Journal of the Society for Psychical Research, 2021
Vol. 85, No. 2, 65–90,
www.spr.ac.uk

ASSOCIATIVE REMOTE VIEWING PROJECTS: ASSESSING RATER RELIABILITY AND FACTORS AFFECTING SUCCESSFUL PREDICTIONS

By Debra Lynne Katz, Igor Grgić, Patrizio Tressoldi and T.W. Fendley

ABSTRACT

Associative Remote Viewing (ARV) is a psi-based methodology used by individuals and for-profit organizations to predict such things as sporting-event outcomes, stock market moves, and for research purposes. Documented studies have shown the successful application of psi using ARV to predict future events, leading to profits, and unsuccessful applications, leading to losses. To better understand the contributing factors, 86 completed ARV trials, which included 220 remote viewing transcripts for individual sporting or financial events, were collected. Three teams of judges operating under blind conditions — some working independently, some working as teams — repeated the process of judging, scoring, and predicting, while keeping all other variables stable. To gauge inter-rater reliability, the new scores and predictions were compared to the original scores and predictions, as well as to each other. Rating variance was clearly demonstrated. Judges were in 100% agreement in only six (6.9%) of 86 trials. In seventeen trials (19.7%), eight of nine judges agreed with each other. Original judges did better than all new judges, and judges with more experience obtained statistically significant higher hit rates than less experienced judges. The results were virtually the same for the two ranking scales used. This project points to a variety of factors in need of further testing, both in future ARV projects and in parapsychology projects that involve independent judging of tasks and photosets.

Introduction

Remote viewing (RV) refers to the ability of a perceiver (a "viewer") to describe or give details about a target that is inaccessible to normal senses due to distance, time, or shielding. While some use the term as a synonym for clairvoyance, Ingo Swann (1993) defined remote viewing not as a type of psychic phenomenon, but rather as a type of experiment in which intuitive faculties could be put to use. From 1970 to 1973, Swann and his fellow researchers at the American Society for Psychical Research (ASPR) carried out thousands of double-blind trials, using an assortment of target materials and approaches to explore concepts related to nonlocal perception, out of body travel, and telepathy (Mitchell, 1987). Swann and yet another team of researchers, who developed the psycho-energetics lab at Stanford Research Institute (SRI) in coordination with U.S. military and other intelligence agencies from 1972 to the mid-1990s, further explored RV (May & Marhawa, 2018; Targ, 2007).

Early published reports of remote viewing experiments at SRI — followed by those at Princeton Engineering Anomalies Research Laboratory (PEAR) that attempted to replicate SRI's results over a period of two decades — demonstrated that remote viewing could be used to produce descriptive, accurate, and useful accounts and sketches of real locations, events, objects, people, and photographs (Puthoff & Targ, 1976; Puthoff et al., 1977; Puthoff & Targ, 2005; Swann, 1996). Following the declassification and disbandment of these programmes, remote viewing moved from institutional settings into the public sector. Former research and military remote viewing personnel and their students, and subsequently practitioners, businesses, and social groups, have attempted to maintain the integrity of remote viewing, which — unlike many other intuitive practices — was built on scientific principles (Katz et al., 2018).

Remote viewers access their intuitive-based perceptual and sensory abilities, including, but not limited to, clairvoyance, via intentional and systematic processes. Some of these, such as Controlled Remote Viewing (Smith, 2014, Williams, 2019) and its many derivatives (Knowles, 2017; Vivanco, 2016), are highly organized and rely on somatic responses. They have been designed to produce a lot of data while decreasing analytical interference from logic-based processes. Other approaches, such as Extended Remote Viewing (Atwater, 2001; Herlosky, 2015, Morehouse, 2011), are less structured and allow the viewer to move into a deeper meditative state similar to that of shamanic journeying (Storm, 2019).

While remote viewing projects vary in their design and procedures, they often involve blinding protocols, randomization, separation of roles, and the recording of impressions during the sessions onto paper transcripts either by participants themselves (Nobel, 2018) or those monitoring/interviewing them (Schwartz, 2019; Williams, 2017). Remote viewing data usually includes words and sketches, and sometimes maps or three-dimensional modelling (Smith, 2005).

Remote viewing can be used in a variety of ways, depending on the type of information sought. To be useful, sessions must be properly recorded, reported, and analyzed. Early analysis methods for remote viewing within a laboratory setting mirrored those used within Ganzfeld experiments (Storm et al., 2010; 2012) and dream ESP experiments (Storm et al., 2017). They often involved matching tasks in which a remote viewer's impressions would be compared to several photos in the set, with the aim of choosing the correct photo. This practice allowed for easy statistical analysis, although a major criticism was, and remains, that it reduced the potentially rich qualitative data into a single "hit" or "miss" result. Another drawback is that the extra photos comprising the judging set become a source of noise or distraction for the participant. Brown (2005) and others have observed that remote viewers may provide excellent descriptions of the wrong photo, speculating this may be because the other photos were more interesting to the viewer or easier to describe or name. This has been referred to as "displaced psi" or "displacement". Brown and others have observed this can happen with both participants acting as self-judges or with independent judges, but suggest it is less likely to occur with independent judges.

Associative Remote Viewing

Stephan Schwartz created Associate Remote Viewing (ARV) as a predictive method for real-life events with binary or multiple potential outcomes. He ran experiments at his Mobius lab and tested ARV in partnership with colleagues from SRI's psycho-energetics programme (Schwartz, 1977–1979; Schwartz, 2007; Schwartz, 2020). ARV has a specific and rather complex double-blinded set of procedures used to bypass the typical pitfalls of having remote viewers tune into potential outcomes of a future event (such as sporting events or stock market fluctuations) in which the possible winning options are too familiar or too similar to each other for judges to determine which outcome is being described (Katz et al., 2019).

For example, if the weekly goal is to predict the winner of a football game, a remote viewer directly tuning in to the game's end (whether blind to the task or aware of it) might perceive helmeted, muscular men in tight uniforms running across a finish line with an egg-shaped ball. While descriptive, this could obviously describe both teams. ARV protocols were developed to circumvent these kinds of challenges and to allow for blind, repeated trials over time, essentially turning what could easily become forced-choice tasks between binary outcomes into free-response tasks. This is accomplished through pairing photos different from each other in every aspect to two possible outcomes, and then having a remote viewer tune into the single photograph they will see in the future (called their future feedback photo).

Understanding a Typical ARV Trial

In ARV trials, viewers don't need to know anything about the overall project itself, nor do they need to be privy to the individual events or possible outcomes. All they need to know is their task, which is to use their psi abilities to describe a photo they will see at a specified future date. Viewers may approach this task using a variety of processes. Prior to the event, however, they must provide the judge a written transcript consisting of words and sketches. Each target photo is paired with a potential binary outcome before the event's start time. The judge must decide which photo the viewer is likely to see in the future as feedback on the winning side.

To do this, the judge compares the transcript to one photo at a time and scores it according to a predetermined scale, such as the SRI seven-point confidence ranking scale (Appendix B). This enables judges to specify how confident they are a transcript matches a particular photo in the set. Ideally, one photo would have a high score and the other a very low score. Low scores for both photos indicate little psi is present. High scores for both photos indicate something has gone awry, such as photos in the set being too similar to each other. Sometimes this happens if the game is cancelled or if there is a tie, which would mean the viewers don't have a single target to tune into. It's been noted/theorized that if one photo is particularly numinous, while another is dull or even confusing, viewers may tune into the wrong photo in the set, although this assertion has not been formally tested (Brown, 2015).

Predictions vs. Passes

Sometimes multiple viewers produce transcripts for a single event. In such cases, a project manager may take into consideration whether the transcripts collectively describe one photo over another. If confidence is high, a prediction will be issued. If confidence is low, a "pass" will be called (Katz et al., 2018). Passes may also be made due to procedural errors. For example, if the event is a stock trade, occasionally the trader might make an error setting up the trial or fail to enter a timely trade into the online system.

In applied ARV projects, passes are typically not factored into the overall statistics beyond being recorded (Katz et al., 2018). Only outcomes of actual predictions are tallied to arrive at the hit/miss ratio. In such projects, the end goal generally is to use the prediction to make money, and no money is won or lost with a pass. While passing can prolong a project with a set number of predictions, it has been a commonly held assumption by ARV applied project managers that passing can guard against misses, and ultimately lost funds, in trials that are wagered on (McMoneagle & May, 2016).

Whether a formal prediction or a pass is issued, following the event the manager sends a feedback photo to the viewer. The feedback photo is always the one associated with the winning outcome. This completes the "feedback" or "retrocausal loop," closing out that trial (Rosenblatt et al., 2015). To date, it's been felt that — particularly in ARV trials — feedback is important since it sets apart that which the viewers are describing from the photos in the judging set.

However, a recent study suggested otherwise. Müller, Müller, and Wittmann (2019) attempted to predict the binary (up vs. down) course of the German stock index DAX with the Associative Remote Viewing (ARV) method. Thirty-eight of 48 predictions were correct, resulting in a highly significant hit rate of 79.16% (p = 2.3 × 10–5, binomial distribution, $B_{48}(1/2)$; $z = 3.897$; ES = 0.56). They compared trials for when viewers received feedback vs. when they simply had focused on the photo attached to the winning outcome without feedback and did not find a difference between the two groups. This latter finding demonstrates that the widely accepted view about the importance of feedback in ARV trials warrants further testing. Still, we were not able to ascertain whether their viewers were directed (or directed themselves) to focus on the feedback itself, or simply on the photo attached to the winning outcome, which might alter the results. Another finding was that session quality correlated with volatility of the stock index, in that "the viewer's perceptions were clearer and less ambivalent when the stock index also had a larger point difference at the end of the prediction period" (p. 326).

Prior ARV Research

The present researchers have provided an extensive history of ARV research in two recent studies (Katz et al., 2018; Katz et al., 2019). Therefore, included here are highlights of only a few relevant ARV studies regarding financial gain and loss. In 1982, Keith Harary and Russell Targ used ARV to make nine consecutive forecast changes in closing prices of the silver futures market, yielding earnings of more than $100,000 (Harary & Targ, 1985). The following year, Harary and Targ repeated the experiment, but were unsuccessful across all nine trials. Some speculated that shortening the time interval between trials, which resulted in the remote viewer having to perform a subsequent trial before receiving feedback for the preceding one, may have impaired performance (Houck, 1986; Targ, 2012). In a recently published interview (Katz & Bulgatz, 2019), Targ explained that in at least one trial for this experiment, during the judging process they discovered the photos in the judging set were too similar. Their project design did not specify a way to deal with such a situation, meaning at that time they didn't issue passes, so they were forced to issue a prediction that had to be wagered on. Also the researchers failed to heed the viewer's sense that their choice was not correct.

In 1985, Targ and his co-researcher repeated their experiment by using a "redundancy protocol". Viewers participated in only one ARV trial per day. Passes were called if a score of 4 was not reached on a 0–7 confidence rating scale with a two-point spread between scores, and if the remote viewers both "accurately described photos in discrepant directions". Twelve of 18 trials resulted in predictions and from these, seven forecasts were recorded as trades even though no monies were wagered. All but one of these were correct (Targ et al., 1995).

In 1982, Puthoff used ARV to predict the daily outcome of silver futures contracts for 30 consecutive days. Seven remote viewers took part in 12 to 36 trials over the entire series. Each day predictions were made using consensus judging. Twenty-one of the 30 trades were hits, yielding profits of $250,000 (Puthoff, 1984).

In 2012, Smith, Laham, and Moddel (2014) conducted an experiment with ten University of Colorado college students, all inexperienced remote viewers, who pooled their responses together to make seven successful predictions of the Dow Jones Industrial Average (DJIA) of seven attempted — (binomial probability test, $p < .01$). They used an ARV protocol, but created their own three-point ranking scale (see Appendix A), which enabled judges to determine the strength of the remote viewing data and whether the data was strong for one photo or more than one. Their $10,000 actual investment yielded a $16,000 gain, with a total balance of $26,000 at the end of Trial 5 (Smith et al., 2014).

In another project, Greg Kolodziejzyk (2015), acting as a single operator over a 13-year period, used a unique computer-based approach to the ARV protocol. His project combined remote viewing, logic, and knowledge of the stock market, and yielded a profit of $146,587.30. Ongoing informal research is being done by the Applied Precognition Project (APP), an organization comprised of individuals and groups that use ARV predictions for wagering and stock trades, as well as for educational and research purposes (Rosenblatt et al., 2015).

APP's activities could be characterized as Participatory Action Research (Kindon et al., 2007). APP groups and members interact with parapsychological researchers, sometimes on a formal basis (e.g., to conduct formal experiments) and sometimes on an informal basis (e.g., at biannual conferences, in webinar presentations, social media groups, and to conduct informal experiments). APP members follow blinding and randomization procedures, and keep track of basic statistics, all per traditional parapsychological standards.

Some of APP's 1,163 members (Rosenblatt, 2019) are very experienced at using their nonlocal perception to describe photographs they will see in the future, as well as at creating photo pairings or sets, rating sessions, and overall project management (Fendley, 2016; Rosenblatt et al., 2015). One APP member created a new scoring system (Poquiz, 2014) and another created publicly available software for making binary and multiple-choice predictions (Grgić, 2019).

In 2018, Katz, Grgić and Fendley reported on a year-long project conducted by APP in which 60 remote viewers contributed 177 predictions generating 240 foreign exchange (FOREX) executed trades. It resulted in an actual loss of $52,186 in funds pooled together by viewers and APP's management. The project relied heavily on the Kelly wagering strategy, which based the size of wagers on a composite hit rate of more than 60% for existing APP groups. High wagers on misses resulted in significant losses early in the project. The researchers' subsequent examination of the data showed many of the groups participating in the year-long project previously had hit rates below chance for similar financial predictions. The authors also found that — as in the early Harary and Targ (1985) experiment — "too many predictions may have been made in too short a time-span" (p. 44).

Most recently, Katz, Smith, Graff, Bulgatz and Lane (2019) conducted a year-long, double-blind study using dreaming as a precognitive tool within an ARV protocol. A cohesive group of seven experienced remote viewers (the APP Sublime RV group) participated in 56 trials in which they attempted to have precognitive dreams of a future feedback photo. Their protocols, mirroring RV protocols, required them to produce a written transcript upon awakening that included descriptor words and sketches. A single judge served as project manager. She rated the transcripts using the SRI seven-point scale, pooled the transcripts for each trial to make an overall group prediction, and wagered on a sporting event. Five of the seven remote viewers/dreamers consistently produced dreams at will, resulting in 278 transcripts. Two dreamers had high individual hit rates (76 percent on 17 trials and 64 percent on 25 trials). With 56 trials, 28 group predictions yielded 17 hits and 11 misses, which a binomial test showed insufficient to reject the chance hypothesis. Nevertheless, the overall monetary gain was almost 400 percent of the initial stake.

Inter-rater Reliability in RV and ARV Projects

According to Stemler (2004, p.9), "inter-rater reliability is one of the most important concepts of educational and psychological measurement. Without demonstrating that two independent judges can be reliably trained to rate a particular behavior, our hope for achieving objective measurement of behavioral phenomena is diminished." Milton (1997; 1985) conducted a survey of judging practices used in 85 free-response studies reported from 1964 through 1985 in five parapsychology journals. In only about 4 percent of these studies were judges trained in any systematic way. She concluded that either having no instructions or a failure to report what instructions were given might imply a lack of importance being attached to the judging process within the field.

Our study examined some of the judging factors affecting ARV. In remote viewing projects that require comparing transcripts to photos to determine the best match, variability may exist on a micro and macro level. At the far end of the micro level would be the evaluation of a transcript's words, phrases, and sketches. Raters may define words differently. Many words in the English language have multiple meanings. For example, the word "light" could refer to luminosity or weight or shading.

In remote viewing, sketches may very closely represent detailed aspects of a photograph, or they may be simplistic. For example, a viewer draws a sketch of a rectangular shape with four circles attached and other details resembling a steering wheel, doors, and headlights, accompanied by words like "vehicle." That may be easy for a rater to judge as matching the feedback photo of an automobile, but what if the viewer drew a rectangle floating above two circles? One rater may feel the circles represent the car's wheels and credit it as a match, while another may feel the sketch is not descriptive enough.

The judge must take each word and shape into consideration to give an overall score based on the rating scale used. If it is a binary choice (yes/no, up/down, etc.), the judge must compare the transcript to both photos to determine which one the viewer was most likely describing as the future feedback photo (paired to the winning outcome). In theory, a strong transcript should make this matching task easy, and often it is. But what if the viewer listed the word "red" and both photos have this colour? Even with a simple transcript and a simple photograph, a rater faces dozens of perceptual tasks and decisions, and some of these invariably lead to attempts to interpret what the viewer meant or experienced at the time.

At the far end of the macro level of assessment for ARV projects are decisions of whether to issue a prediction or call a pass, and ultimately whether to place a wager on that prediction. To complicate things even more, many RV and ARV projects use more than one participant per trial. Ideally participants' scores will favour the same photo. When this is not the case — particularly if there are many viewers — the judge may complete additional analyses to arrive at a prediction.

While researchers have not systematically analyzed the extent of the problem of inter-rater or inter-analyst variability, they have made plenty of observations and attempted to overcome it (Honorton, 1975; Humphrey et al., 1986; Humphrey et al., 1988; Jahn et al.,1980; Targ et al., 1977). May et al. (1990, p. 194) concluded, "If multiple analysts are used, additional problems arise concerning inter-analyst reliability. If an individual analyst judges a number of responses in a series, within-analyst consistency becomes an individual problem". In addition to inter-analyst variability, photo selection and photo orthogonality are also often cited as playing a role in misses within formal remote-viewing studies. Photo orthogonality is defined as having selected photo targets that are as dissimilar from each other as possible (May et al., 1990).

While Child was focused on dream ESP research, his observations regarding nonindependence could easily apply to remote viewing projects with multiple viewers. Not only was he concerned about reliability/variability between raters, but he also observed how the ordering of photos in a set (earlier decision-making) could impact a rater's subsequent choices. He noted, "If a judge is presented with a set of transcripts and a set of targets and is asked to judge the similarity of each target to each transcript, the various judgments may not be completely independent. If one transcript is so closely similar to a particular target that the judge is confident of having recognized a correct match, the judge (or percipient, of course) may minimize the similarity of that target to the transcripts judged later...Nonindependence would create no bias toward either positive or negative evidence of correspondence between targets and transcripts, but it would alter variability and thus render inappropriate some standard tests of significance" (Child, 1985, p. 1223).

Attempts to Increase Inter-intra Rater Reliability

Among the challenges of judging remote viewing transcripts are how much consideration to give correct data compared to incorrect data, and how much value to place on single words, simple sketches, or common words vs. highly specific or less-common words and complex sketches. Over the years, researchers involved in SRI's psycho-energetics programmes sought to develop analysis methods that could provide greater confidence that the impressions of any individual transcript were not just random bits of data or logic-based content, but rather coming from what would possibly be considered part of the remote viewer's extrasensory perceptual and sensory faculties. Also of concern was inter- and intra-rater reliability. This led to the development of what is commonly referred to as the SRI seven-point confidence-ranking (CR) scale, also known as the "Targ" scale (Targ et al., 1995).

As presented in Appendix B, to receive a CR score of 5 required "good matching, unambiguous, unique" impressions; a 6 required "good analytic (naming the target) with relatively little incorrect information"; and a 7 required excellent correspondence, "naming the target, with no incorrect data". Therefore, a high rating indicated the presence of psi and less likelihood the results were due to chance than CR scores of 3.5 or less.

Another approach to improve rater reliability developed by May et al. (1990) focused on decreasing the role of humans in the judging process. After reviewing a viewer's transcript and answering a series of questions on a coding sheet, the human coder input responses into a computer programmed to issue a figure of merit (FOM) score. A high FOM score could only be achieved if one photo had strong correspondence and the other low correspondence. The FOM would not be high if both photos had strong correspondence or if both had poor matches. May and others have used the program for both RV research and applied ARV projects (Bierman, 2016).

While early preliminary trial results seemed promising, formalized studies producing conclusive results have not been forthcoming. The present researchers have been personally involved in projects involving this system, and results have varied. Much is still left to the human coder's discretion, particularly when it comes to evaluating the remote viewer's sketches. Essentially, the same inter-rater variability issues remain.

Jahn, Dunne, and Juan at PEAR (1980) attempted conceptual replication of SRI's remote viewing program, calling it "precognitive remote perception" (PRP) since it involved multi-sensory systems and faculties, not just visual ones. This program spanned two decades, consisted of more than 650 trials, and demonstrated overall significance. During this time, a variety of adjustments to judging, viewing, and analysis protocols were made, with exploration of variables such as "ex post facto vs. participant-encoded descriptions, agent-chosen versus randomly-assigned targets, single vs. multiple percipients". Ultimately, they found "most of these factors were not strong modulators of the scoring" (Nelson, 2017). According to Nelson:

> Originally, the efficacy of the remote perception was determined by human judging — rating a set of targets including the actual target and several decoys. PEAR replaced this with a protocol where the "agent" at the scene filled out a binary descriptor list indicating whether each of 30 elements were present or absent from the scene. The "percipient" encoded his or her experience in a narrative and sketches, but also using the same descriptor list, and the subsequent analysis compared the two lists yielding a score which reflected the relative accuracy of the perception. (p.1)

Modern Remote Viewing Research and Issues in Rater Reliability

Inter- and intra-rater reliability remain a concern for researchers currently involved in ARV projects. After running a series of informal tests and observing continued discrepancies in raters' scores, Alexis Poquiz designed a method of scoring that takes rater variability into account. Every word and sketch in a transcript is evaluated to arrive at an overall hit/miss ratio. When asked about his ranking scale, Poquiz (2012) explained in private email correspondence:

> The Poquiz Methodology has developed into a computational approach to qualitatively and quantitatively evaluate a remote viewing session. At its very core, judging remote viewing sessions is subjective because judges may differ in their evaluation of a given perception. Arriving at a true score is not possible; we can only approximate the score of a session. The Poquiz Methodology acknowledges this subjective nature by borrowing on the concepts of variance, standard deviation, and uncertainty. Rather than providing a definitive score, it produces a base score and establishes a range that attempts to isolate the true score, between a defined minimum and maximum... (p. 1)

The Poquiz Methodology was first reported on the internet, social media, at various remote viewing conferences, and formally in a project conducted by Katz, Beem and Fendley (2015). Remote viewers were tasked with describing a microscopic organism, specifically a bacteriophage. Three biologists were recruited to rate their sessions using the Poquiz system of scoring (Poquiz, 2014), which required them to individually assess every word and sketch, and then to subtract the number of correct responses from incorrect ones to derive an overall hit rate.

The data sheets were inadvertently lost, and about a month later, the raters were asked to repeat their rating tasks. Soon after these were completed, the original data sheets were found. The researchers compared the two and found all raters had changed some of their responses. As many as 50 percent of one rater's responses on the two sheets were different.

The Poquiz system has undergone several iterations (Katz & Knowles, 2021). It was used in a remote viewing project designed to predict the outcome of a U.S. presidential election (Katz et al., 2015). In this project, the researchers served as raters. Their protocol mandated them to always do their ratings together, and to be in 100 percent agreement on each descriptor before moving on to the next. They found this process to be very laborious but revealing. Some of the most contentious words seemed "relative" to the embodied experience, perspective or characteristics of the remote viewer. For example, words like "tall" or "lighter complexion" or "compassionate" could have different meanings for different people. While the project's sample size was too small to determine significance, it revealed specific challenges in judging remote viewers' perceptions regarding particular people.

ARV and Variance in Judging

In early 2016, Igor Grgić, one of the present researchers, conducted two informal studies of ARV trials, comparing the scores given by multiple judges. I. Grgić (personal communication, July 20, 2016) found:

> There are large differences. The scoring was based on the SRI seven-point scale. For almost all of the transcripts, the gap between the lowest score and the highest score (assigned by the judges when judging the same photo target) was 2.42 on average. The maximum score difference that was found was 4.0 and a gap of 2.5 is a rule of thumb...on a group level, final group predictions were not in 100% agreement by all judges. Mostly, the majority of judges agreed and called a prediction for the same side, some called a pass, but once there was a call for the opposite side. On an individual transcript level, most of the time all judges agreed on the side they had picked. If not in agreement, then in most cases the judges would pass on a particular individual pick. There was also an example of one judge picking the opposite side from the other judges… of course all this is based on a low sample size and further research is needed.

From the above-referenced studies, it is clear several researchers developed a variety of systems in response to their personal and collective observations of inconsistencies and problems in judging. The present project demonstrated the extent of this problem across multiple trials.

Objective

This study's overall objective was to examine factors and practices leading to successful predictions (hits) in ARV trials. In particular, the focus was to assess the extent of consistency in judging across multiple ARV trials. This type of information can guide not only ARV project managers and researchers, but also those designing experimental parapsychology projects that include independent-judging practices for determining best matches between transcripts and photos.

Unlike many projects that seek to test participants' abilities by running them through a series of new psi-related trials, the present project sought to test the efficacy and consistency of raters by having them rejudge and issue predictions for already-completed ARV trials, while operating under double-blind conditions. Outcomes of these trials were already known, and researchers had access to all viewers' transcripts, photo pairings, and original raters' scores. That allowed a comparative analysis of judges' scores, predictions, and outcomes related to both the original and the new raters. It also allowed exploration of a number of variables that might affect the end goal of making successful predictions.

Preregistration

The project was preregistered with the Koestler Parapsychology Unit in April 2017. Seven hypotheses were registered. All of these were tested, as specified in the original registration, with the exception of Hypothesis seven, which was deemed too complex to evaluate and was therefore cut from this paper. Another minor change from the preregistered proposal concerned Hypothesis three, which contained two unrelated statements that should never have been combined. Therefore, the second statement was made into Hypothesis four in the current paper, shifting the numbers up.

Following preregistration but prior to any analysis being performed, four more hypotheses were developed and tested. These were conceived after reviewing results of another study — the ARV dream project (Katz et al., 2019). Given the present project was identified as an exploratory study, with the aim to learn from results rather than to prove an effect, it was felt this was acceptable. Therefore, if one compares the present hypotheses to those listed in the registry, one will find that hypotheses one through six are the same, with seven through ten consisting of additional hypotheses. Again, these were tested prior to analysis, and therefore would not be considered post hoc analysis. Only one post hoc analysis was performed, which had to do with comparing prediction rates with judging experience. These three categories of hypotheses — preregistered, post-registration and post hoc — are presented below.

Hypotheses

All comparative statistical analyses of proportions of correct hits were carried out using a test of binomial proportions using the method of approximation via normal distribution available online at http://vassarstats.net/binomialX.html. Comparisons between independent proportions (e.g., Hypothesis two) were carried out following the method suggested by Newcombe and Altman (2000).

Hypothesis one. Differences on prediction level. Based on past informal experiments and a literature review, wide differences in the hit rate will be found in the judging among 1) original predictions, 2) new predictions of six Single New Judges (SNJ), and 3) new predictions of Teams of Two Judges (TTJ).

Hypothesis two. Judging. Original Judges' (OJ) hit rates will be higher than Single New Judges' (SNJ) hit rates using the same method, possibly due to their familiarity with the viewers and history with the remote viewers.

Hypothesis three. Ranking scale performance. The SRI seven-point CR scale will be more effective than the UC three-point scale. While a variety of ranking scales used in applied and experimental ARV projects are worthy of consideration, the SRI seven-point scale continues to be the most popular. It was used by the original project managers/judges who provided data. The second scale we chose to use has been dubbed the UC three-point scale, which was conceived of by researchers who ran a successful series of ARV trials at the University of Colorado with novice remote viewers.

Hypothesis four. Consensus team judging. Consensus team judging will be more effective, resulting in more hits, than individual judging.

Hypothesis five. Individual performance (judges). Some Single New Judges (SNJ) will do better at judging than other SNJs, resulting in higher hit rates. Higher hit rates can be due to such things as: a) passing on a prediction for which the other SNJ had a miss, or b) making a correct prediction (and having a hit) in a particular trial where the other SNJ passed or predicted the opposite/wrong side.

Hypothesis six. Multiple viewers. Predictions based on contributions from multiple viewers (consensus viewing) will yield better results (more hits) than predictions based on a single remote viewer's transcripts.

Hypothesis seven. Individual performance (viewers). Some remote viewers will outperform others even when their sessions are part of a group of viewers, while some will underperform, bringing down the success rate for the overall group.

Hypothesis eight. Spread between scores. The spread between scores for each photo of at least two points on the SRI seven-point scale should yield more hits. Project managers with the Applied Precognition Project informally agree a prediction should only be made when there is at least a two-point difference between scores for each photo. To date, however, this has not been formally tested.

Hypothesis nine. Minimum prediction threshold. Predictions when at least one of the photos earns a minimum confidence ranking of 3.5 or higher on the seven-point scale should yield more hits than vice versa. A CR score of less than 3.5 could be due to chance. For example, if a transcript is compared to one photo and scores two on a seven-point scale, and then is compared to the other photo and scores a three, the informal and largely untested rule is that the manager would pass rather than make a prediction, which means no wagering will occur.

Hypothesis ten. Passes. More passes will produce a higher hit rate (more successful predictions) and a lower miss rate (fewer unsuccessful predictions). Consequently, we expect a positive correlation between passes and hits.

Method

Materials

A call went out to ARV project managers to provide data from completed series of ARV trials meeting the following criteria:

- The original project had to follow well-established scientific protocols that included double blinding.
- It included at least 10 trials/predictions all using the same protocol.
- The protocol had to be fully known and defined for each series.
- All session data used to make original prediction had to be available. This data included: all original transcripts, photo pairings, scores, predictions, and outcomes.
- All viewers had to be trackable. The viewer's name, identifier number or code word had to be attached to their scores so individual performance could be assessed and evaluated.
- All sessions had to be independently judged, meaning the trials could not include self-judging where viewers may have been exposed to all photos in the set.

- Any formal or informal rules applied to the predictions and a statement regarding how closely those were followed had to be accessible.

Although several researchers initially responded, only two project managers provided the required data. One of those was one of the present study's co-researcher (Grgić), who is a long-time ARV project manager. The other manager was co-founder of the Applied Precognition Project (Chris Georges), who had used a single remote viewer, co-researcher of the present study (Katz), in his original series of trials that yielded profits of $1,311 from an initial investment of $200 placed by an independent trader. The latter were interested in discovering if their initial success could be replicated by new judges in terms of the hit-to-miss ratio, which included six hits, two passes, and a single miss. Full data for one of these trials was unavailable so eight trials were included in the present study.

Combined trial data from both managers and five ARV series consisted of 86 separate ARV events/predictions. Forty trials used single remote viewers, who generated one transcript per trial, while 47 included multiple viewers (two to six), whose transcripts were independently scored but then collectively assessed to arrive at a single prediction per trial. The 86 events/predictions consisted of 220 transcripts. The five series of trials selected to be rejudged are described in Table 1. All involved standard binary ARV protocols.

Table 1.

Five series that met the predefined criteria were selected to be rejudged

Series Name	# of Viewers per trial	Original Judge	Methods/notes	# of trials/events
P7B, 2015	Average 4	Igor Grgić	7-pt. Scale, Grgić judging	27 ARV events with 110 transcripts
P7B, 2016	Average 3.5	Igor Grgić	7-pt. Scale, Grgić judging	20 ARV events with 71 transcripts
Red Dwarf, 2014*	One (1)	Igor Grgić	7-pt. Scale, Grgić judging*	14 ARV events with 14 transcripts
Zero-One, 2016	One (1)	Igor Grgić	7-pt. Scale, Grgić judging	17 ARV events with 17 transcripts
APPI Team, 2016**	One (1)	Chris Georges	7-pt. Scale, Georges judging**	8 ARV events with 8 single transcripts

** This series was selected because of its low hit rate.*
*** This series was selected because of its high hit rate, which originally also doubled wagered earnings. Method included a bonded pair of experienced judge and viewer.*

Design

Using double-blind protocols, 10 new judges were recruited to evaluate and score the original transcripts and to issue their own predictions for the original event, while keeping other variables stable. Six of these judges operated independently, while four were placed into teams of two. Three of the six single judges and one of the judging pairs used the same scoring method as the original judges to arrive at their predictions. These judges used the SRI seven-point Confidence Ranking (CR) scale developed at SRI in the early 1970s (Targ et al., 1995) (Appendix B).

The other three single judges and the other team of judges used a different scoring method to arrive at their predictions — the University of Colorado (UC) three-point ranking scale (Smith et al., 2014). It incorporates two separate measures that enable judges to score the transcripts as having a low, medium, or high match with each photo option while indicating if the scores were equally low, high, or the same for each option (Appendix A). The goal was to have eight new sets of scores and predictions to compare to those issued by the judges/project managers during the original trials.

Participant characteristics

Nine of the 10 new judges practiced remote viewing themselves and were, therefore, toward the far end of the openness and belief spectrum, although no formal testing assessed this. The lead researchers knew the judges had prior ARV judging experience, although its extent varied. Ideally, only highly experienced judges with project management experience would have been chosen, but too few volunteered.

One judging pair was comprised of experienced remote viewers/judges who had been friends for at least two years, and the other was a long-time married couple. The husband in this pairing was the only judge who did not have prior judging experience beyond evaluating his wife's transcripts; he also had the least amount of personal remote viewing experience. The wife was an experienced remote viewer and a long-time student of clairvoyance and energy healing, but she had less experience than the other judges with scoring/ranking ARV transcripts.

Judges' preparation

All judges signed a participation, ethics, and confidentiality agreement. All participated in a survey about their past experience with rating remote viewing transcripts. Judges remained blind to the project's overall objective, as well as to the original events, scores, predictions, and outcomes.

All judges underwent at least two separate orientation/training sessions, conducted via webinar, to ensure they understood their tasks. This included individualized instruction regarding their assigned scoring methods and mock group practice judging sessions until consistency had been achieved. Judges were advised their batches would be randomized and different from each other's and all were given strict instructions to not discuss their scores or work with other judges. For each trial, judges were instructed to score every photo on the basis of its own merit in relation to each remote viewer's transcript. They were told to only compare the scores and observe the photos side by side after separately generating scores for both photos.

Judges assigned to use the SRI seven-point ranking scale were advised during training of the general rule to pass unless one photo had a score of 3.5 or higher. These judges were also told it was good practice for predictions to have a two-point spread or better between scores, but since the original judges did not always follow this rule, it was left to the new judges' discretion. Team judges were instructed to come to a consensus with every choice. If they could not come to a consensus after extensive discussion, they were advised to call a pass for that prediction. In order to increase the likelihood that individuals comprising teams would be comfortable expressing opinions that might conflict with their partner's, couples were chosen based on prior relationships.

Judging tasks distributed between the New Judges

Computerized randomization procedures were used so all judges received the same trial data, but in different ordering. Each of the six Single New Judges (SNJ) analyzed and rated all 220 transcripts for 86 ARV events/predictions, thereby generating 86 new ARV predictions. Three used the UC three-point scale, and the other three used the SRI seven-point scale. Within each of the two Team of Two Judges (TTJ), the judges rated together, but independently of the other TTJ. Both TTJ rejudged and rated all 220 transcripts for 86 ARV events/predictions, so each TTJ generated 86 new ARV predictions. One TTJ used the UC three-point scale, and the other TTJ used the SRI seven-point scale. Overall, the rejudging resulted in eight new predictions for each of the 86 original ARV predictions (see Table 2).

Table 2.

Comparison of new judges' predictions to originals

Original or New	Scale Type	Single/Team	Abbreviation	n = predictions
Original Judges	(7-point scale)	Single Judge	OJs	86
New Judge 1	(3-point scale)	Single Judge	SNJ1	86
New Judge 2	(3-point scale)	Single Judge	SNJ2	86
New Judge 3	(3-point scale)	Single Judge	SNJ3	86
New Judge 4	(7-point scale)	Single Judge	SNJ4	86
New Judge 5	(7-point scale)	Single Judge	SNJ5	86
New Judge 6	(7-point scale)	Single Judge	SNJ6	86
Team Judges 1	(3-point scale)	Team judges	TTJ1	86
Team Judges 2	(7-point scale)	Team judges	TTJ2	86

Judging schedule

The ten judges were given five months to judge 220 transcripts. They received one of five randomized batches of up to 20 completed trial data. Batches included digital folders of photocopies of viewers' transcripts (cleaned of viewers' real names), along with judging sets and ranking sheets. Judges returned their score/prediction sheets to researchers via email before receiving the next batch. Judges were encouraged to spread their tasks evenly throughout the month to, as closely as possible, emulate the original judges' behaviours, who had typically carried out one to three trials per week. Because they worked unsupervised from their homes, however, researchers had no way to control how many trials they judged in a single sitting.

Overall project timeline and budgeting

The project met all prescribed deadlines. All original material (i.e., sessions, photos, etc.) was received from original project managers by May 2017. New judges were recruited by June 2017. Judges received judging packets by July 2017. Judges completed judging by December 2017. The project received a $3,000 PEAR award in early 2018. Each judge was paid

$300, and $100 went toward materials and printing costs. Analysis was completed by early June 2018. Results were reported in June at the combined 2018 IRVA/SSE conference and also in August at the 2018 Parapsychological Association conference.

Analysis methods

All comparative statistical analyses of proportions of correct hits were carried out using a test of binomial proportions using the method of approximation via normal distribution available online at http://vassarstats.net/binomialX.html. Comparisons between independent proportions (e.g., Hypothesis two) were carried out following the method suggested by Newcombe and Altman (2000).

Results

Results from preregistered hypotheses

Hypothesis one. Differences on prediction level. As predicted, we observed large hit-rate differences among 1) Original Judges, 2) new predictions of six Single New Judges (SNJ), and 3) new predictions of Teams of Two Judges (TTJ). See Table 3.

As shown in Table 3, the mean of hits percentage of the six Single New Judges is 51.5%, range 45%–59%, vs. the 65% of the original studies. This difference is statistically significant (independent proportions comparison) with p = .055 (one-tailed).

Further, the difference between the mean of hits percentages obtained by the two teams of judges (51.5%) and the original ones (65%) is evident. The statistical difference is equal to p = .087 (one-tailed).

Table 3.

Comparison of Original to New Judges' hit-miss-pass rates for Single/Team, 7-pt vs. 3-pt, level of experience

Ranking	Judge ID	Single/ Team	Ranking system	Hit rate	Experienced judge?	n = Passes	n = Hits	n = Misses
0	Original	Single	7-pt	65%	Yes	38	31	17
1	SNJ4	Single	7-pt	59%	Yes	23	37	26
2	SNJ1	Single	3-pt	55%	Yes	42	24	20
3	SNJ6	Single	7-pt	54%	Yes	51	19	16
4	TTJ1 Friends	Team	3-pt	54%	Yes	51	19	16
5	TTJ2 Married	Team	7-pt	49%	No	27	29	30
6	SNJ3	Single	3-pt	49%	No	27	29	30
7	SNJ2	Single	3-pt	47%	No	27	28	31
8	SNJ5	Single	7-pt	45%	No	19	30	37

Hypothesis two. Judging. As predicted, Original Judges' (OJ) hit rates were higher than Single New Judges' (SNJ) using the same method even if this difference was statistically different only with respect to the performance of the SNJ2. OJ hit rate: 64.58%: three SNJs: 58.73% (*p* = .53); 44.78% (*p* = .01) and 54.29% (*p* = .24) hit rates.

Hypothesis three. Ranking scale performance. Our hypothesis that the SRI seven-point CR scale would be more effective than the UC three-point scale was not confirmed. No appreciable difference in performance resulted from using the SRI seven-point CR scale vs. the UC three-point scale: (see Table 3).

- Three-point scale judges' results: 100 hits and 97 misses for an average hit rate of 50.75%
- Seven-point scale judges' results: 115 hits and 109 misses for an average hit rate of 51.34%

Hypothesis four. Consensus team judging. Our hypothesis that consensus team judging would be more effective, resulting in more hits than individual judging, was not confirmed. Consensus judging teams did not prove to be consistently more effective than single judges. The team of two new judges who used a three-point scale had a hit rate of 54.29%. This is compared with the single new judges who used the same scale and had a Mean: 50.38%; DS = 3.7%. The team that used a seven-point scale had a hit rate of 49.15%. This is compared with the entire group of single new judges who had a Mean: 52.6%; DS = 7.1%.

Hypothesis five. Individual performance (judges). Single New Judges. As predicted, some Single New Judges (SNJ) did better at judging than other SNJs, resulting in higher hit rates. Results indicated a particular judge obtained more ARV hits and/or fewer ARV misses than all other judges. SNJ4 produced the highest number (37) of hits among all new SNJ while maintaining an average miss rate. This resulted in the highest hit rate (58.73%) among all new judges, including the two TTJ. Only the OJ had a better hit rate (64.58%).

Hypothesis six. Multiple viewers. Predictions based on contributions from multiple viewers (consensus viewing) yielded better results than predictions made by single remote viewers, as hypothesized. Group predictions (contributions from multiple viewers): 69.56% hit rate (16 hits, 7 misses); Solo predictions (single remote viewer): 60.0% hit rate (15 hits, 10 misses). However, this difference is not statistically significant (*p* = .69).

Results from post-registered hypotheses, but prior to analysis

Hypothesis seven. Individual performance (viewers). As hypothesized, some remote viewers outperformed others even when their sessions were part of a group of viewers, while some underperformed. Researchers assessed how often a single viewer outperformed others so if only that viewer's session had been used, the prediction would have resulted in a hit. In 47 group predictions having multiple viewers, 30 events resulted in misses and passes. Of these, one viewer outperformed others in 19 events (63.3%). If only that viewer's session had been used, the prediction would have resulted in a hit.

Conversely, researchers analyzed this series involving multiple viewers to see how often a single viewer derailed the prediction so if their session had been eliminated, it would have resulted in a different prediction. Of the 30 group predictions that produced passes and misses (of 47 group predictions), a single viewer derailed 14 of the predictions. If their session been eliminated, a different prediction would have been made, resulting in a pass rather than a miss or a hit instead of a pass.

However, it is doubtful these differences can be applied to reduce passes and misses because most of the remote viewers had individual hit rates of 60% to 70%, and a judge couldn't predict which viewer would outperform or under-perform in each particular group trial.

Hypothesis eight. Spread between scores. Some evidence supported the hypothesis that a spread of at least two points between scores (using the seven-point scale) would yield better results. When individual viewers' scores were analyzed across all trials (those with both single and multiple viewers), 181 hits and 138 misses occurred when a two-point spread was observed, resulting in a percentage of 56.7%. This is 3% higher hit rate than when the one-point spread was applied (53.3%), However, this was not statistically significant (Table 4).

Table 4.

Comparison of number of predictions leading to hits and misses for each judge based on when they honoured a >1 or a >2 spread between CR scores when issuing predictions for SRI 7-pt scale

Sides Diff (Solo + Group)	CRDiff. OJ	CRDiff SNJ4	CRDiff. SNJ5	CRDiff. SNJ6	CRDiff. TTJ2	Totals
>2 - Hits	36	46	42	28	29	181
>2 - Misses	23	33	35	24	23	138
>1 - Hits	76	82	70	57	53	338
>1 - Misses	49	65	83	47	52	296
Sides Diff (Solo)	CRDiff. OJ	CRDiff SNJ4	CRDiff. SNJ5	CRDiff. SNJ6	CRDiff. TTJ2	
>2 - Hits	10	9	8	5	7	39
>2 - Misses	6	11	12	7	5	40
>1 - Hits	18	15	14	12	11	71
>1 - Misses	13	16	20	13	11	73
Sides Diff (Group)	CRDiff. OJ	CRDiff SNJ4	CRDiff. SNJ5	CRDiff. SNJ6	CRDiff. TTJ2	
>2 - Hits	26	37	34	23	22	142
>2 - Misses	17	22	23	17	18	97
>1 - Hits	58	67	56	45	42	268
>1 - Misses	36	49	63	34	41	223

Hypothesis nine. Minimum prediction threshold. When at least one of the photos earns a minimum confidence ranking of 3.5 or higher on the seven-point scale, it should yield more hits than vice versa. Results did not support this hypothesis. Trials with CR scores of 3.5 or higher had virtually the same number of hits and misses (322 hits and 323 misses).

Hypothesis ten. Passes. We predicted more passes would produce a higher hit rate (more successful predictions). Consequently, we expected a positive correlation between passes and hits. The correlation between passes and misses was positive, but not statistically significant: *rho* = .42; *p* = .13 (one-tailed). Differently, the correlation between passes and hits turned out negative and statistically significant: *rho* = –.78; *p* = .013 (two-tailed).

Results from post hoc analysis

Judging Experience. As shown in Table 3, judges with more experience obtained higher hit rates: Mean = 57.4; SD = 4.7; Range: 54–65. Conversely, judges with less experience obtained the following results: Mean = 47.5; SD = 1.9; Range 45–49. This difference was also statistically significant with a *p* = .009 at the Mann-Whitney U test.

Discussion

This exploratory study examined various factors and practices leading to successful Associative Remote Viewing (ARV) predictions. Often the success or failure of RV projects is attributed to the skill of the viewers, with misses attributed to a lack of psi. However, if psi ability was all that mattered, our study with new judges using the same data should have had similar results as the trials with the original judges. This was not the case.

We came away with four key findings that may be useful to other project managers. First, consistency between raters' scores and predictions was remarkably low, demonstrating ARV outcomes may be highly dependent on the decision-making and performance of raters. This low correspondence between the original and new judges occurred despite pretrial training to establish inter-rater reliability and with only two photos in each set.

Our second key finding was that passing, a feature unique to applied ARV projects, may not improve hit rates, as commonly believed by ARV practitioners. Most experimental parapsychology projects do not issue passes because their goal is solely to determine whether there is a psi effect. In an applied remote viewing project to determine on which option to wager, where success is determined not only by the hit/miss ratio but monies earned through wagering, the common logic is that a pass can protect against loss of income. The nonparametric Spearman rho correlations between passes and hits show a clear, strong inverse relationship. In other words, the more passes, the fewer hits.

This finding initially perplexed us. From simply eyeballing the pass to hit/miss ratio, we could clearly see the original judges passed more than the new judges (38 vs. 34), and the original judges' hit rate of 64.58% was higher than all new judges' hit rates. The SNJ with the fewest passes (19) had the lowest hit rate (44.8%). A closer examination of Table 3 revealed those were merely distractions from the bigger picture. The two judges, including a single judge and pair of judges, with the highest number of passes (51) had the lowest number of hits (19) and of misses (16). The new judge with the highest hit rate of all SNJ (58.73%) had the second lowest number of passes (23). In other words, this SNJ had 11 fewer passes and 10 more hits than average.

Something else was going on beyond merely passing. Did following the "two-point spread" criteria make a difference? Judges who mostly followed the criteria of passing if the spread between the CR scores for the photos was two points or greater did have a higher hit rate overall than when the one-point spread rule was applied (181 hits and 138 misses). But the 3% difference in the hit rates between the two- and one-point spreads (56.7% and 53.3% respectively) was not statistically significant.

Within the Applied Precognition Project community, it is believed, based on anecdotal findings, that scores of 3.5 or better on a seven-point scale show a greater presence of psi and are "wagerable". Our third key finding was that the overall data did not support *that* hypothesis but a confidence ranking of five or greater was an indicator of success for trials with multiple viewers. Taken as a whole, trials by solo and multiple viewers with CR scores of 3.5 or higher showed an equivalent number of hits and misses (322 hits and 323 misses). But the data showed about 20% more hits than misses occurred in trials with multiple viewers when judges issued a CR score of five or higher using the seven-point scale (Table 5). We must issue a cautionary note here, since solo viewers with the same CR 5+ scores had an equal rate of hits and misses (26 and 26), not a hit-rate increase.

One factor not explored in this study or in the aforementioned ARV dream study was how judges dealt with discrepancies between transcripts from multiple viewers within a group. Several times in the present study, original judges and new judges gave equally high CR scores to transcripts for both targets within the same trial (e.g., CR 5s for both Side A and Side B). Our guidelines called for a pass in such cases, but raters did not always follow the guidelines. Another unexplored factor was how judges arrived at an aggregate CR score for a group of multiple viewers to make a prediction. Parameters on how much weight to give to high- and low-ranking transcripts should be established and evaluated. These very important considerations were not part of the present study nor were they discussed in the literature we reviewed.

Our fourth key finding, discovered during post hoc analysis, was that judges with more experience obtained statistically significant higher hit rates than less experienced judges. While caution should be taken in making inferences given the small sample size, this still warrants some consideration. Our most experienced judges were the original judges. What gave them an advantage? It could have been they were simply more experienced than the new judges. However, other factors may have been involved, as well. For example, sometimes the original judges created the photo pairings based on an understanding of the viewers' interests and strengths. Viewers emailed their transcripts directly to the original judges, and the judges had the ability to track both viewer progress and their own, which may have influenced decision-making about subsequent predictions.

In contrast, the new judges remained blind to the viewers' identity and to their own progress until all trials were complete. Without any kind of feedback loop, the new judges didn't know if they were making good choices. Therefore, they were unable to learn or adjust their decision-making or behaviours accordingly. While working with Osis and Mitchell (1981) at the ASPR, Swann (1996) found having a feedback loop to be important for remote viewing performance, and perhaps it could be true for judging performance, as well.

Table 5.

Comparison of number of predictions leading to hits and misses for each judge based on their assigned CR Scores of >3.5 and >5

Absolute Score (Solo + Group)	OJ	SNJ4	SNJ5	SNJ6	TTJ2	Totals
>=5 Hits	24	20	13	10	33	100
>=5 Misses	17	14	18	3	32	84
>5 Hits	13	6	5	6	19	49
>5 Misses	10	2	7	2	17	38
>=3.5 Hits	69	86	67	35	65	322
>=3.5 Misses	62	67	89	33	72	323
Absolute Score (Solo)						
>=5 Hits	4	6	6	4	6	26
>=5 Misses	3	7	7	1	8	26
>5 Hits	2	1	3	4	3	13
>5 Misses	3	1	4	0	5	13
>=3.5 Hits	15	14	13	8	14	64
>=3.5 Misses	13	16	22	7	11	69
Absolute Score (Group)						
>=5 Hits	20	14	7	6	27	74
>=5 Misses	14	7	11	2	24	58
>5 Hits	11	5	2	2	16	36
>5 Misses	7	1	3	2	12	25
>=3.5 Hits	54	72	54	27	51	258
>=3.5 Misses	49	51	67	26	61	254

NOTE: This takes all scores > = 3.5 regardless if the Judge picked a Pass. For example: Judge might have scored Side A = 4.5 and Side B = 4.0 and his pick could be a Pass (because the scores are too close) — but regardless of that, this example is counted as 'Side A' pick and counted as a 3.5 Hit

We suspect some new judges may also have experienced judging fatigue. Original judges often judged one to two trials per week, whereas some of the new judges did several sessions in one sitting, despite researchers' attempts to restrict this type of behaviour. This may have led to rushing and paying less attention to detail. Viewer fatigue has been a concern since Harary and Targ's second (and failed) attempt in predicting the silver futures market in the mid-1980s (Larson, 1984). Similarly, the stress on APP judges in Project Firefly to make 240 foreign exchange trades in less than a year provided another example — affecting judges rather than viewers — of too many predictions in too short a timespan (Katz et al., 2018). While we can't know the actual effect of "binge-judging/predicting" in the present study, it again raises the question and points to judging fatigue as a factor to consider when designing protocols.

We expected to find differences in successful prediction hit rates between those who used a seven-point vs. three-point ranking scale. We did not. The sample size for both groups was quite small and therefore further testing is needed.

Earlier studies had suggested consensus team judging would be more effective than single judges. Our two new teams of two judges (TTJs) performed comparably to the single new judges (SNJs), not better. However, much remains to be explored there. One of the project's pair of judges had less judging experience than all the others. Also, one of the judges on the other team admitted she sometimes felt pressured to give in to the other judge to avoid conflict, which defeated the purpose of consensus judging.

Our study obviously attempted to test many variables at once — two-point or one-point spread between CRs, group or solo transcripts, original or new judges, single or team judges. While all may be relevant factors, a simpler study solely on the effects of passing could provide a more definitive answer. This could be accomplished by reusing data from already completed trials where judges issued passes when appropriate and keeping all parameters stable except for requiring predictions on all trials (no passes).

To summarize, none of the various factors we studied was, in itself, a clear indicator for success or failure. Together, they point to the difficulty of achieving intra-rater reliability, which Stemler (2004, p. 9) called "one of the most important concepts of educational and psychological measurement". However, the results do clearly demonstrate that some judges getting more hits than others wasn't due to the level of psi exhibited by the viewers — that remained constant — but rather to the judges' ability and decision-making processes.

Finally, while we assume judges' decisions, with the aid of training and scoring systems, were logic-based, we also can't rule out the possibility of their own psi. Judges were instructed to not intentionally use their intuition, but much of the presentiment literature posits that psi functions at an unconscious level (Radin, 2004). The Experimenter Psi Hypothesis is that psi-conducive experimenters (or PIs) influence the outcome of their experiments by imposing their own psi. This "experimenter effect" has been noted by several researchers, including White (1977) and Palmer (1997). While discussing the experimenter effect in his presidential address to the Parapsychological Association, Honorton (1976, p. 220) said, "I think it is interesting to note that many of the successful experimenters in psi research have also been successful subjects". In the present study, all but one of our judges, both original and new, were remote viewers or clairvoyant practitioners. All but one of the highest performing judges were at least self-described intermediate to advanced remote viewers, with the lowest performing judges referring to themselves as "beginner" remote viewers.

Conclusion

Unlike most projects of this kind, our study made use of data from already-completed ARV trials, with the outcomes of predictions already known. Rather than having to allocate time and resources to gathering data from participants, this design allowed for resources to be focused almost exclusively on the analysis phase. While not identical replications, the rejudged trials closely followed the original judging procedures, allowing a number of variables to be tested. This approach could prove useful for future experiments, particularly ones that seek to test whether effects were due exclusively to participant performance or to design-related considerations.

The present researchers had difficulty finding project managers willing to provide data from completed ARV trials. We contacted those who had published past studies, as well as two researchers who had completed lengthy series of trials but never formally published their results. Several researchers said they no longer had access to the requested data. Others said their data was not kept in an organized manner that would allow for easy access and sharing. Some did not respond or they agreed to participate but never followed through.

Given the extensive time and resources that go into any project, we hope this study helps motivate researchers to save their data in a manner that would make it accessible for ongoing assessment. This should not be in conflict with most institutional Internal Review Board requirements, providing measures are taken by original researchers to exclude participants' names from all shared data. Use of an open-source scientific platform might help toward this aim.

Acknowledgments

This project received the 2018 PEAR award from the Parapsychological Association.

Mapleton, OR 97453 Debra Katz
debra@debrakatz.com

REFERENCES

Atwater, F. H. (2001). *Captain of my ship, master of my Soul: Living with guidance*. Hampton Roads Publishing.

Bierman, D. (2014. August 8–11). Can psi sponsor itself? Simulations and results of an automated ARV-Casino experiment. [Conference Session]. Proceedings of the 56th Parapsychological Association Convention. Viterbo, Italy. *Journal of Parapsychology*, 77(2), 159–164. https://www.rhine.org/images/jp/JPv77n2.pdf.

Child, I. L. (1985). Psychology and anomalous observations: The question of ESP in reams. *American Psychologist*, *40*, 1219–1230.

Fendley, T. W. (2016). WWC Group sets the pace in Associate Remote Viewing. *Eight Martinis Magazine*, 14.

Grgić, I. (2019). RV Studio software. http://www.arv-studio.com

Harary, K., & Targ, R. (1985). A new approach to forecasting commodity futures. *Psi Research*, *4*, 79–85.

Herlosky, J. (2015). *A sorcerer's apprentice: A skeptic's journey into the CIA's project Stargate and remote viewing*. Trine Day Publishing.

Honorton, C. (1975). Objective determination of information rate in psi tasks with pictorial stimuli. *Journal of the American Society for Psychical Research*, *69*, 353–359.

Honorton, C. (1976). Has science developed the competence to confront the claims of the paranormal? Presidential Address, Parapsychological Association. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.) 1975., *Research in Parapsychology*, 199–223. Scarecrow.

Humphrey, B. S., May, E. C., & Utts, J. M. (1988). Fuzzy set technology in the analysis of remote viewing. *Proceedings of the 31st Annual Convention of the Parapsychological Association*, 378–394.

Humphrey, B. S., Task, V. V., May, E. C., & Thomson, M. J. (1986). Remote viewing evaluation techniques, *SRI International, Menlo Park, CA, Final Report*, Project 1291, SRI/GF-029.

Jahn, R. G., Dunne, B. J., & Jaun, E. G. (1980). Analytical judging procedure for remote perception experiments. *Journal of Parapsychology*, *44*, 207–231.

Katz, D., Bulgatz, M., & Fendley, T. (2015). Remote Viewing the outcome of the 2012 Presidential election: An expedition into the unexplored territory of remote viewing and rating human subjects as targets within a binary protocol. *Aperture*, Spring/Summer, 46–56.

Katz, D. L., Beem, L., & Fendley, T. W. (2015). Explorations into remote viewing microscopic organisms. *Aperture*, *26*, Fall/Winter, 42–49.

Katz, D. L., Grgić, I., & Fendley, T. W. (2018). An ethnographical assessment of ProjectFirefly: A year long endeavor to create wealth by predicting FOREX currency moves with associative remote viewing. *Journal of Scientific Exploration, 32*(1), 21–54.

Katz, D. L., Smith, N., Graff, D., Bulgatz, M., & Lane, J. (2019). The associative remote dreaming experiment: A novel approach to predicting future outcomes of sporting events. *Journal of the Society for Psychical Research, 83*(2), 65–84.

Kindon S., Pain, R., & Kesby, M. (2007). *Participatory action research approaches and methods: Connecting people, participation and place.* Routledge.

Knowles, J. (2017). *Remote Viewing from the ground up*. Create Space for MPRV Publishing.

Kolodziejzyk, G. (2015). 13-year associative remote viewing experimental results. *The Journal of Parapsychology, 76*(2), 349–368.

Larson, E. (1984). Did psychic powers give firm a killing in the silver market? — And did greed ruin it all? Californians switch over to an extrasensory switch. *Wall Street Journal.* (Eastern edition). New York, N.Y.: Oct 22.

May, E. C. & Marhawa, 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.

May, E., Utts, J., Humphrey, B., Luke, W., Frivold, T., & Trask, V. (1990). Advances in remote viewing analysis. *Journal of Parapsychology*, *54*, 193–294.

McMoneagle, J. & May, E. C. (2016, June 13–16). Recommendations for ARV projects. *Applied Precognition Project Annual Conference*, Las Vegas, NV.

Milton, J. (1997). Meta-analysis of free-response ESP studies without altered states of consciousness. *Journal of Parapsychology, 61*(4), 79 plus.

Milton, J. (1985). *A survey of free response judging practices.* CIA documents.

Morehouse, D. (2011). *Remote viewing: The complete user's manual for coordinate remote viewing. S*ounds True Publishing.

Müller, M., Mülller, L., & Wittmann, M. (2019). Predicting the stock market: An associative remote viewing study. *Zeitschrift für Anomalistik*. Band 19, 326–346.

Nelson, R. D. (2017). Princeton engineering anomalies research (PEAR)'. *Psi Encyclopedia. The Society for Psychical Research.*

Newcombe, R. G. & Altman, D. G. (2000). Proportions and their differences. *Statistics with confidence*, *2*, 45–56.

Nobel, J (2018). *Natural remote viewing: A practical guide to the mental martial art of self-discovery*. Amazon Digital Services, LLC.

Palmer, J. (1997). The challenge of experimenter psi. *European Journal of Parapsychology, 11,* 10–125.

Poquiz, A. (2012). Alexis Poquiz (Dung Beetle) scoring system. Private email correspondence with Debra Katz.

Poquiz, A. (2021). Alexis Poquiz (Dung Beetle) scoring system in Katz, D. & Knowles, J. Associative Remote Viewing (Eds.), *The art and science of predicting outcomes for sports, financials, elections and the lottery*. Chapter 11. *Living Dreams Press*.

Puthoff, H. E. (1984). *ARV (Associative Remote Viewing) applications: Research in parapsychology*, 121–122. ScarecrowPress, Inc.

Puthoff, H. E., & Targ, R. (1976). A perceptual channel for information over kilometer distances: Historical perspective and recent research. *Proceedings of the IEEE, 64*(10). doi: 10.1109/PROC.1976.10370.

Rosenblatt, M. (2019) Private email correspondence dated February 1, 2020, to researcher (Katz), providing membership information, for Applied Precognition Project.

Rosenblatt, R., Knowles, J. & Poquiz, A. (2015). Applied Precognition Project (APP) and a summary of APP-2014. *Connections Through Time*, 38.

Schwartz, S. A. (2019). The location and reconstruction of a Byzantine structure in Marea, Egypt, including a comparison of electronic remote sensing and remote viewing. *Journal of Scientific Exploration, 33*(3), 451–480.

Schwartz, S. A. (2020). The origins of ARV. *Mindfield*: *Bulletin of the Parapsychological Association, 12*(1), 5–15.

Smith, C., Laham, D., & Moddel, G. (2014). Stock market prediction using associative remote viewing by inexperienced viewers. *Journal of Scientific Exploration*, *2*8(1), 7–16.

Smith, D. (2014). CRV — *Controlled Remote Viewing: Manuals, collected papers & information to help you learn Controlled Remote Viewing*. Amazon Digital Services, LLC.

Smith, P. H. (2005). *Reading the enemy's mind: inside star gate — America's psychic espionage program*. Tom Dougherty Associates LLC.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating inter-rater reliability. *Journal of Practical Assessment, Research & Evaluation, 9*(4), 1–11.

Storm, L., Sherwood, S. J., Roe, C. A., Tressoldi, P. E., Rock, A. J., & Di Risio, L. (2017). On the correspondence between dream content and target material under laboratory conditions: A meta-analysis of dream-ESP studies, 1966–2016. *International Journal of Dream Research,10(*2), 120–140.

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.

Storm, L. (2019). Imagination and reactance in a psi task using the imagery cultivation model and a fuzzy set encoded target pool. *Journal of Scientific Exploration, 33*(2), 197–212.

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), 471–485.

Swann, I. (1993). On remote viewing, UFO's and extraterrestrials. *Fate Magazine*. September,73–82.

Swann (1996) The real story. The American prophecy project. https://ingoswann.com/biomind-1.

Targ, R. (2012). *The reality of ESP: A physicist's proof of psychic abilities*. Quest Books.

Targ, R., Katra, J., Brown, D., & Wiegand, W. (1995). Viewing the future: A pilot study with an error-detecting protocol. *Journal of Scientific Exploration*, *9*, 367–380.

Targ, R., Puthoff, H. E., & May, E. C. (1977). State of the art in remote viewing studies at SRI. *Proceedings of the International Conference of Cybernetics*.

Vivanco, J. (2016). *The time before the secret words: On the path of remote viewing, high strangeness and zen*. Amazon Digital Services LLC.

White, R. A. (1977). The influence of experimenter motivation, attitudes, and methods of handling subjects on psi test results. In B. B. Wolman (Ed.), *Handbook of parapsychology*, 273–301. Van Nostrand Reinhold.

Williams, L. L. (2019). *Boundless: Your how to guide to practical remote viewing — phase one (A How-To series to learn controlled remote viewing book 1).* Amazon digital services.

Williams, L. L. (2017). Monitoring: A guide for remote viewing & professional intuitive teams. Amazon Digital Services.

Appendix A

University of Colorado Three-Point Confidence Ranking (CR) Scale

A-1 Low similarity to the “Side A” target and with little similarity to the “Side B” target.

A-2 Good (medium) similarity to the “Side A” target and with usually low or no similarity to the “Side B” target.

A-3 Excellent (high) similarity to the “Side A” target, with usually low or no similarity to the “Side B” Target.

B-1 Low similarity to the “Side B” target, with little similarity to the “Side A” target.

B-2 Good (medium) similarity to the “Side B” target and with usually low or no similarity to the “Side A” target.

B-3 Excellent (high) similarity to the “Side B” target, with usually low or no similarity to the “Side A” Target.

M-0 Means it can’t be judged between the two and the transcript has low or no similarity to both targets.

M-1 Means it can’t be judged between the two and the transcript has medium similarity to both targets.

M-2 Means it can’t be judged between the two and the transcript has high similarity to both targets.

Appendix B

SRI Seven-Point Confidence Ranking (CR) Scale

7 Excellent correspondence, including good analytical detail (e.g., naming the target), and with essentially no incorrect information.

6 Good correspondence with good analytical information (e.g., naming the function of the target), and relatively little incorrect information.

5 Good correspondence with unambiguous unique matchable elements, but some incorrect information.

4 Good correspondence with several matchable elements intermixed with some incorrect information.

3 Mixture of correct and incorrect elements, but enough of the former to indicate that the viewer has made contact with the target.

2 Some correct elements, but not sufficient to suggest results beyond chance expectation.

1 Little correspondence.

0 No correspondence.