	15,100	3,782	0.251	< .001	0.251	[0.242, 0.259]
All trials	1,235	4,432,173	900,881	0.203	< .001	0.203	[0.202, 0.204]

Note. Hit rate expected under chance = 0.200. For each subset, the number of users is the count of participants who contributed at least the indicated number of trials. Binomial *p*-values test whether the observed hit rate differs from 0.200. The weighted mean accounts for each user's contribution by total trials; over users, averages converge to the same omnibus rate across all trials. Panel B presents the same analyses restricted to the post-hoc Top Performer subset (users exceeding chance uncorrected; see Methods).

Descriptive hit rates by trial subset are presented in Table 2 for both the full 75% retention sample and the Top Performer subset. In the full sample, hit rates were slightly above chance across all subsets (ranging from 0.201 to 0.216) and highly significant given the very large sample sizes, though effect magnitudes were small. Among Top Performers, hit rates were substantially elevated and declined monotonically as more trials were included: from 0.498 on the first trial alone, to 0.435 across the first 5 trials, 0.377 across the first 10, 0.317 across the first 25, 0.251 across the first 100, and 0.203 across all trials. This pattern is consistent with a front-loaded effect but also with selective retention, because Top Performers were defined by above-chance performance overall, the pattern of early elevation may partly reflect this selection rather than a genuine front-loaded effect.

Participants Characteristics

Means, standard deviations, and group comparisons with two-tailed Welch's two-sample *t*-tests are presented in Table 3. To provide descriptive transparency, the table summarizes participants' self-reported beliefs and experiences separately for Top and Not Top performers. Although these comparisons were exploratory and not preregistered, they offer context for understanding whether the Top Performer subset differed meaningfully on background variables. Of the 15 self-report measures, two showed statistically significant differences after FDR correction. Top performers reported slightly more frequent meditation practice than non–top performers. Top performers also rated their view of time slightly closer to "raging waterfall" than "placid pool" (see Table 3). Although these differences were statistically reliable, the effect sizes were very small ($|d| \approx 0.07$, with confidence intervals extending nearly to zero), and are unlikely to reflect substantively meaningful psychological distinctions.

Research Question 1: Hit Rate Relationship With Run Length in All Participants

The ribbon plot illustrating the relationship between hit rate and run length is displayed as Figure 2. This analysis examines aggregate hit rates as a function of run length. Because the task offered only four planned run lengths 5, 10, 25, or 100 trials, runs ending at other lengths above 25 necessarily reflect within-run stopping from planned 100-trial attempts; runs ending at 5, 10, or 25 may represent either completed planned runs or optionally stopped attempts from longer planned lengths (see Task Description). Thus, any runs consisting of a different number of trials corresponded to a run in which the participant elected to stop the run before completion (or there was a glitch in the data encoding that stopped the run). Results for runs ending at the four planned lengths (5, 10, 25, and 100 trials) conformed to chance: none differed significantly from 0.2 in the run-level analysis underlying Figure 2 (Supplemental Table 4a). The trial-level binomial tests in Supplemental Table 4b flagged lengths 10 and 100 as significant, but the corresponding deviations were negligible (hit rates of 0.1994 and 0.2007; Cohen's $h \approx 0$), reflecting the very large number of trials at these common stopping points rather than a meaningful departure from chance.

Table 3

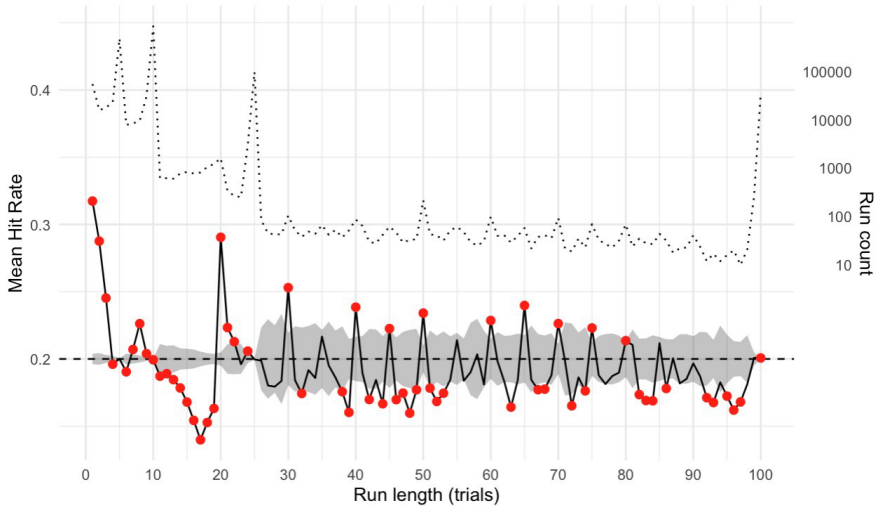
Participant Beliefs and Experiences

Degree to which you…	NOT Top Performers *N* = 45,487		Top Performers *N* = 1,235		Welch's Two-Sample *t*-test		Cohen's *d* [95% CI]
	Mean	*SD*	Mean	*SD*	*df*, *t*	*p*	
1. believe in "psychic" phenomena (none, absolute)	4.23	0.99	4.24	0.93	1200, -0.44	.66	-0.01 [-0.07, 0.05]
2. have had precognitive experiences (none, extensive)	3.45	1.16	3.48	1.13	1194, -0.75	.45	-0.02 [-0.08, 0.04]
3. practice meditation (none, extensive)	2.57	1.28	2.66	1.30	1188, -2.24	.03*	-0.07 [-0.31, -0.01]
4. practice martial arts or yoga (none, extensive)	2.03	1.25	2.09	1.29	1185, -1.65	.10	-0.05 [-0.11, 0.01]
5. consider yourself creative (none, extremely)	3.89	1.03	3.90	1.01	1192, -0.35	.72	-0.01 [-0.07, 0.05]
6. are lucky (none, extremely)	3.21	1.11	3.23	1.07	1195, -0.91	.36	-0.03 [-0.09, 0.03]
7. trust in your intuition (never, always)	3.95	0.90	3.98	0.86	1196, -1.30	.20	-0.04 [-0.10, 0.02]
8. trust in a religious faith (none, absolute)	2.86	1.48	2.89	1.46	1193, -0.81	.42	-0.02 [-0.08, 0.03]
9. have a sense of the spiritual (none, absolute)	3.99	1.12	4.00	1.11	1191, -0.35	.72	-0.01 [-0.07, 0.05]
10. are enthusiastic about sports (none, extremely)	2.59	1.37	2.62	1.37	1192, -0.53	.60	-0.02 [-0.08, 0.04]
11. work as a scientist (none, extensive)	1.83	1.21	1.80	1.18	1193, .04	.30	0.03 [-0.03, 0.09]
12. are trained in remote viewing (none, extensive)	1.32	0.80	1.32	0.81	1188, -0.07	.95	0.00 [-0.06, 0.06]
13. actively remote viewing (none, extensive)	1.43	0.92	1.48	0.96	1183, -1.42	.16	-0.04 [-0.10, 0.01]
14. have participated in psi experiments (none, extensive)	1.54	1.02	1.55	1.05	1180, -0.30	.77	-0.01 [-0.07, 0.05]
15. view time metaphorically as a (placid pool, raging waterfall)	3.01	1.20	3.10	1.21	1187, -2.38	.02*	-0.07 [-0.13, -0.01]

Note. Groups compared via Welch's two-sample *t*-test, equal variances not assumed, applying Satterthwaite approximation for degrees of freedom. Reported *p*-values are unadjusted. Asterisks indicate predictors that remained significant after false discovery rate (FDR) correction for multiple comparisons ($p < .05$). The *N* for Top Performers differs from the *N* in Table 1 because not all participants opted to complete these items.

Figure 2

Ribbon Plot of Hit Rate as a Function of Run Length



Note. In this ribbon plot, each value along the X axis corresponds to all runs submitted for that given run length. The solid line shows the mean hit rate on the left-sided y-axis, and circles indicate run lengths where the mean hit rate was significantly different from 0.2 after FDR correction for these 100 tests. The gray ribbon corresponds to the null envelope, or mean hit rate of 0.2 +/- 1.96 SE, FDR corrected. The dotted line shows the log of the number of runs submitted for each given run length, indicated by the right-sided y-axis.

The ribbon plot in Figure 2, with significant values listed in Supplemental Table 4, reveals five patterns in the average hit rate (represented by the black line), which are outlined in the following five paragraphs. In brief, these five patterns include that the mean hit rate (1) was nonsignificant for complete runs, (2) decreased continually over run lengths 1–4, (3) 11–19, and (4) 65–97, and (5) spiked upward on every fifth run length beginning at 20. In addition to hit rates, Figure 2 includes the log of the number of runs submitted at each length (dotted line), providing context for how often each run length occurred. This allows deviations from chance to be interpreted in light of the frequency with which participants stopped at those run lengths. As expected, there were much greater run counts at 5, 10, 25, and 100.

Pattern 1: At the four planned run lengths 5, 10, 25, or 100 trials, the mean hit rate was not significantly different from 0.2, though it should be noted that runs ending at 5, 10, or 25 are a mixture of completed planned runs and runs optionally stopped at those lengths from longer

planned attempts. Deviations from chance emerged primarily at run lengths outside the planned options, which necessarily reflect optional stopping. The log of run counts revealed four pronounced spikes at these planned run lengths, confirming that they were by far the most common stopping points. By contrast, incomplete runs showed a much more variable distribution of frequencies, with some lengths rarely attempted and others clustering around round numbers. These patterns provide important context for interpreting the fluctuations in hit rate described below.

Pattern 2: The second pattern was a decreasing hit rate for run lengths 1–4. Overall, runs consisting of only one, two, or three trials showed hit rates significantly above chance, with a descending magnitude. This pattern continued for runs with four trials, where the mean hit rate is significantly less than 0.2. Runs with six and eight trials were significant, but no clear pattern was observed for run lengths between six and ten trials.

Pattern 3: Run lengths from 11 to 19 were significantly lower than chance, with a general pattern of decreasing hit rate. This pattern reversed dramatically at runs with 20 trials, for which the average hit rate was significantly above chance.

Pattern 4: Run lengths from 65 to 97 also revealed another pattern of decreasing hit rates. The overall pattern observed for these run lengths above 65 can be seen as the additive effects of patterns 1, 4, and 5 (discussed below).

Pattern 5: A recurring upward spike in hit rate was observed beginning at runs of 20 trials and reappearing at every fifth run length thereafter (e.g., 30, 35, 40, 45, 50, etc.). The log of run counts showed the same periodic pattern, indicating that these upward deviations in hit rate coincided with points where participants disproportionately chose to stop. In other words, the peaks in performance were not randomly distributed but systematically aligned with round-numbered stopping points, suggesting that stopping behavior was closely linked to the observed fluctuations in hit rate.

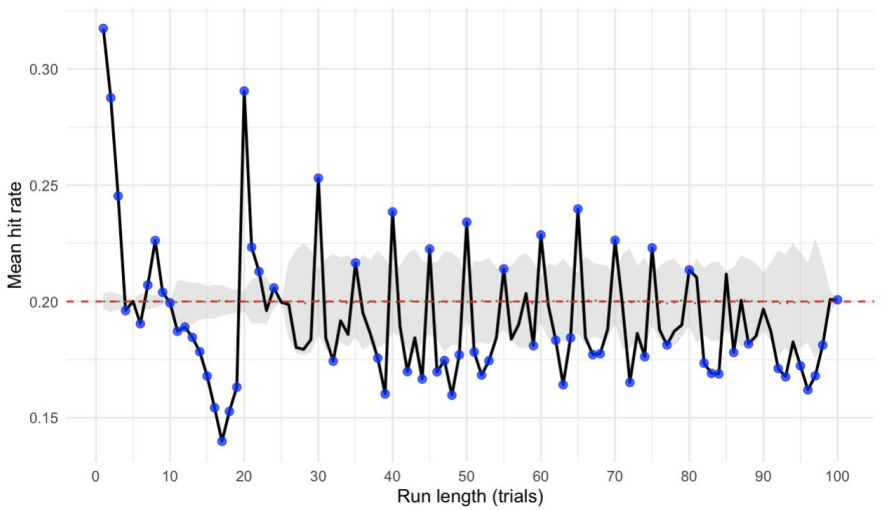
Finally, there were overlapping additive effects of the above patterns. Of note, the upward spiking of the Pattern 5 was not seen for runs of length 25 or 100 (see the Pattern 1), but was observed for the remainder of trials (i.e., 20, 30, 35, 40, 45, 50, etc.). For run lengths above 65, fluctuations in hit rate followed both the gentle decrease in hit rate of Pattern 4 and the spiking of Pattern 5. This blending of patterns was most interesting for run lengths between 92 and 97, where the positive spiking was observed for run lengths of 95 (and surprisingly, more so for 94). However, the magnitude of the decreasing hit rate overpowered the spike, so that even with the positive spike in hit rate at 95, the hit rate there was in fact significantly below chance. In other words, all run lengths between 91 and 98 were significantly below chance (except 94), such that the positive spiking at 95 was masked by the decreasing trend from 65 to 97.

To summarize, when aggregated across all participants, completed runs of 5, 10, 25, or 100 trials showed no significant deviations from chance expectation ($M \approx 0.20$ hit rate). In optionally stopped runs, systematic patterns emerged: very short runs of 1–3 trials were above-chance but declined toward chance by four trials; runs of 11–19 trials were significantly below chance; and runs beginning at 20 trials showed recurrent spikes at every fifth run length, such as 30, 35, 40, 45, 50. These effects diminished for very long runs (> 80 trials), where overall performance dipped below chance.

Following the reviewers' suggestion, we compared early-trial hit rates between runs that were optionally stopped and runs that continued. Runs stopped after a single trial ($N = 54937$) had a mean hit rate of 0.317, significantly higher than the 0.197 observed for continuing runs ($N = 1{,}601{,}472$), $t(57{,}721.13) = 59.91$, $p < .001$, Cohen's $d = 0.30$, 95% CI [0.29, 0.31]. Runs stopped after 2 trials ($N = 15{,}871$) showed a significant gap (0.287 vs. 0.198, $d = 0.32$, 95% CI [0.30, 0.33]). Runs stopped after 3 trials ($N = 17{,}840$) showed a hit rate of 0.244 vs. 0.198 for continuing runs ($d = 0.20$, 95% CI [0.19, 0.22]). These differences are modest and consistent with the interpretation that optional-stopping behavior selectively retains lucky early outcomes, although they are also consistent with genuine early-trial effects.

Figure 3

Ribbon Plot of Hit Rate, Monte Carlo null envelope



Note. Blue circles indicate empirical outside 95% null envelope.

Figure 3 displays the empirical mean hit-rate curve alongside the 95% null envelope from the Monte Carlo simulation. Several features of the run-length pattern fall within the null envelope and are therefore adequately explained by optional stopping under chance-level performance: the below-chance dip across run lengths 11–19, much of the variability across run lengths 30–100, and many of the round-number points. In contrast, the elevations at run lengths 1, 2, and 20 lie well above the null envelope, and several below-chance points in the 65–97 range fall below the lower envelope. These deviations are not reducible to optional stopping alone and indicate that some features of the empirical curve, particularly the early-trial elevation and the run-length-20 peak, require additional explanation.

To further test whether round-number patterns reflect selective stopping based on performance, we compared cumulative hit rates at each unambiguous round-number stopping point (lengths 30 through 90 in increments of 5) between runs that ended at that length and runs that continued past it (Supplemental Table 10). Run lengths 5, 10, and 25 could not be tested because they coincide with planned run lengths and cannot be classified unambiguously; run length 95 was excluded due to insufficient runs ending at that length. Of the 13 comparisons tested, 8 were significant after FDR correction. Effects were strongest at run lengths 30 and 65 (Cohen's $d = 0.69$ and 0.80, respectively) and attenuated at higher run lengths (no significant effects at run lengths ≥ 80). Across significant comparisons, runs that ended at the round number had cumulative hit rates 2–5 percentage points higher than runs continuing past, consistent with users selectively stopping after favorable performance.

Predictors of Top Performers' Success

An overview of the four preregistered logistic multiple regression models applied to the post-hoc Top Performer subgroup is displayed in Table 4. Two outcomes were modeled (first trial; all trials), each with two model specifications: main effects only, and main effects plus the preregistered interactions.

Table 4

Trial-Level Analysis: Overview of Predictors Across Top Performer Models

Predictor	First Trial: Main Effects	First Trial: Main + Interactions	All Trials: Main Effects	All Trials: Main + Interactions
Intercept	2.71 **	0.39	1.40 ***	1.54
Belief in Psi	1.02 (.004)	1.72	0.92 (-.014) ***	0.93
Precog	0.98 (-.005)	3.67	1.04 (.006) ***	1.07
Meditation	0.91 (-.022)	2.78	1.02 (.004) ***	1.09
Total Trials	1.00 (.000195)	0.89	1.00 (-.000044) ***	1.00
Opt Stop Ratio	3.45 (.292) ***	3.37 ***	2.01 (.113) **	3.85 ***
Belief × Precog	–	0.71	–	0.98
Belief × Meditation	–	0.77	–	0.98
Precog × Meditation	–	0.49	–	0.92
Belief × Total Trials	–	1.00	–	1.00
Precog × Total Trials	–	1.00	–	1.00 ***
Meditation × Total Trials	–	1.00	–	1.00
Belief × Precog × Meditation	–	1.18	–	1.02
Belief × Total Trials × Opt Stop	–	1.00	–	1.00 ***
Precog × Total Trials × Opt Stop	–	1.00	–	1.00 ***
Meditation × Total Trials × Opt Stop	–	1.00	–	1.00 ***

Note. The values are odds ratio (OR), delta probability (*Δp*; the change in probability for a one-unit change in the predictor; therefore intercepts and interactions do not have *Δp*; *Δp* values are also omitted in models with interaction terms because they reflect linear approximations that may be unreliable in those models, see OR for effect direction), and *p*-value levels, with FDR correction (Benjamini–Hochberg) applied to non-intercept predictors within each model. Em dashes (—) indicate that a term is not included in that model. The All Trials models assess all trials from each subject regardless of run structure. The First Trial models assess just the first trial (implicitly, from the first run) submitted by each user. Asterisks indicate the traditional significance thresholds (.05, .01, .001) after adjusting for FDR. Given the very large sample size, significance may be reached for effects that are substantively small; see *Δp* column for effect magnitudes.

Top Performers First Trial

Two logistic models were run for the top performers' first trials only analyses: (1) main effects only (Supplemental Table 5) and (2) main effects plus the preregistered interactions, including three-way terms (Supplemental Table 6).

In all both models, optional stopping (OptStop) was the only significant predictor (besides the intercept); no Psi-type, behavioral, or interaction term reached significance after FDR correction in the preregistered model. In the main-effects model, OptStop was associated with a Δp of 0.29, indicating that a user with 100% optionally stopped runs had, on average, a 0.29 higher probability of hitting on their first trial than a user with 0% optionally stopped runs.

Top Performers All Trials

Two models were conducted for the top performers' all trials analyses: (1) main effects only (Supplemental Table 7), and (2) the preregistered interaction model including three-way terms (Supplemental Table 8).

In the main-effects model, all five predictors were significant after FDR correction, although effect magnitudes were small ($|\Delta p| \leq .014$ for all predictors except OptStop, which had $\Delta p = 0.11$). In the preregistered interaction model (Supplemental Table 8), only the main effect of optional stopping reached significance; main effects of Belief, Precog, Meditation, and Total Trials were no longer significant once preregistered interactions were included. Among interactions, three of the preregistered three-way terms reached significance: Belief × Total Trials × OptStop, Precog × Total Trials × OptStop, and Meditation × Total Trials × OptStop, along with the Precog × Total Trials two-way interaction. As noted in Methods, however, statistical significance at this scale does not imply substantive importance; effect magnitudes were generally small.

In summary, across both first-trial and all-trial analyses, optional stopping was the strongest predictor of success among top performers. In the main effects only models for all trials, Belief, Precog, Meditation, and Total Trials each showed small but statistically reliable associations with hit probability ($|\Delta p| \leq .014$). However, when the preregistered interactions were added, these main effects were no longer significant, while three-way interactions among each Psi-type variable, Total Trials, and Optional Stopping reached significance. This pattern suggests that the influence of belief, prior precognitive experience, and meditation on top performers' performance depends jointly on cumulative task experience and stopping behavior, rather than operating as independent traits. Notably, the main-effect Δp for OptStop in the preregistered all-trials interaction model was extreme ($\Delta p \approx 2.97$), reflecting limitations of the linear approximation when interaction terms are included; the odds ratio (OR = 3.85, $p < .001$) provides a more interpretable estimate of OptStop's role in this specification.

Discussion

This preregistered study examined over 25 million trials from the Quick Remote Viewing task to investigate how psi performance unfolds across run lengths, stopping behaviors, and individual differences. Results reveal that apparent performance fluctuations were not uniform but emerged intermittently, shaped by pacing, engagement, and stopping strategies. From a conservative standpoint, these dynamics may largely reflect known response biases, such as optional stopping. If psi effects exist, they are likely subtle and embedded within these behavioral patterns.

Patterns of Performance in All Participants

On the surface, participants appeared to "know" when to stop, with elevated hit rates after one or two trials and recurring peaks at every fifth run length beginning at 20 (e.g., 30, 35, 40, 45, 50) indicating that performance interacted with self-paced stopping. However, much of this apparent prescience is reproducible from stopping behavior alone. Runs ending at the planned lengths (5, 10, 25, 100) conformed to chance, whereas runs at other lengths, which by definition reflect within-run optional stopping, showed systematic above- and below-chance fluctuations that faded in very long runs (>80 trials). These dynamics could result from expectancy or learning effects but also resemble classic reports of unstable or "psi missing" (Kennedy, 2003). Monte Carlo simulations imposing only optional-stopping rules clarify how much of this structure is reproducible as a behavioral artifact. The 11–19 trough and much of the round-number variability fell within the simulated null envelope, suggesting these patterns are adequately explained by optional stopping under chance-level performance. However, the elevations at run lengths 1, 2, and 20 exceeded the null envelope, indicating that these specific features cannot be reduced to stopping behavior alone. This elevation pattern resembles classic reports of unstable or front-loaded performance in the psi literature (Kennedy, 2003), though we cannot rule out residual stopping-related artifacts at these specific run lengths in this particular study.

At the participant level, variability dominated: three below-chance stretches (4, 11–19, 65–97 trials) alternated with short bursts above chance (1–3 trials, round numbers). Rather than field-wide decline, such patterns likely reflect context-dependent fluctuations driven by participant behavior and task dynamics (Tressoldi & Storm, 2024b). Notably, this finding pertains to study-level effect sizes aggregated across decades of research. These results may differ from the within-individual fluctuations observed here, which reflect performance dynamics at the participant level. Instead, these results suggest that participant-level variability and task context may be primary drivers (cf. Tressoldi & Storm, 2024b, study-level), reframing forced-choice psi tasks as behaviorally dynamic systems where procedure and stopping strategies shape observed outcomes.

Individual Differences and Top Performers

No individual met preregistered significance after multiple-comparison correction, yet in a pool of >46,000 users, even rare, stable deviations may be meaningful. Conservative false-discovery procedures minimize spurious results but also obscure potential "black-swan" performers. Exploratory analyses identified a small high-performer subgroup whose behavior merits confirmation in preregistered, fixed-length retests. A two-stage design, broad screening followed by focused retesting, could better evaluate stability across sessions.

Among these Top Performers, optional stopping was the strongest predictor of success, particularly for first trials (Δp = 0.29 for first trials), with other predictors contributing only marginally ($|\Delta p| \leq .014$). In the interaction model for all trials, the pattern was qualitatively similar: stopping behavior again carried the largest effect, with smaller contributions from belief, meditation, and prior precognitive experience emerging only through their three-way interactions with Total Trials and Optional Stopping. These findings suggest that situational and behavioral factors, more than enduring traits, shape observed outcomes.

An alternative interpretation of the three-way interactions involving Total Trials and Optional Stopping (Belief × Total Trials × OptStop, Precog × Total Trials × OptStop, Meditation × Total Trials × OptStop; see Supplemental Table 8) is worth acknowledging. Participants who complete very large numbers of trials have ample opportunity to refine informal stopping rules, such as stopping after perceived "hot" streaks, at psychologically salient milestones, or when early performance seems favorable. These refinements may interact with trait-level differences in belief, meditative practice, and prior precognitive experience. The result may be inflated hit rates through outcome-dependent stopping rather than enhanced psi accuracy that varies as a function of those traits. This perspective is compatible with the broader behavioral-dynamics framing of this study and does not require assuming that psi ability changes as a function of experience or as a function of belief, meditation, or prior precognitive experience.

Methodological Implications

Psi research traditionally relies on pooled binomial tests assuming trial independence (Jahn & Dunne, 1987; Utts, 1991). In the QRV context, such aggregation can mask sequential clustering and behavioral dynamics. Once deviations appear, independence assumptions weaken, and apparent effects may arise from expectancy, fatigue, or stopping biases. Alternatively, if psi exists, its operation may be intermittent and state-dependent, intertwined with ordinary variability. Mixed-effects and repeated-measures models (Storm et al., 2010; Tressoldi & Storm,

2024b) are better suited to capture these temporal fluctuations and individual differences than fixed-effect binomial tests.

Because massive datasets yield minuscule *p*-values even for trivial deviations (Wagenmakers et al., 2012), reporting effect sizes, confidence intervals, and Bayesian estimates are essential. Preregistration further constrains analytic flexibility yet aligns psi research with open-science norms (Nosek et al., 2018).

Implications for Parapsychology

At a broader level, this large-scale QRV study underscores both the opportunities and challenges inherent in modern, open-ended online psi testing. Compared with traditional fixed-length laboratory tasks, at-will designs introduce substantial methodological complexity, particularly through optional stopping and unstandardized participant behavior, highlighting the need for refined analytic approaches rather than simply replicating classic forced-choice methods. The influence of stopping behavior and participant variability underscores the limits of examining only aggregated trial data and supports the integration of participant-level modeling as a standard practice (Tressoldi & Storm, 2024b). Rather than weakening evidential standards, this approach complements pooled binomial tests with analyses that investigate the mechanisms driving anomalous cognition. Framing psi research in terms of both proof and process complements pooled binomial tests with participant-level models that investigate mechanisms, while highlighting challenges inherent in open, online testing. This integrative perspective also aligns with contemporary movements in consciousness studies that emphasize linking first-person experience, interactive dynamics, and third-person measurement (Cardeña, 2018; Kripal, 2019).

Interpreting Psi Dynamics

The QRV data are compatible with multiple interpretations. The dominant role of optional stopping behavior could reflect a purely psychological bias (expectancy, fatigue, or preferences for round-number runs), or it could reflect subtle psi-related intuitive or embodied cues influencing the moment of decision (i.e., the decision to stop as itself a form of psi). The present data cannot distinguish these possibilities. This is a question best answered with preregistered designs combining phenomenological and physiological measures.

A more skeptical reading of the same pattern is also possible. The absence of effects in runs of 100 trials, the only run length that can be unambiguously identified as a completed, fixed-

length attempt (see Methods, Task Description), is itself a substantive finding. It is compatible with at least three interpretations: that psi manifests primarily under novelty or uncertainty and does not survive pre-committed task structures (Kennedy, 2001, 2003); that psi operates by influencing *when* participants act rather than *what* outcome occurs, such that fixed-length runs eliminate the relevant degree of freedom (decision augmentation theory; May et al., 1995); or that apparent effects in optionally stopped runs reflect stopping-related artifacts rather than anomalous cognition. These interpretations converge on a common methodological recommendation: fixed-length, preregistered designs are essential for distinguishing candidate psi signals from behavioral artifacts.

Regression analyses also indicate a front-loaded effect, with higher probabilities of success on the very first trial. Under a psi interpretation, this would be consistent with both a genuine peak early in performance (Kennedy, 2003; Mossbridge & Radin, 2018) and with selective retention of early hits into the Top Performer subgroup. Such episodic patterns mirror findings from spontaneous cases and Ganzfeld experiments, where psi appears briefly and context-dependently (Wahbeh et al., 2018; Tressoldi & Storm, 2024a). Descriptive hit rates across nested trial subsets (Table 2) further illustrate this front-loaded pattern among Top Performers, where hit rates declined monotonically from approximately 50% on the first trial to chance levels across all trials. However, this decline is also consistent with selective retention: because Top Performers were defined by overall above-chance performance, users whose early trials were hits are over-represented, particularly at shorter subsets. The present data cannot distinguish a genuine front-loaded psi effect from this selection dynamic.

Trait-level predictors, such as belief, meditation, or prior precognitive experience, were weak. This may reflect limitations of the brief self-report measures used at registration, or it may indicate that stable individual differences play a smaller role than situational and behavioral factors in this paradigm. Performance declined with heavy repetition, consistent with fatigue or boredom, but also with proposals that psi is transient or self-limiting (Kennedy, 2001). Future work should distinguish between ordinary motivational decline and possible self-limiting psi dynamics.

Limitations

The QRV dataset has several limitations to be considered when reviewing the results. Participants did not provide demographic information, such as age, gender, or race, which prevented examining how these variables might moderate psi performance. Because usernames were self-generated, some individuals may have registered multiple times under different aliases.

Although the dataset's massive scale (>25 million trials) likely minimizes the effects of duplication on pooled outcomes, it reduces certainty in participant-level and repeated-measures analyses. As noted in the Methods, the dataset records the number of trials completed in each run but not the participant's chosen planned length. This prevents clean separation of completed planned runs from optionally stopped runs at the 5-, 10-, and 25-trial lengths, and conclusions about performance at these specific lengths should therefore be interpreted with that constraint in mind. Some analytic choices, such as the post-hoc redefinition of top performers and the use of log-transformed run counts, were not preregistered but were adopted to maintain feasibility and interpretability. The preregistered fixed-effect logistic regression approach does not formally account for the nested structure of trials within users. Mixed-effect models with random intercepts for user and appropriately scaled predictors would be a more appropriate alternative for confirmatory replications.

Brief registration questionnaires captured handedness, creativity, RV training, and belief/experience measures, but these were single-item indicators that may not have been able to detect subtle relationships. Self-selection likely produced a sample skewed towards believers, limiting variance and reducing power for trait–performance correlations. The group comparison in Table 2, though not preregistered, was included for descriptive transparency and showed that small numerical differences between Top and Not Top performers were unlikely to explain the preregistered findings. Future studies could include fuller multi-item psychological scales, demographic variables, and targeted modeling to clarify moderators such as handedness, creativity, and RV training. Finally, uncontrolled device, browser, and environmental differences may have introduced noise or subtle biases not accounted for in the current analyses.

Future Directions

These findings highlight key directions for future research. Mixed-effects and multilevel models remain essential for accommodating individual variability and session dynamics while retaining sensitivity to meaningful effects (Tressoldi & Storm, 2024b). Repeated-measures designs can help distinguish stable individual differences from noise and clarify how psi unfolds over time.

Going beyond aggregate hit rates, trial-level analyses of effect sizes and temporal structure will be critical for revealing subtle patterns obscured in pooled data. Addressing biases from optional stopping may require sequential or adaptive statistical approaches that legitimately accommodate ongoing data collection. Combining these methods with signal detection theory (SDT) can separate perceptual sensitivity from decision criteria while remaining valid under optional stopping. SDT offers a powerful framework for assessing whether participants extract

information even when guessing incorrectly, helping determine whether anomalous cognition reflects genuine information transfer or shifts in decision-making (Storm et al., 2010; Utts, 1991; Anderson, 2015; Román et al., 2022).

Monte Carlo simulations also provide valuable baselines for modeling complex stopping behaviors, sequential dependencies, and noise influences that violate trial independence (Varvoglis et al., 2019). Measures such as autocorrelation can reveal whether hits and misses cluster over time, clarifying whether effects arise from session-level dynamics rather than isolated trials (Fischer & Whitney, 2014).

More elaborate temporal analyses remain for future work. Sequential Bayesian updating could track how evidence for above-chance performance accumulates trial by trial within each user, without the Type I inflation of repeated frequentist testing. Cross-session stability analyses could test whether users who perform above chance in one session continue to do so in subsequent sessions, distinguishing stable individual differences from session-level noise.

Future studies should incorporate demographic, cultural, and biological moderators (e.g. race, sex, handedness, social role) and psychological factors such as belief systems and personality traits to test whether these systematically influence psi expression (Storm et al., 2010). Expanding recruitment beyond psi-interested populations and integrating these contextual layers will move the field toward a more relational understanding of psi as a human, rather than purely statistical, phenomenon. Continued methodological refinement and broader collaboration will be essential for translating insights from large-scale datasets into the next generation of study designs.

Conclusions

Analyses of over 25 million QRV trials revealed that performance dynamics reflect complex interactions among stopping behavior, task context, and individual variability. Group-level results were driven by optional stopping, while exploratory top-performer analyses suggested a context-dependent interplay among belief, experience, and meditation that influenced performance only through their joint interactions with task experience and stopping behavior, rather than as independent predictors. These findings underscore that psi, if present, may emerge episodically and in interaction with behavioral states, calling for preregistered, fixed-length, participant-level designs capable of distinguishing subtle signals from the noise of human behavior.

Acknowledgments

The authors thank the Institute of Noetic Sciences (IONS) for hosting and maintaining the GotPsi platform and for providing administrative and technical support over the years of data collection. We are especially grateful to the thousands of participants who engaged with the online task and contributed to this long-term citizen-science project. The authors also acknowledge the assistance of IONS research staff and volunteers who supported aspects of data management and manuscript preparation.

Disclosures

Author Contributions

Helané Wahbeh conceptualized the study, developed the preregistration and analytic plan in collaboration with Michael Kriegsman, and led the interpretation of results and manuscript preparation.

Michael Kriegsman performed all data cleaning, statistical analyses, and figure generation, and contributed to data interpretation and manuscript editing.

Beth Glick contributed to the drafting and revising of the Introduction and Discussion sections and assisted with manuscript preparation through multiple iterations.

Arnaud Delorme co-managed the GotPsi website and dataset, reviewed the preregistration and analytic framework, and provided substantive feedback on the final manuscript.

Dean Radin created and maintained the GotPsi website, contributed to the long-term data collection, reviewed the preregistration and statistical plan, and provided critical revisions of the manuscript.

AI Use Disclosure

Portions of this manuscript were prepared with the assistance of OpenAI's ChatGPT, which was used to support drafting, summarization, and editorial refinement of text throughout the project. The authors guided all content generation and are solely responsible for the final versions of all analyses and interpretations.

Correspondence concerning this article should be addressed to Helané Wahbeh, ND, MCR, Institute of Noetic Sciences; 7250 Redwood Blvd, Ste 208, Novato, CA 94945-327; hwahbeh@noetic.org

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This research received no external funding.

Open Data

Data processing code and analysis scripts will be shared upon publication (with privacy safeguards).

References

Anderson, N. D. (2015). Teaching signal detection theory with pseudoscience. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00762

Baptista, J., Derakhshani, M., & Tressoldi, P. E. (2015). Explicit anomalous cognition: A review of the best evidence in Ganzfeld, forced choice, remote viewing and dream studies. In E. Cardeña, J. Palmer, & D. Marcusson-Clavertz (Eds.), *Parapsychology: A handbook for the 21st century* (pp. 192–214). McFarland & Company, Inc., Publishers.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Cardeña, E. (2018). The experimental evidence for parapsychological phenomena: A review. *American Psychologist, 73*(5), 663–677. https://doi.org/10.1037/amp0000236

Dalkvist, J., Mossbridge, J., & Westerlund, J. (2014). How to remove the influence of expectation bias in presentiment and similar experiments: A recommended strategy. *Journal of Parapsychology, 7*8(1), 80–97.

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience, 17*(5), 738–743. https://doi.org/10.1038/nn.3689

Hyman, R. (1995). Evaluation of program on "anomalous mental phenomena." *Journal of Parapsychology, 59*, 321–351.

Jahn, R. G., & Dunne, B. J. (1987). *Margins of reality: The role of consciousness in the physical world*. Jovanovich.

Kennedy, J. (2001). Why is PSY so elusive? A review and proposed model. *The Journal of Parapsychology, 65*, 219–246.

Kennedy, J. (2003). The capricious, actively evasive, unsustainable nature of psi: A summary and hypotheses. *The Journal of Parapsychology, 67*(1), 53–75.

Kennedy, J. (2013). Methodology for confirmatory experiments on physiological measures of precognitive anticipation. *The Journal of Parapsychology, 77*(2), 237.

Kennedy, J. (2014). Letter on methodology for presentiment studies. *Journal of Parapsychology, 78*(2), 273–274.

Kennedy, J. (2016). Is the methodological revolution in psychology over or just beginning. *Journal of Parapsychology, 80*(2), 156–168.

Kripal, J. J. (2019). *The flip: Epiphanies of mind and the future of knowledge*. Bellevue Literary Press.

May, E. C., Utts, J., & Spottiswoode, S. (1995). Decision Augmentation Theory: Toward a model of anomalous mental phenomena. *Journal of Parapsychology, 59*(3), 195–220.

Mossbridge, J., & Radin, D. (2018). Precognition as a form of prospection: A review of the evidence. *Psychology of Consciousness: Theory, Research, and Practice, 5*(1), 78–93. https://doi.org/10.1037/cns0000121

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Palmer, J. (1985). *An evaluative report on the current status of parapsychology*. U. S. Army Research Institute for the Behavioral and Social Sciences.

R Development Core Team, R. (2023). R: A language and environment for statistical computing [Computer software]. *R Foundation for Statistical Computing*. http://www.R-project.org/

Rhine, J. B. (1934). *Extra-sensory perception*. Boston Society for Psychic Research.

Román, C. A. F., DeLuca, J., Yao, B., Genova, H. M., & Wylie, G. R. (2022). Signal detection theory as a novel tool to understand cognitive fatigue in individuals with multiple sclerosis. *Frontiers in Behavioral Neuroscience, 16,* 828566. https://doi.org/10.3389/fnbeh.2022.828566

Storm, L., & Tressoldi, P. E. (2020). Meta-analysis of free-response studies 2009-2018: Assessing the noise-reduction model ten years on. *Journal of the Society for Psychical Research, 84*(4), 193–219.

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010). Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin, 136*(4), 471.

Storm, L., Tressoldi, P. E., & Di Risio, L. (2012). Meta-analysis of ESP studies, 1987-2010: Assessing the success of the forced-choice design in parapsychology. *Journal of Parapsychology, 76*(2), 243–273.

Tressoldi, P. E., & Storm, L. (2024a). Stage 2 Registered Report: Anomalous perception in a Ganzfeld condition-A meta-analysis of more than 40 years investigation [version 4; peer review: 2 approved, 1 not approved]. *F1000Research, 10*, 234. https://doi.org/10.12688/f1000research.51746.4

Tressoldi, P. E., & Storm, L. (2024b). The myth of the decline effect in psi research: The empirical evidence. *Journal of Scientific Exploration, 38*(3), 461–465. https://doi.org/10.31275/20243313

Utts, J. (1991). Replication and meta-analysis in parapsychology. *Statistical Science, 6*(4), 363–378.

Utts, J. (1996). An assessment of the evidence for psychic functioning. *Journal of Scientific Exploration, 10*(1), 3–30.

Varvoglis, M., Bancel, P.A., Bailly, J.-P., Boban, J., & Ahmed, D.S. (2019). The Selfield: Optimizing precognition research. *Journal of Parapsychology, 83*(1), 13–24. https://doi.org/10.30891/jopar.2019.01.02

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H.L.J., & Kievit, R.A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wahbeh, H., Radin, D., Delorme, A., & Kriegsman, M. (2025). Probing top performers in a forced choice precognition task [Preregistered study]. *KPU Study Registry*. https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_Registry_1094.pdf

Wahbeh, H., Radin, D., Mossbridge, J., Vieten, C., & Delorme, A. (2018). Exceptional experiences reported by scientists and engineers. *Explore: The Journal of Science and Healing, 14*(5), 329–341. https://doi.org/10.1016/j.explore.2018.05.002

Zusammenfassung

Untersuchung von Top-Performern bei einer Forced-Choice-Hellsehaufgabe

Diese präregistrierte Studie analysierte mehr als 25 Millionen Durchgänge einer webbasierten „Forced-Choice"-Remote-Viewing-Aufgabe, um Muster der hellseherischen Leistung bei allen Teilnehmern sowie bei einer Untergruppe der besten Teilnehmer zu untersuchen. Auf aggregierter Ebene entsprachen Durchgänge, die nach den vier geplanten Längen (5, 10, 25 oder 100 Trials) endeten, den Zufallserwartungen. Im Gegensatz dazu zeigten Durchgänge mit optionalem Abbruch („optional stopping") systematische Schwankungen: Kurze Durchgänge (1–3 Versuche) lagen über dem Zufallsniveau, bevor sie abfielen; Durchgänge mit einer Länge von 11–19 Versuchen lagen unter dem Zufallsniveau; und Durchgänge ab einer Länge von 20 Versuchen zeigten wiederkehrende Spitzenwerte über dem Zufallsniveau bei jeder fünften Durchgangslänge (z. B. 30, 35, 40, 45, 50), die bei mehr als 80 Versuchsdurchgängen abnahmen. Eine an die empirische Abbruchverteilung angepasste Monte-Carlo-Simulation verdeutlichte, inwieweit diese Muster allein durch ein optionales Abbruchverhalten reproduziert werden konnten, wobei ein Großteil des Musters der Anzahl der Versuchsdurchgänge – einschließlich des Tiefpunkts bei 11–19 und der Variabilität bei runden Zahlen – innerhalb der simulierten Nullhüllkurve lag. In explorativen Analysen der Top Performer – post-hoc definiert als jene 1.235 Nutzer (2,64 %), die das unkorrigierte Zufallsniveau übertrafen, nachdem kein Nutzer das präregistrierte, FDR-korrigierte Kriterium erfüllt hatte – wurden der Glauben an Psi, frühere präkognitive Erfahrungen, Meditation, die Gesamtzahl der Versuchsdurchgänge sowie der

optionale Abbruch als Prädiktoren untersucht. Der optionale Abbruch war der Prädiktor, der sowohl bei den ersten Versuchen als auch über alle Versuche hinweg am konsistentesten mit Treffern assoziiert war; zudem zeigte er Interaktionseffekte mit den Glaubensüberzeugungen, früheren präkognitiven Erfahrungen und Meditation in Verbindung mit der kumulativen Erfahrung in der Aufgabe. Die Effektstärken waren gering (Δp und Cohens d lagen für die meisten Prädiktoren nahe null), und die Ergebnisse werden als explorativ interpretiert. Die Befunde legen nahe, dass die Ergebnisse auf Gruppenebene primär den optionalen Abbruch und damit verbundene Verhaltensdynamiken widerspiegeln, während die Analysen der Top Performer differenziertere, wenn auch schwache, kontextabhängige Zusammenhänge zwischen Glaubensüberzeugungen, Erfahrung und Verhalten aufzeigen. Diese Befunde unterstreichen die methodischen Herausforderungen groß angelegter, offener Online-Tests sowie den Wert präregistrierter Ansätze auf Teilnehmerebene in Kombination mit Benchmark-Simulationen, um Verhaltensartefakte von potenziellen Psi-Signalen zu unterscheiden.

Schlüsselbegriffe: Hellsehen, Forced-Choice-Aufgabe, Optional Stopping, Top-Performer, individuelle Unterschiede, Präregistrierung