Journal of Anomalous Experience and Cognition

JAEX VOLUME 4 ISSUE 1 2024

ISSN: 2004-1969

Journal of Anomalous Experience and Cognition

Volume 4, Issue 1, 2024

ISSN: 2004-1969

Journal of Anomalous Experience and Cognition

(JAEX)

The *Journal of Anomalous Experience and Cognition (JAEX)* provides a forum for the rigorous, multidisciplinary study of anomalous experience and cognition. Anomalous experience refers to spontaneous or induced unusual but not necessarily pathological experiences, such as mystical and out-of-body experiences. Anomalous cognition refers to rigorous multidisciplinary research that seeks to improve our understanding of psycho-physical relations. It includes the hypotheses that organisms can be affected by spatially or temporally distant stimuli -unmediated by the senses or reason- and that intentions can directly affect physical systems, as well as related attitudes, beliefs, and other variables.

There are no charges for article submission or processing. JAEX uses the Open Journals System at Lund University (OJS), supported by Lund University's Library. JAEX's publisher is the Parapsychology Foundation (PF).

Online and printed issues have ISSNs, and all published items have a Crossref DOI (digital object identifier), are archived in PKP PN, and are indexed in the Directory of Open Access Journals (DOAJ), Google Scholar, and Researchgate from the journal's inception. There are ongoing applications to other indexes.

Layout Design
Nikolaos Koumartzis, SEIKILO Cultural Organization

Publishers:
CERCAP/PF, Etzel Cardeña, Ph. D., Department of Psychology, Lund University, Allhelgona kyrkogata 16a, Lund, 22100, Sweden.

Parapsychology Foundation, 315 East 72 Street, #10L, New York, NY, USA

JAEX is partly supported by a generous annual donation from the IHPV Foundation.

Permanent article identifier: DOI

Archives digitally identified in PKP PN

For additional information:
https:// journals.lub.lu.se/jaex

Table of Contents

EDITORIAL

Celebrating Three Consciousness Trailblazers: Jeanne Achterberg, Ruth-Inge Heinze, and Stanley Krippner

Etzel Cardeña

Lund University

Abstract: Following the previous homage to Charles T. Tart and considering that Stan Krippner has celebrated more than 90 birthdays, this issue celebrates Stan along with two consciousness trailblazers who collaborated with him while they were alive: Jeanne Achterberg and Ruth-Inge Heinze. Jeanne was a very influential pioneer in mind/body medicine, particularly imagery in healing. Ruth-Inge Heinze had a deep commitment to the study of shamanism and promoted its interdisciplinary study in many ways. Finally, Stan Krippner is the quintessential liminal figure with major contributions to the study of dreams, anomalous cognition, personal mythology, and other consciousness areas.

Keywords: Consciousness, altered states of consciousness, shamanism, mind-body medicine, mental imagery, anomalous cognition, parapsychology, Jeanne Achterberg, Ruth-Inge Heinze, Stanley Krippner

The previous *JAEX* issue celebrated the life and work of altered states of consciousness researcher and theoretician Charles T. Tart. I now turn to three other pioneers: Jeanne Achterberg, Ruth-Inge Heinze, and Stanley Krippner. The first two passed away some years ago, but when I decided to honor Stan in his 90+ years, I thought of how much women have been undervalued in all fields, including the study of consciousness. I decided to balance *JAEX's* celebrations of Charley and Stan with a recognition of two (out of many) outstanding women who, incidentally, had some association with Stan.

https://doi.org/10.31156/jaex.26220

Jeanne Achterberg: Woman as Healer/Researcher/Author

I start with Jeanne Achterberg, Ph. D. (1949-2012). She added her research skills and multidisciplinary knowledge to those of other women who have soothed ailing bodies and souls as shamans, nurses, or other type of healers (Achterberg, 1990). Her focus was on mind-body medicine -which studies and uses the effects of psychological variables on medical conditions- particularly guided imagery, which she researched for decades. *Time* magazine named Jeanne one of six innovators of alternative (complementary might be more appropriate) medicine for the 21st century (Lemonick, 2009). Recently, cognitive-behavioral therapy approaches seem to have "discovered" the importance of imagery but without crediting her contributions or those of the centuries-old study of hypnosis.

Jeanne was faculty at Southwestern Medical School, Saybrook University, and the Institute of Transpersonal Psychology (ITP), co-chaired a panel on mind/body interventions for the Office of Alternative Medicine of the NIMH (National Institutes of Mental Health), advised the Office of Technology Assessment of the US Congress, and served as senior editor for the *Journal of Alternative and Complementary Medicine.* She is also known for her collaboration with the Simontons on psychological effects on cancer outcome (e.g., Achterberg et al., 1977), and books on imagery in healing and women as healers (Achterberg, 1985, 1990). She also co-authored an overview of anomalous healing experiences (Krippner & Achterberg, 2014) and a creative study on distant healing intention and the brain function of recipients (Achterberg et al., 2005).

Our paths crossed a few times at the NIMH, Esalen, and the then called Institute of Transpersonal Psychology. (ITP; now Sofia University). She was a daring but careful and demanding scholar and did not tolerate sloppiness and arrogance masquerading as spirituality. I recall a conversation we had at ITP deploring how a student was boasting that s/he would complete her dissertation without using a single reference. Her luminous presence can be gleaned from an interview available online (Mishlove, no date).

Ruth-Inge Heinze: Playful Tamer of Shamans

Ruth-Inge Heinze and Stanley Krippner

Another vibrant woman who did not suffer fools gladly was Ruth-Inge Heinze (1919-2007), an indefatigable actor and anthropologist who organized yearly meetings on shamanic scholarship, interspersed by rituals conducted by experts from all over the world. She had had a successful acting career in Berlin collaborating with the likes of Lote Lenya and Peter Lorre before having to emigrate to the US, where she got a Ph. D. in anthropology and became faculty at the University of California, Berkeley, the California Institute of Integral Studies (CIIS), and Saybrook Graduate School (Anonymous, 2007). She was an expert in Southeast Asia shamanism and published, among others, about 15 (!) proceedings of her organized International Conferences on Shamanism and Alternate Modes of Healing, *Shamans of the 20th Century* (Heinze, 1990), and *The Nature and Function of Rituals* (Heinze, 2000). I probably met Ruth-Inge through Stan and we got along fabulously because we shared careers as actors; she published a couple of my papers discussing ritualistic and shamanic practices in 20th century performance (Cardeña, 1987, 1991). Ruth-Inge could be fierce with those she disliked, but I was fortunate to receive instead her child-like playfulness and open heart. With very different personal styles, Jeanne and Ruth-Inge unveiled the importance of imagery and visions for individual and communal healing.

Stanley Krippner: The Liminal Psychologist

And now to my decades-long friend and colleague Stanley Krippner. Stan is the most liminal person I know, straddling along noetic, cultural, disciplinary, and personal frontiers. I met him while I was doing my Ph. D. at the University of California, Davis, in the late 80s and our paths crossed many more times in conferences on shamanism,

dreaming, and parapsychology. I could see how skillful he was as a presenter, neither rushing nor dragging his presentations, and ending always precisely in time.

He was always generous with his time even when I was a mere grad student and he already a recognized figure. We have been friends since, although in the bargain he got, for better or worse, a friend who can also be a critic and does not think that he is a holy being, as some seem to do...

Stan has contributed to many areas in consciousness and we have co-authored various publications, foremost *Varieties of Anomalous Experience* (Cardeña, Lynn, & Krippner, 2000, 2014), which made history by being published by the American Psychological Association in 2000.

Stan's legacy includes supporting international communities and authors interested in the study of alterations of consciousness, anomalous cognition, and shamanism. The writeup for the American Psychological Association's 2002 Award for Distinguished Contributions to the International Advancement of Psychology credited his: "...efforts to expand the frontiers of the psychological study of consciousness... service interpreting indigenous traditions of world cultures for Western audiences (Anonymous, 2002)." He published an important overview of shamanism in response to this honor (Krippner, 2002).

From his work on dreaming and anomalous cognition, I would single out the ground-breaking studies on dream telepathy at the Maimonides Center (Ullman et al., 1973) and his 10 volumes or so of *Advances in Parapsychology Research* (with various co-editors). Stan, thanks to his bohemian side and acquaintance with many artists and musicians such as The Grateful Dead, also crossed the borders between art, science, and creativity, as in his studies of use of psychedelics and creativity (1985), and a telepathy test in a concert of the Grateful Dead (Krippner et al., 1973). His mind remains generous and creative, despite the many years weighing on his body.

All students of consciousness are in debt to these three singular and courageous consciousness explorers.

References

Achterberg, J. (1985). *Imagery in healing: Shamanism and modern medicine.* Shambhala.

Achterberg, J. (1990). *Woman as healer.* Shambhala.

Achterberg, J., Cooke, K., Richards, T., Standish, L. J. , Kozak, I., & Lake, J. (2005). Evidence for correlations between distant intentionality and brain function in recipients: a functional magnetic resonance imaging analysis. *Journal of Alternative and Complementary Medicine,*

11(6):965-671.

Achterberg, J., Lawlis, G. F., Simonton, O. C., & Matthews-Simonton, S., (1977). Psychological factors and blood chemistries as disease outcome predictors for cancer patients. *Multivariate Experimental Clinical Research*, *3*(3), 107-122.

Anonymous (2002). Stanley C. Krippner: Award for Distinguished Contributions to the International Advancement of Psychology (2002). *American Psychologist, 57*(11), 960–961. https://doi.org/10.1037/0003-066X.57.11.960

Anonymous (2007). Ruth-Inge Heinze. *San Francisco Chronicle*. https://www.legacy.com/us/obituaries/sfgate/name/ruth-inge-heinze-obituary?id=24620445&__cf_chl_tk=ptxyVf32us.OXB_CoLZlZNTkz4FCWmmpLA4hul_G79o-1714378721-0.0.1.1-1450

Cardeña, E. (1991). Jerzy Grotowski. A shaman director. In R. I. Heinze (Ed.), *Shamans of the XXth century* (pp. 86-92). Irvington.

Cardeña, E. (1987). The magical flight: Shamanism and theatre. In R. I. Heinze (Ed.), *Proceedings of the Third International Conference on the Study of Shamanism and Alternate Ways of Healing* (pp. 291-304). A-R Editions.

Cardeña , E., Lynn, S. J., & Krippner, S. (2000). *Varieties of anomalous experience: Examining the scientific evidence,* American Psychological Association.

Heinze, R-I. (1990). *Shamans of the 20th century.* Irvington.

Heinze, R-I. (2000). *The nature and function of ritual: Fire from heaven.* Praeger.

Krippner, S. (1985). Psychedelic drugs and creativity. *Journal of Psychoactive Drugs, 17*(4), 235–245. https://doi.org/10.1080/02791072.1985.10524328

Krippner, S. C. (2002). Conflicting perspectives on shamans and shamanism: Points and counterpoints. *American Psychologist, 57*(11), 962–977. https://doi.org/10.1037/0003-066X.57.11.962

Krippner, S., & Achterberg, J. (2014). Anomalous healing experiences. In E. Cardeña, S. J. Lynn, & S. Krippner (Eds.), *Varieties of anomalous experience: Examining the scientific evidence* (2nd ed., pp. 273–301). American Psychological Association. https://doi.org/10.1037/14258-010

Krippner, S., Honorton, C., & Ullman, M. (1973). An experiment in dream telepathy with "The Grateful Dead." *Journal of the American Society of Psychosomatic Dentistry and Medicine, 20*, 9 - 17.

Lemonick, M. D. (2001). Alternative medicine/guided imagery: Mind over malignancies. *Time.* https:// content.time.com/time/subscriber/article/0,33009,999711,00.html

Mishlove, J. (no date). *Imagery in healing. Part one.* https://www.youtube.com/watch?v=TMyi4QOr6qU

Ullman, M., & Krippner, S., with Vaughan, A. (1973). *Dream telepathy: Experiments in nocturnal extrasensory perception*. Macmillan.

Célébration de Trois Pionniers de la Conscience: Jeanne Achterberg, Ruth-Inge Heinze, et Stanley Krippner

Etzel Cardeña

Résumé: Après que nous ayons déjà rendu hommage à Charles T. Tart, et en prenant en considération que Stan Krippner a fêté ses plus de 90 ans, ce numéro rend hommage à Stan ainsi qu'à deux pionniers de la conscience qui ont collaboré avec lui de leur vivant: Jeanne Achterberg et Ruth-Inge Heinze. Jeanne a été une pionnière très influente dans le domaine de la médecine corps/esprit, en particulier en ce qui concerne l'imagerie mentale et la guérison. Ruth-Inge Heinze était profondément attachée à l'étude du chamanisme et a encouragé son étude interdisciplinaire de multiples façons. Enfin, Stan Krippner est la quintessence de la figure liminale, avec des contributions majeures à l'étude des rêves, de la cognition anomale, de la mythologie personnelle et d'autres domaines de la conscience.

Translation into French by Antoine Bioy, Ph. D.

Zur Feier von drei Wegbereitern der Bewusstseinsforschung: Jeanne Achterberg, Ruth-Inge Heinze, und Stanley Krippner

Etzel Cardeña

Zusammenfassung: Nach der vorangegangenen Hommage an Charles T. Tart und in Anbetracht der Tatsache, dass Stan Krippner mehr als 90 Geburtstage gefeiert hat, wird er in dieser Ausgabe zusammen mit zwei Wegbereiterinnen der Bewusstseinsforschung gefeiert, die zu Lebzeiten mit ihm zusammengearbeitet haben: Jeanne Achterberg und Ruth-Inge Heinze. Jeanne Achterberg war eine sehr einflussreiche Pionierin auf dem Gebiet der Geist-Körper-Medizin, insbesondere der Bilderwelt in der Heilung. Ruth-Inge Heinze engagierte sich sehr für das Studium des Schamanismus und förderte dessen interdisziplinäre Erforschung auf vielfältige Weise. Stan Krippner schließlich ist der Inbegriff des Grenzgängers, der wichtige Beiträge zur Erforschung von Träumen, anomaler Kognition, persönlicher Mythologie und anderen Bewusstseinsbereichen geleistet hat.

Translation into German by Eberhard Bauer, Ph. D.

Celebrando Três Pioneiros da Consciência: Jeanne Achterberg, Ruth-Inge Heinze, e Stanley Krippner

Etzel Cardeña

Resumo: Seguindo a homenagem anterior a Charles T. Tart e considerando que Stan Krippner já completou mais de 90 anos, o presente estudo celebra Stan junto de duas pioneiras da consciência que colaboraram com ele enquanto estiveram vivas: Jeanne Achterberg e Ruth-Inge Heinze. Jeanne foi uma pioneira muito influente na medicina mente/corpo, especialmente no uso de imagens voltadas para cura. Ruth-Inge Hein-

ze revelara um compromisso profundo com o estudo do xamanismo e promoveu seu estudo interdisciplinar de diversas maneiras. Por fim, Stan Krippner é a figura primordial por excelência a trazer grandes contribuições para o estudo dos sonhos, cognição anômala, mitologia pessoal e outras áreas da consciência.

Translation into Portuguese by Antônio Lima, Ph. D.

Una Celebración de Tres Pioneros de la Consciencia: Jeanne Achterberg, Ruth-Inge Heinze, y Stanley Krippner

Etzel Cardeña

Resumen: Tras el homenaje previo a Charles T. Tart y tomando en consideracòn que Stanley Krippner ha cumplido más de 90 años, este número celebra a Stan junto con dos pioneras de la consciencia que colaboraron con él: Jeanne Achterberg y Ruth-Inge Heinze. Jeanne fue muy influyente en el campo de la medicina mente/cuerpo, especialmente la sanación usando imágenes mentales. Ruth-Inge Heinze se enfocó en el estudio del chamanismo y promovió su estudio interdisciplinario de muchas formas. Por último, Stan Krippner es una figura liminal por excelencia, con importantes contribuciones al studio de los sueños, la cognición anómala, la mitología personal, y otras áreas de la consciencia.

Translation into Spanish by Etzel Cardeña, Ph. D.

In Memoriam

Robert Rosenthal (1933-2024), Affable Giant of Psychology and Champion of Honest Science

Etzel Cardeña

Lund University

One of the 100 most eminent psychologists in the 20th century (Haggebloom et al., 2002) Harvard and University of California-Riverside Professor Robert Rosenthal, died from an aneurysm on January 5. He is primarily known for his research on the "Pygmalion effect" or "Rosenthal Effect." In a nutshell, he showed that giving false feedback to teachers about randomly chosen students resulted in the latter achieving more than their peers because of the way they were (differentially) treated by the teachers, a type of self-fulfilling prophecy. He also co-created modern meta-analysis with Gene Glass. He was a delight to interact with, always responsive and kind when I approached him. One of his colleagues, Dave Funder, described him as "about the kindest, least pretentious person you'd ever meet" (https://news.ucr.edu/articles/2024/01/10/psychology-research-giant-robert-rosenthal-has-died).

In 2016, I had the honor to interview Bob (as he liked to be called) for an article in *Mindfield*, which I edited at that time (Rosenthal, 2016). He started his career as an experimentalist in high school when he was 14 or 15, using a set of Zener cards that Duke University parapsychology J. B. Rhine had sent him in response to a letter! Rosenthal mentioned in the interview that he was impressed by parapsychology's methodology

Journal of Anomalous Experience and Cognition
2024, Vol. 4, No. 1, pp. 13-15

https://doi.org/10.31156/jaex.26221

and analytic approaches. Incidentally, one of the students he tested was Russel Targ, the laser physicist who later became a major developer of remote viewing (e.g., Targ, 2004).

Bob considered himself an outsider to parapsychology but one who defended it against those who think it impossible, as did recently two open-minded editors of a mainstream journal who decided to publish a paper considering its quality rather than their own a-priori beliefs (Della Sala & Grafman, 2024; for examples of parapsychology-censoring editors see Cardeña, 2015).

Rosenthal developed statistical techniques to analyze psi research (e.g., Rosenthal & Rubin, 1989) and was asked by the National Research Council to review five areas of potential human enhancement. Despite finding that ganzfeld (monotonous sensory stimulation) psi research followed the most rigorous protocols of those areas, he and his co-author (Dr. Monica J. Harris) were "pressured to withdraw our… positive evaluation of psi so that… reports on other areas would be published" (Rosenthal, 2016, p. 26). They refused and he wrote a letter in *Psychological Science* mentioning the pressure and his conclusion that "we found the typical methodological quality of the Ganzfeld experiments to be superior to the typical quality of the four other areas" (Rosenthal, 1990, p. 329). The resulting book by the National Research Council (Druckman & Swets, 1988. P. iii) shamefully does not even include their names as part of the committee.

Decades before the current uproar about questionable research practices (QRP), he wrote a paper analyzing the psi ganzfeld database at that point and discussed replication and the problems of multiple testing, selective reporting, and others. His overall conclusion was that the ganzfeld psi supportive data could not be explained away by these flaws, although they likely reduced its effect size (Rosenthal, 1986).

In 1993, he helped Dean Radin get a grant from Harvard's Richard Hodgson's Memorial Fund (Radin, personal communication, January 22, 2024). Hodgson was one of the foremost original psychical researchers and he would likely be aghast to know that Harvard voted to allocate the money of his fund to "the acquisition of materials in the occult sciences" (https://sites.harvard.edu/harvard-library-funds/home/all-funds/415_513603/), among them items on "the global conspiracy," "classic true crime stories," and so on.

Considering Bob's eminence and support for psi research, next time someone states that scientists do not support psi, you may ask that person whether Rosen-

Journal of Anomalous Experience and Cognition (JAEX) ta 2024, Vol. 4, No. 1, pp. 13-15

thal and hundreds of other modern eminent scientists (https://psi-encyclopedia.spr.ac.uk/articles/eminent-people-interested-psi) are not real scientists. Bob and Dr. Harris gave me permission to publish their report and, with gratitude, I do so in this issue.

References

Cardeña, E. (2015). The unbearable fear of psi: On scientific censorship in the 21st century. *Journal of Scientific Exploration, 29*, 601-620.

Della Sala, S., & Grafman, J. (2024). Why we publish papers reporting findings we may not believe. *Cortex, 172*, A1–A2. https://doi.org/10.1016/j.cortex.2024.01.002

Druckman, D., & Swets, J. A. (1988). *Enhancing human performances: Issues, theories, and techniques.* National Academy Press.

Haggbloom, S. J., Warnick, R., Warnick, J. E., Jones, V. K., Yarbrough, G. L., Russell, T. M., Borecky, C. M., McGahhey, R., Powell, J. L., Beavers, J., & Monte, E. (2002). The 100 most eminent psychologists of the 20th Century. *Review of General Psychology, 6*(2), 139-152. https://doi.org/10.1037/1089-2680.6.2.139

Rosenthal, R. (1986). Meta-analytic procedures and the nature of replication: The ganzfeld debate. *Journal of Parapsychology, 50*, 315-336.

Rosenthal, R. (1990). Letter to the editor. *Psychological Science, 1,* 329.

Rosenthal, R. (2016). Reflections: Robert Rosenthal. *Mindfield, 8*(1), 25-26.

Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin, 106*, 332-337.

Targ, R. (2004). *Limitless mind: A guide to remote viewing and transformation of consciousness.* New World Library.

2024, Vol. 4, No. 1, pp. 16-17

Journal of Anomalous Experience and Cognition (JAEX)

In Memoriam:

H. H. Jürgen Keil (1930–2024)[1]

Jim B. Tucker

University of Virginia

Jürgen Keil, a parapsychologist who researched topics such as psychokinesis (PK) and past-life memories, died on January 7, 2024, at the age of 93. Keil was born and raised in Germany before immigrating to Australia as a young man. Once there, he attended the University of Tasmania at Hobart, obtaining both an undergraduate degree and a Ph.D. in psychology. He then spent his career as a faculty there, eventually retiring as an Emeritus Professor.

Keil only became interested in parapsychology when, as a student, he was introduced by chance to psychokinesis. While studying psychology, he was also working as a fitter and turner to earn extra income when the head of the physics department asked him to help construct a microbalance apparatus to be used in PK experiments. Those experiments produced statistically significant results, and having "observed phenomena in a sophisticated scientific setting which strongly suggested PK" (Keil, 1987, p. 2), he wanted to explore more.

This led to a collaboration with Gaither Pratt, a noted figure in the field, and a long, productive career in parapsychology. Highlights include work with Nina Kulagina, a Russian woman who appeared capable of what is now referred to as macro-PK,

1 Address correspondence to: Jim B. Tucker. M. D., University of Virginia Division of Perceptual Studies, jbt8n@uvahealth.org

https://doi.org/10.31156/jaex.26036

seeming to move small objects and cause other directly observable PK effects using only her mind. Following studies by Russian scientists, Keil and his colleagues used various controls with Kulagina and concluded that a good case could be made that her PK abilities were genuine (Keil et al., 1976).

In later years, Keil joined the work on cases of the reincarnation type (CORT), the term Ian Stevenson used to describe children who report memories of a previous life. Keil published numerous papers on the topic, both with colleagues and independently. He conducted one of his more important studies with Stevenson. Keil reinvestigated 15 cases that Stevenson had studied some 20 years before, to determine if the reports by the families had become exaggerated over time. When Keil and Stevenson compared notes, they found that the cases had not become stronger in the intervening years (with one exception). Instead, 11 of the 15 had become weaker, as the families gave fewer details. Thus, the findings contradicted the idea that families over time credit the children with more knowledge of the past life than they had actually demonstrated (Stevenson & Keil, 2000).

All told, Keil made significant contributions in multiple areas of parapsychology. In so doing, he was not afraid to push against prevailing thought—outside of parapsychology to be sure, but in it as well. For example, in the CORT research, he accepted that the cases indicated a paranormal process but argued against the proposition that reincarnation was the best way to explain them, as Stevenson had concluded. Instead, he suggested that the child had a connection with thought pools associated with the previous person. These pools had persisted but the term *survival* was not justified (Keil, 2010). Right or wrong on that issue, the ability Keil demonstrated throughout his career to question rather than assume is one that all scientists can aspire to emulate.

References

Keil, J. (1987). *Gaither Pratt: A life for parapsychology*. McFarland.

Keil, J. (2010). Questions of the reincarnation type. *Journal of Scientific Exploration, 24*(1), 79-99.

Keil, H. H. J., Herbert, B., Ullman, M., & Pratt, J. G. (1976). Directly observable voluntary PK effects: A survey and tentative interpretation of available findings from Nina Kulagina and other known related cases of recent date. *Proceedings of the Society for Psychical Research, 56*(210), 197–235.

Stevenson, I., & Keil, J. (2000). The stability of assessments of paranormal connections in reincarnation-type cases. *Journal of Scientific Exploration, 14*(3), 365–382.

Parapsychology[1]

Monica J. Harris Robert Rosenthal

Harvard University

Abstract: Monica J. Harris and Robert Rosenthal were commissioned by the National Research Council to conduct meta-analyses and review five areas of potential human enhancement. Despite finding that ganzfeld psi research followed the most rigorous protocols they were pressured to withdraw their supportive evaluation of psi, but refused to do so. This is their original report, with only very minor formatting changes. Among other things, Harris and Rosenthal concluded that "it would be implausible to entertain the null given the combined *p* from these 28 studies… when the accuracy rate expected under the null is 1/4, we estimate the obtained accuracy rate to be about 1/3." They were then asked to analyze the effect of potential design and procedure flaws and, after doing so, they concluded that: "Our analysis of the effects of flaws on study outcome lends no support to the hypothesis that ganzfeld research results are a significant function of the set of flaw variables."

Keywords: Ganzfeld; parapsychology; anomalous cognition; meta-analysis; Robert Rosenthal

There are two transition points in the recent history of parapsychology. At each point parapsychology advanced to a new level of more rigorous research and scientific respectability, though neither point earned for it full acceptance as a respectable field of scientific inquiry (Boring, 1962; Murphy, 1962, 1963; Truzzi, 1981). The first point was in 1882 when the Society for psychical Research was founded in London by a group primarily from Cambridge University. Among the distinguished presidents of this Society were William James, Henri Bergson, Lord Rayleigh, and C.D. Broad (Schmeidler, 1968). The second point was in 1927 when William McDougall, newly arrived at Duke University, was joined by J. B. Rhine (Boring, 1950; Schmeidler, 1968). It was Rhine who established the basic procedures of parapsychological research that are still employed

1 The authors are deceased but had earlier granted the editor of *JAEX* permission to publish this contribution. Because their original paper did not have an abstract, the journal editor added it for indexing purposes and left the paper, formatting and all, almost exactly as originally written.

https://doi.org/10.31156/jaex.26222

today. His best known method required the subject to guess which one of five designs was the "target" stimulus. Since the probability of guessing the correct design was .20, any subject's "psi" ability could be evaluated for statistical significance by comparing the obtained success rate with the .20 expected under the null hypothesis.

Parapsychological investigations cover a wide variety of phenomena including: telepathy (e.g., guessing a design being viewed by another); clairvoyance (e.g., guessing a design not being viewed by another); precognition (e.g., guessing a design not yet selected); psychokinesis (e.g., trying to influence the fall of a pair of dice); survival after death (e.g., reincarnation). The. first three of these are often referred to generically as ESP (extrasensory perception). Because the types of research subsumed under the topic of parapsychology range so widely, and because of the sheer number of parapsychological investigations, we have confined our discussion to a focused domain of parapsychological inquiry: the ganzfeld experiments.

Ganzfeld Experiments

In these experiments subjects typically are asked to guess which of four stimuli had been "transmitted" by an agent or sender with these guesses made under conditions of sensory restriction (usually white noise and an unpatterned visual field). There were several strong reasons for the selection of this domain of parapsychological research:

1. The domain is of recent origin so that even the earliest studies managed to avoid some of the older problems found in parapsychological research (Hansen & Lehmann, 1895; Kennedy, 1938, 1939; Moll, 1898; Rosenthal, 1965, 1966; Warner & Raible, 1937).
2. Because of the recency of the research, access to original data was more likely than for some of the older areas (Rao, 1985).
3. The domain is considered an especially promising area of parapsychological inquiry (Hyman, 1985; Rao, 1985).
4. Investigations in this area have been carried out by respected researchers (Hyman, 1985).
5. The area has been the subject of recent sophisticated public debate by eminent investigators and critics of the area (Honorton, 1985; Hyman, 1985; Rao, 1985).
6. As an outgrowth of this debate, two formal meta-analyses of this area have become available (Honorton, 1985; Hyman, 1985).

Meta-Analytic Results

Five indices of "psi" success have been employed in ganzfeld research (Honorton, 1985). One criticism of research in this area is that some investigators employed several such indices in their studies and failed to adjust their reported levels of significance (p) for the fact that they had made multiple tests (Hyman, 1985). Since most studies employed a particular one of these five methods, the method of direct hits, Honorton focused his meta-analysis on just those 28 studies (of a total of 42) for which direct hit data were available.

The method of direct hits scores a success only when the single correct target is chosen out of a set of t total targets. Thus the probability of success on a single trial is $1/t$ with t usually = 4 but sometimes 5 or 6. The other methods, employing some form of partial credit, appear to be more precise in that they use more of the information available. Although they differ in their interpretation of the results, Honorton (1985) and Hyman (1985) agree quite well on the basic quantitative results of the meta-analysis of these 28 studies. This agreement holds both for the estimation of statistical significance (Honorton, 1985, p.58) and of effect size (Hyman, 1985, p. 13).

Stem-and-Leaf Display

Table 3 shows a stem-and-leaf display of the 28 effect slze estimates based on the direct hits studies summarized by Honorton (1985, p. 84). The effect size estimates shown in Table 3 are in units of Cohen's h which is the difference between (a) the arcsine transformed proportion of direct hits obtained and (b) the arcsine transformed proportion of direct hits expected under the null hypothesis (i.e., $1/t$). The advantage of h over j, the difference between raw proportions, is that all h values that are identical are identically detectable while all j values that are identical (e.g., .65-.45 and .25-.05) are not equally detectable (Cohen, 1977, p. 181).

The stem-and-leaf display of Cohen's h values is shown on the left and the display is summarized on the right. Tukey (1977) developed the stem-and-leaf plot as a special form of frequency distribution to facilitate the inspection of a batch of data. Each number in the data batch is made up of one stem and one leaf, but each stem may serve several leaves. Thus, the stem .1 is followed by leaves of 3, 8, 8 representing the numbers .13, .18, .18. The first digit is the stem; the next digit is the leaf. The stem-and-leaf display functions as any other frequency distribution but the original data are retained precisely.

TABLE 3 Stem-and-Leaf Plot and Statistical Summary of Ganzfeld Studies Employing Criterion of Direct Hits

Cohen's h

Stem	Leaf
1.4	4
1.3	3
1.2	
1.1	
1.0	
.9	
.8	
.7	3
.6	
.5	8
.4	0 2 2 2 4
.3	1 2 2 4 4 7 8
.2	2
.1	3 8 8
.0	7 7 9
-.0	5
-.1	0
-.2	
-.3	2
-.4	0
-.5	
-.6	
-.7	
-.8	
-.9	3

Summary Statistics	
Maximum	1.44
Quartile$_3$ (Q_3)	.42
Median (Q_2)	.32
Quartile$_1$ (Q_1)	.08
Minimum	-.93
$Q_3 - Q_1$	.34
$\hat{\sigma}$: [.75($Q_3 - Q_1$)	.26
S	.45
Mean[a]	.28
N	28
Proportion positive sign	.82
Z of proportion positive	3.40
Combined Stouffer Z	6.60
t test of mean Z	3.23
Correlation between h and Z	.86
Correlation between h and raw i	.98
Number of studies with mean Z = 0.00 required to bring combined results to p > .05	423

Confidence Intervals*	from	to
80%	.17	.39
95%	.11	.45
99%	.04	.52
99.9%	-.03	.59

* where N = 28 studies

[a] Unweighted; weighted mean = .23

Distribution. From Table 3 we see that the distribution of effect sizes· is unimodal with the bulk of the results (80%) falling between -.10 and .58. The distribution is nicely symmetrical with the skewness index ($g1$ = .17) only 24% of that required for significance at $p < .05$ (Snedecor & Cochran, 1980, pp. 78-79, 492). The tails of the distribution, however, are too long for normality with kurtosis index g_2=.2.04, p=.02. Relative to what we would expect from a normal distribution, we have studies that show larger positive and larger negative effect sizes than would be reasonable. Indeed, the two largest positive effect sizes are significant outliers at $p < .05$, and the largest negative effect size approaches significance with a Dixon index of .37 compared to one of .40 for the largest positive effect size (Snedecor & Cochran, 1980, pp. 279-280, 490). The total sample of studies is still small; however, if a much larger sample showed the same result, that would be a pattern consistent with the idea that both strong positive results ("psi") and strong negative results ("psi-missing") might be more likely to find their way into print or at least to be more available to a meta-analyst.

Effect size. The bulk of the results (82%) show a positive effect size where 50%

would be expected under the null (p = .0004). The mean effect size, h, of .28 is equivalent to having a direct hit rate of .38 when .25 was expected under the null. The 95% confidence interval suggests the likely range of effect sizes to be from .11 to .45, equivalent to accuracy of guessing rates of .30 to .46 when .25 was expected under the null hypothesis.

Significance testing. The overall probability that obtained accuracy was better than the accuracy expected under the null was a p of $3.37/10^{11}$ associated with a Stouffer z of 6.60 (Mosteller & Bush, 1954; Rosenthal, 1978a, 1984).

File drawer analysis. A combined p as low as that obtained can be used as a guide to the tolerance level for null results that never found their way into the meta-analytic data base (Rosenthal, 1979, 1984). It has long been believed that studies failing to reach statistical significance may be less likely to be published (Sterling, 1959; Rosenthal, 1966). Thus it may be that there is a residual of nonsignificant studies languishing in the investigators' file drawers. Employing simple calculations, it can be shown that, for the current studies summarized, there would have to be 423 studies with mean p =.50, one-tailed, or z = 0.00 in those file drawers before the overall combined p would become just > .05.

That many studies unretrieved seems unlikely for this specialized area of parapsychology (Hyman, 1985; Honorton, 1985). Based on experience with meta-analyses in other domains of research (e.g., interpersonal expectancy effects) the mean or effect size for nonsignificant studies is not 0.00 but a value pulled strongly from 0.00 toward the mean Z or mean effect size of the obtained studies (Rosenthal & Rubin, 1978).

Comparison to an Earlier Meta-Analysis

We felt it would be instructive to compare the results of the ganzfeld research meta-analysis by Honorton (1985) to the results of an older and larger meta-analysis of another controversial research domain --that of interpersonal expectancy effects (Rosenthal & Rubin, 1978). In that analysis, eight areas of expectancy effects were summarized; effect sizes (Cohen's d roughly equivalent to Cohen's h) ranged from .14 to 1.73 with a grand mean d of .70. Honorton's mean effect size (h =-.28) exceeds the mean d of two of the eight areas (reaction time experiments [d =.17]; and studies employing laboratory interviews [d =.14]).

The earlier meta-analysis displayed the distribution of the Z's associated with the obtained p levels. Table 4 shows a comparison of the two meta-analyses' distri-

butions of *Z*s. It is interesting to note the high degree of similarity in the distributions of significance levels. The total proportion of significant results is somewhat higher for the ganzfeld studies but not significantly so ((1) =1.07, *N* = 373, *p* = .30, =.05).

Interpretation of Meta-Analytic Results

Although the results of the meta-analysis are clear, the meaning of these results is open to various interpretations (Truzzi, 1981). The most obvious interpretation might be that at a very low *p*. and with a fairly impressive effect size, the ganzfeld psi phenomenon has been demonstrated. However, there are rival hypotheses that will need to be considered. many of them put forward in the recent detailed evaluation of the ganzfeld research area by Hyman (1985).

Procedural Rival Hypotheses

Sensory leakage. A standard rival hypothesis to the hypothesis of ESP is that sensory leakage occurred and that the receiver was knowingly or unknowingly cued by the sender or by an intermediary between the sender and receiver. As early as 1895, Hansen and Lehmann (1895) had described "unconscious whispering" in the laboratory and Kennedy (1938, 1939) was able to show that senders in telepathy experiments could give auditory cues to their receivers quite unwittingly. Ingenious use of parabolic sound reflectors made this demonstration possible. Moll (1898), Stratton (1921), and Warner and Raible (1937) all gave early warnings on the dangers of unintentional cueing (for summaries see Rosenthal, 1965a, 1966). The subtle kinds of cues described by these early workers were just the kind we have come to look for in searching for cues given off by experimenters that might serve to mediate the experimenter expectancy effects found in laboratory settings (Rosenthal, 1966, 1985).

By their nature, ganzfeld studies tend to minimize problems of sensory cueing. An exception occurs when the subject is asked to choose which of four (or more) stimuli had been "sent" by another person or agent. When the same stimuli held originally by the sender are shown to the receiver, finger smudges or other marks may serve as cues. Honorton has shown, however, that studies controlling for this type of cue yield at least as many significant effects as do the studies not controlling for this type of cue.

Recording errors. A second rival hypothesis has nearly as long a history. Kennedy and Uphoff (1939) and Sheffield and Kaufman (1952) both found biased errors of recording the data of parapsychological experiments. In a meta-analysis of 139,000

recorded observations in 21 studies, it was found that about 1% of all observations were in error and, that of the errors committed, twice as many favored the hypothesis as opposed it (Rosenthal, 1978b). While it is difficult to rule recording error out of ganzfeld studies (or any other kind of research) their magnitude is such that they could probably have only a small biasing effect on the estimated average effect size (Rosenthal, 1978b, p. 1007).

Intentional error. The very recent history of science has reminded us that while fraud in science is not quite of epidemic proportion it must be given close attention (Broad & Wade, 1982; Zuckerman, 1977). Fraud in parapsychological research has been a constant concern, a concern found justified by periodic flagrant examples (Rhine, 1975). In the analyses of Hyman (1985) and Honorton (1985), in any case, there appeared to be no relationship between degree of monitoring of participants and the results of the study.

Statistical Rival Hypotheses

File drawer issues. The problem of biased retrieval of studies for any meta-analysis was described earlier. Part of this problem is addressed by the 10 year old norm of the Parapsychological Association of reporting negative results at its meetings and in its journals (Honorton, 1985). Part of this problem is addressed also by Blackmore who conducted a survey to retrieve unreported ganzfeld studies. She found that 7 of her total of 19 studies (37%) were judged significant overall by the investigators. This proportion of significant results was not significantly (or appreciably) lower than the proportion of published studies found significant. A problem that seems to be a special case of the file drawer problem was pointed out by Hyman (1985). That was a possible tendency to report the results of pilot studies along with subsequent significant results when the pilot data were significant. At the same time it is possible that pilot studies were conducted without promising results, pilot studies that then found their way into the file drawers. In any case, it is nearly impossible to have an accurate estimate of the number of unretrieved studies or pilot studies actually conducted. Chances seem good, however, that there would be fewer than the 423 results of mean z = 0.00 required to bring the overall combined g to > .05.

Multiple testing. Each ganzfeld study may have more than one dependent variable for scoring a success. If investigators employ these dependent variables sequentially until they find one significant at $p < .05$ the true p will be higher than .05 (Hyman, 1985). Although a simple Bonferroni procedure can be used to adjust for this problem

2024, Vol. 4, No. 1, pp. 18-33
Journal of Anomalous Experience and Cognition (JAEX)

(e.g., by multiplying the lowest obtained by the number of dependent variables tested) this adjustment is quite conservative (Rosenthal & Rubin, 1983). The adjustment can be made with greater power if the investigators are willing to order or to rate their dependent variables on a dimension of importance (Rosenthal & Rubin, 1984, 1985). Most useful, however is a procedure that uses all the data from all the dependent variables with each one weighted as desired so long as the weighting is done before the data are collected (Rosenthal & Rubin, 1986).

Randomization. Hyman (1985) has noted that the target stimulus may not have been selected in a truly random way from the pool of potential targets. To the extent that this is the case the *p* values calculated will be in error. Hyman (1985) and Honorton (1985) disagree over the frequency in this sample of studies of improper randomization. In addition, they disagree over the magnitude of the relationship between inadequate randomization and study outcome. Hyman felt this relationship to be significant and positive; Honorton felt this relationship to be nonsignificant and negative. Since the median *g* level of just those 16 studies employing random number tables or generators (z=.94) was essentially identical to that found for all 28 studies it seems unlikely that poor randomization procedures were associated with much of an increase in significance level (Honorton, 1985, p. 71).

Statistical errors. Hyman (1985) and Honorton agree that six of the 28 studies contained statistical errors. However, the median effect size of these studies (h=.33) was very similar to the overall median (h=.32) so that it seems unlikely that these errors had a major effect on the overall effect size estimate. Omitting these six studies from the analysis decreases the mean *h* from .28 to .26. Such a drop is equivalent to a drop of the mean accuracy rate from .38 to .37 when .25 is the expected value under the null.

Independence of studies. Because the 28 studies were conducted by only 10 investigators or laboratories, the 28 studies may not be independent in some sense. While under some data analytic assumptions such a lack of independence would have implications for significance testing, it does not in the ganzfeld domain because of the use of trials rather than subjects as the independent sampled unit of analysis. The overall significance level, then, depends on the results of all trials, not the number of studies, of subjects or investigators (any of which may be viewed as fixed rather than random).

However, the lack of independence of the studies could have implications for the estimation of effect sizes if a small proportion of the investigators were responsible for

all the nonzero effects. In that case the average of the investigators' obtained effects would be much smaller than the average of the studies' obtained effects.

In an extreme example the median effect size of a sample of studies could be .50 while the median effect size of a sample of investigators could be zero because very few investigators obtained any nonzero effect. That did not turn out to be the case for the ganzfeld domain. The median effect size (*h*) was identical (.32) for the 28 studies and the 10 investigators or laboratories. The mean effect sizes, however, did differ somewhat with a lower mean for labs (.23) than for studies (.28). The proportions of results in the positive direction were very close: .82 for studies and .80 for labs.

It is of interest to note that investigators did differ significantly from one another in the magnitude of the effects they obtained with $F(9, 18) = 3.81$, $p < .01$, intra-class $p = .63$. There was little evidence to suggest, however, that those investigators tending to conduct more studies obtained higher mean effect sizes; the f(I, 18) testing that contrast was 0.38, $p = .54$, $r^2 = .14$.

Conclusion

On the basis of our summary and the very valuable meta-analytic evaluations of Honorton (1985) and Hyman (1985), what are we to believe? The situation for the ganzfeld domain seems reasonably clear. We feel it would be implausible to entertain the null given the combined *p* from these 28 studies. Given the various problems or flaws pointed out by Hyman and Honorton, the true effect size is almost surely smaller than the mean *h* of .28 equivalent to a mean accuracy of 38% when 25% is expected under the null. We are persuaded that the net result of statistical errors was a biased increase in estimated effect size of at least a full percentage point (from 37% to 38%). Furthermore, we are persuaded that file drawer problems are such that some of the smaller effect size results have probably been kept off the market. If pressed to esti-

2 After preparation of this paper we learned of a possible problem in the randomization procedures employed by the investigator contributing the largest number (9) of Ganzfeld studies to the set of 28 summarized in this section. Accordingly we constructed Table 4a to investigate the effect on the mean and median effect sizes of omitting all the studies conducted by this investigator. The top half of Table 4a shows this effect when we consider the 28 studies, as the units of analysis. Omitting the 9 questioned studies lovers the mean effect size from .28 to .26 and does not change the median effect size which remains at .32. The lower half of Table 4a shows this effect when we consider the 10 investigators as the units o£ analysis. Omitting the investigator in question lowers the mean effect size from .23 to .22 but raises the median effect size from .32 to .34. It seems clear that the questioned randomization of the 9 studies of this investigator cannot have contributed substantially to an inflation of the overall effect size.

mate a more accurate effect size we might think in terms of a shrinkage of *h* from the obtained value of .28 to perhaps an *h* of .18. Thus, when the accuracy rate expected under the null is 1/4, we estimate the obtained accuracy rate to be about 1/3.[2]

Postscript

We have been asked to respond to a letter from the committee listing questions about the presence and consequences of methodological flaws in the ganzfeld studies discussed by Honorton (985), Hyman (1985), Hyman and Honorton (1986), Rosenthal (1986), and by the present authors of this paper. Our response is in two parts. In part 1, we examine the likely effects of flaws on the meta-analytic results of the ganzfeld studies. In part 2, we examine the results of a series of new studies designed to address the flaws discussed by Hyman and Honorton in their individual and joint papers.

Flaw Effects

The committee has called our attention to possible flaws in the randomization procedure employed by Sargent and his colleagues. In its letter it noted that Honorton agreed with Hyman about the assignment of these randomizations flaws to the Sargent study. However, Honorton states in two letters that this agreement was not reached (personal communications of November 25, 1987, and January 5, 1988). Apparently, experts on the ganzfeld research disagree on whether the Sargent studies' randomization procedures are flawed given all the evidence available to both, evidence which is summarized in papers by Blackmore (1987), Harley and Matthews (1987), and Sargent (1987).

For purposes of this postscript and the following data analyses, we are going to assume that Hyman is correct in his assignment of randomization flaws and all other flaws he assigned in his 1985 paper. The heart of the matter is the relationship of flaws to research results and that is what our analyses are designed to investigate. In a 1986 manuscript, Hyman suggested that the relationship of flaws to study outcomes should be examined in a multivariate manner. Accordingly, that is the nature of our analyses in our first pass effort to examine the likelihood that methodological flaws are driving the results of the ganzfeld studies to an appreciable degree.

Canonical analysis. The most general of the multivariate procedures examines the maximum relationship that can be found between two sets of variables, for example, a set of predictor variables and a set of outcome variables. In our analysis

the predictor variables were Hyman's (1985) flaw variables of documentation (DOC), feedback (FB), randomization (R), security (SEC), single target (ST), and statistical analysis (STAT), all coded as 0 if adequate or I if not adequately done or not adequately specified. The outcome variables were the significance level ~ and the effect size Cohen's *h*. The adjusted canonical correlation was only .46, a magnitude that for two-predicted-from-six could have arisen under the null hypothesis 54 times out of 100 $F(12,40) = 0.91$). Interestingly, three of the six flaw variables correlated positively with the flaw canonical variable and with the outcome canonical variable (DOC, FB, R) but three correlated negatively (SEC, ST, STAT). Thus, the canonical analysis gives no support to the hypothesis that the research results are a significant function of the set of flaw variables.

Regression analysis. Separate analyses were also done for each of the outcome variables *Z* and *h*. The battery of predictor variables correlated only .44 with Cohen's *h* ($F(6,2I) = 0.84$, $p = .56$) and .57 with *z* ($F(6,2I) = 1.65$, $p = .18$). For neither of the outcome variables did any of the six predictors account for a significant proportion of variance either in zero-order form or after partialing. Since there were two methods of partialing employed, a total of 36 (3 methods x 6 predictors x 2 outcome variables) *t*'s were computed, none of which reached the .05 level. Regression analyses, therefore, gave no more support than did the canonical analysis to the hypothesis that ganzfeld research results are a significant function of the set of flaw variables.

New Evidence

Hyman (1985) and Honorton (1985) were agreed (Hyman and Honorton, 1986) that new studies were needed that would take account of the flaws they had found in their critiques of earlier research. Since our present paper was completed we have learned of a series of 10 new studies conducted by Honorton, one of the four investigators singled out by the Committee on Techniques for the Enhancement of Human Performance as among the best in the country (Druckman and Swets I 1988, p. 22).

The series of 10 ganzfeld studies yielded a combined *z* of 2.791, $p = .0026$ and a mean *h* of .23. This effect size, based on 10 studies, is only slightly smaller than the mean effect size of Sargent's nine studies ($h = .30$) and is very close to the mean effect size of the remaining 19 studies ($h = .26$; see Table 4a). For the original 28 studies plus the 10 new ones from Honorton's lab I the combined *z* is now 7.10 and the mean effect size is now an *h* of .27. Omitting Sargent's nine studies changes matters very little-- *z* is now 5.74 and $h = .25$. In short, the new evidence based on studies designed

to meet earlier methodological objections is very consistent with the earlier evidence and makes the null hypothesis still more implausible.

Conclusion

Our analysis of the effects of flaws on study outcome lends no support to the hypothesis that ganzfeld research results are a significant function of the set of flaw variables. In addition, a series of 10 new studies designed to control for earlier presumed flaws yielded results quite consistent with the original set of 28 studies.

References

Blackmore, S. (1987). A report of a visit to Carl Sargent's laboratory. *Journal of the Society for Psychical Research, 54*, 186-198.

Druckman, D., & Swets, J. A. (Eds.) (1988). *Enhancing human performance. Issues, theories, and techniques*. Washington, DC: National Academy Press.

Harley, T., & Matthews, G. (1987). Cheating, psi, and the appliance of science: A reply to Blackmore. *Journal of the Society for Psychical Research, 54*, 199-207.

Hyman, R. (1986). *To conclude or not to conclude: A reply to the commentators.* Unpublished manuscript, University of Oregon.

Hyman, R., & Honorton, C. (1986). A joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology, 50*, 351-364.

Rosenthal, R. (1986). Meta-analytic procedures and the nature of replication: The ganzfeld debate. *Journal of Parapsychology*, *50*, 315-336.

Sargent, C. (1987). Sceptical fairytales from Bristol. *Journal of the Society for Psychical Research*, *54*, 208-218.

Table 4.

Proportion of Studies Reaching Critical Levels of Significance for Two Research Areas

Interval for *Z*	Expected Proportion	Expectancy Research[a]	Ganzfeld Research[b]	Difference
Unpredicted Direction				
-1.65 and below	.05	.03	.07	+.04
NotSignificant				
-1.64 to + 1.64	.90	.60	.50	-.10
Predicted Direction				
+1.65 and above	.05	.36	.43	+.07
+2.33 and above	.01	.19	.25	+.06
+3.09 and above	.001	.12	.18	+-06
+3.72 and above	.0001	.07	.04	-.03

[a] ***N = 345*** studies; from Rosenthal & Rubin (1978). [b] ***N*** = 28 studies; from Honorton (1985).

Table 4a.

Effects on Effect Size (h) of Removing Studies by Sargent

	Mean	*Median*
Sargent's Studies (N = 9)	.30	.37
Analysis by Studies		
Including Sargent (*N* = 28)	.28	.32
Omitting Sargent (*N* = 19)	*.26*	*.32*
Difference	.02	.00
Analysis by Investigators		
Including Sargent (*N* = 10)	.23	.32
Omitting Sargent (*N* = 9)	*.22*	*.34*
Difference	.01	-.02

Parapsychologie

Monica J. Harris Robert Rosenthal

Résumé: Monica J. Harris et Robert Rosenthal ont été chargés par le National Research Council (« Conseil National de la Recherche ») de mener des méta-analyses et d'examiner cinq domaines d'amélioration potentielle de l'être humain. Alors que les auteurs ont constaté que les recherches psi utilisant le Ganzfeld avaient la rigueur méthodologique la plus importante, ils ont subi des pressions pour retirer leur évaluation favorable à la psi. Ce que les auteurs ont refusé de faire. Le rapport publié est le rapport original, qui n'a subi que des modifications mineures de mise en forme. Entre autres choses, Harris et Rosenthal ont conclu que «il serait peu plausible d'accepter le résultat nul étant donné les p combinés de ces 28 études... lorsque le taux de réussite attendu pour le résultat nul est de 1/4, nous estimons que le taux de réussite obtenu est d'environ 1/3». Il leur a ensuite été demandé d'analyser l'effet des défauts potentiels de conception et de procédure et, après l'avoir fait, ils ont conclu que: «Notre analyse de l'effet des défauts sur les résultats obtenus ne confirme pas l'hypothèse selon laquelle les résultats issus de la méthode ganzfeld constituent en soi une catégorie parmi l'ensemble des variables possibles de défauts.»

Translation into French by Antoine Bioy, Ph. D.

Parapsychologie

Monica J. Harris Robert Rosenthal

Zusammenfassung: Monica J. Harris und Robert Rosenthal von der Harvard University wurden vom Nationalen Forschungsrat beauftragt, Meta-Analysen durchzuführen und fünf Bereiche des potenziellen Human Enhancement zu überprüfen. Obwohl sie feststellten, dass die Ganzfeld-Psi-Forschung den strengsten Protokollen folgte, wurden sie unter Druck gesetzt, ihre positive Bewertung von Psi zurückzuziehen, was sie jedoch ablehnten. Dies ist ihr Originalbericht mit nur sehr geringfügigen Änderungen in der Formatierung. Unter anderem kamen Harris und Rosenthal zu dem Schluss, dass "es unplausibel wäre, die Null-Hyothese anzunehmen, wenn man die p-Werte aus diesen 28 Studien zusammennimmt... wenn die unter der Null-Hypothese erwartete Genauigkeitsrate 1/4 beträgt, schätzen wir die erhaltene Genauigkeitsrate auf etwa 1/3." Sie wurden dann gebeten, die Auswirkungen möglicher Design- und Verfahrensfehler zu analysieren, und kamen zu folgendem Schluss: "Unsere Analyse der Auswirkungen von Fehlern auf das Studienergebnis unterstützt nicht die Hypothese, dass die Ganzfeld-Forschungsergebnisse eine signifikante Funktion einer Anzahl von Fehlervariablen sind."

Translation into German by Eberhard Bauer, Ph. D.

Parapsicologia

Monica J. Harris Robert Rosenthal

Resumo: Monica J. Harris e Robert Rosenthal foram designados pelo Conselho Nacional de Pesquisa para realizar meta-análises e revisar cinco áreas de potencial aprimoramento humano. Apesar de terem constatado que a pesquisa Psi Ganzfeld seguia os protocolos mais rigorosos, foram pressionados retirar sua avaliação favorável ao psi, mas se recusaram a fazê-lo. Este é seu relatório original, com apenas mudanças de formatação muito pequenas. Entre outras coisas, Harris e Rosenthal concluíram que "seria implausível considerar a hipótese nula dada a combinação de p desses 28 estudos... quando a taxa de precisão esperada sob a hipótese nula é de 1/4, estimamos que a taxa de precisão obtida seja cerca de 1/3." Eles foram, então, solicitados a analisar o efeito de possíveis falhas de design e procedimentos e, depois de o fazê-lo, concluíram que: "Nossa análise sobre os efeitos das falhas no resultado do estudo não dá suporte à hipótese de que os resultados da pesquisa ganzfeld sejam uma função significativa do conjunto de variáveis de falha."

Translation into Portuguese by Antônio Lima, Ph. D.

Parapsicología

Monica J. Harris Robert Rosenthal

Resumen: Monica J. Harris y Robert Rosenthal recibieron el encargo por parte del National Research Council de realizar meta-análisis y revisar cinco áreas de potencial mejora humana. A pesar de descubrir que la investigación psi ganzfeld seguía los protocolos más rigurosos, se les presionó para que retiraran su evaluación de apoyo a psi, pero se negaron a hacerlo. Este es su informe original, con sólo unos pequeños cambios de formato. Entre otras cosas, Harris y Rosenthal concluyeron que "sería inverosímil considerar la nulidad dada la *p* combinada de estos 28 estudios... cuando la tasa de precisión esperada bajo la nulidad es de 1/4, estimamos que la tasa de precisión obtenida es de aproximadamente 1/3". A continuación, se les pidió que analizaran el efecto de posibles defectos de diseño y procedimiento y, tras hacerlo, concluyeron que: "Nuestro análisis de los efectos de los defectos sobre el resultado del estudio no apoya la hipótesis de que los resultados de la investigación ganzfeld sean una función significativa del conjunto de variables de los defectos".

Translation into Spanish by Etzel Cardeña, Ph. D.

Open Data
P-R

Testing the Effects of Personality-Related Beliefs on Micro-PK[1]

Marissa-Julia Jakob* Moritz C. Dechamps* Markus A. Maier

Ludwig-Maximilians-Universität, München

Abstract: *Objective.* This preregistered study investigates mind-matter interactions by testing observer effects on quantum random number generator (QRNG) outcomes mediated by implicit intentions. *Methods.* We evaluated participants' personality traits (PTs), and presented them with goal-related or neutral stimuli based on QRNG outputs. We predicted deviations from chance, with high PT scorers expected to observe more PT-related sentences. We conducted three micro-psychokinesis (micro-PK) experiments for Cluster C's PTs: dependent, avoidant, and obsessive-compulsive. *Results.* The results revealed strong evidence (Bayes Factor > 10) for a micro-PK effect in the dependent PT group, with high scorers observing more sentences addressing their concerns than expected by chance. We did not find strong evidence for the other PT groups or low scorers. *Conclusion.* These findings suggest that intentional observation biases QRNG outcomes related to individuals' implicit concerns, potentially leading to self-fulfilling prophecies. The study's implications are discussed within the Unus Mundus model and the Model of Pragmatic Information.

Keywords: micro-psychokinesis, mind-matter interaction, quantum measurement, intentional observation, personality traits, Unus Mundus, Model of Pragmatic Information

Highlights

- Quantum-based psychophysical correlation models provide an extension of QM in which intentional observations contribute to the formation of macroscopic realities.

1 Shared first authorship**.** Address correspondence to: Marissa-Julia Jakob, Moritz C. Dechamps**,** Department of Psychology, LMU Geschwister-Scholl-Platz 1, 80539 Munich, Germany, **MarissaJulia.Jakob@psy.lmu.de**, **Moritz.Dechamps@psy.lmu.de**

https://doi.org/10.31156/jaex.23809

- In three preregistered micro-PK experiments, a task for three Cluster C PTs (dependent, avoidant, and obsessive-compulsive) evaluated observers´ biases on QRNG outcomes.

- The study partially confirmed the existence of a micro-PK effect in the data: Concerning the PT group "high-dependent," strong evidence for H_1 was found.

Following years of psychical research, the psychologist/philosopher William James concluded that extraordinary phenomena "can never be fully explained away, they can also never be susceptible of full corroboration" (in Wilson, 1985, p. 158). Wilson (1985) subsequently derived James's Law, claiming that the evidence is sufficient "to convince those who are willing to be convinced, but never enough to win over the sceptics" (p. 158–159). This law still applies to various frontier phenomena in scientific research. One particular group of exceptional phenomena involves the emergence of realities based on observers' specific motivational states not mediated by behavioral interactions between individuals and their environments. These purely mind-induced effects on physical settings challenge the causal closure of accepted theories regarding physical reality. One area of research called "micro-psychokinesis" (micro-PK; see Varvoglis & Bancel, 2015) is devoted to a subset of these mind-matter interactions.

Micro-PK applies various experimental settings to tackle the mind–matter relation by investigating intentional observers' influence on deviations from randomness, more recently with a focus on observer effects during quantum measurements. This research has a long tradition (e.g., Jahn et al., 1987; Stanford, 1976; Stanford et al., 1975; Schmidt, 1974) and has shown significant effects: observers' conscious or unconscious motivational states seem to influence the outcomes of a quantum number generator (QRNG) in accordance with their motives or intentions through the sole act of observation. Meta-analyses have supported the efficacy of intentional observations on the probabilities of quantum-based events (Bösch et al., 2006; Radin & Nelson, 1989).

When the unconventionally high heterogeneity of effect sizes in these analyses raised criticism, Varvoglis and Bancel (2015) calculated that an unrealistically large number of non-significant studies would be required in the file drawer if micro-PK effects were merely a phenomenon of publication bias (see also Radin et al., 2006). However, many original results could not be replicated directly or declined over time (e.g., Dechamps et al., 2021; Jahn et al., 2000; Dechamps & Maier, 2019*;* but see Mossbridge & Radin, 2021). These ambiguous findings with regard to direct replications and within-study robustness of micro-PK effects lead to different interpretations: skeptics argue that the initial effects are artifacts of error, random fluctuation, or based on questionable research practices (QRPs; e.g., Alcock, 2003; Wagenmakers et al., 2011), while proponents evaluate the findings as providing sufficient evidence (Cardeña,

2018; Cardeña et al., 2015; Radin, 2006). This debate and the actual status of the empirical evidence both reflect the validity of James´s Law in this field.

As a solution to this debate, Rabeyron (2020) suggested that unusual patterns in experimental psi research should be regarded not as impediments or random fluctuations but opportunities to investigate these effects' true nature (see also Atmanspacher et al., 2002; von Lucadou et al., 2007; von Lucadou, 2015). He specifically argued that preregistrations within psi research might enhance the results' confirmatory value. The present study was preregistered to determine whether psi effects can occur under these circumstances.

Quantum-Based Models of Psychophysical Correlations

Although the indeterminacy principle (quantum randomness, Born, 1926) during quantum measurement is considered an ontic principle in orthodox quantum physics (Greenstein & Zajonc, 2006), several authors (e.g., Mensky, 2014; Pradhan, 2012; Penrose & Hameroff, 2011; Stapp, 2007; Stanford, 1990) propose that intentional observers might be able to influence these quantum probabilities during measurement operations, making an outcome more likely than predicted by the Born rule (Born, 1926). Such quantum-based psychophysical correlation models may be understood as an extension of quantum mechanics, considering intentional observations and their effects on macroscopic reality.

One specific variant, the *Unus Mundus* Model (UMM), proposed by Pauli and Jung (see Atmanspacher, 2014; Atmanspacher et al., 2013), offers an elegant framework to explain psychophysical interactions during quantum measurements. In contrast to alternative observational quantum theories, the dual-aspect feature of the UMM provides a solution to the existing mind–matter gap (see Chalmers, 1995). Pauli and Jung propose that mind and matter form a unity on a deeper level of reality called *unus mundus*. The UMM considers this reality to be pre-conscious and pre-material as well as collective, merging unconscious (supra-)individual information and its corresponding quantum states. It overcomes and precedes the classical subject–object distinction. By implementing such a pre-reality as an interface between the two substances mind and matter, this approach allows us to describe different types of observer-dependent mind–matter interactions. The Cartesian dual realities, conscious experience and classical matter, become two distinct aspects of the world through an act of knowledge transition (from unknown to known), referred to as epistemic split (Atmanspacher, 2020).

According to general quantum theory (GQT, Filk & Römer, 2011), a mathematical formalization of the Pauli-Jung conjecture, conscious observation is the key process to initiate an epistemic split. Prior to this, mind and matter are believed to form acausal and non-local correlations of their pre-conscious and pre-material states in the *unus mundus*. After performing an epistemic split through conscious observation, both substances emerge as distinctive forms—the conscious mind and the macroscopic matter. They cannot interact directly anymore, as they now appear as two different substances. Still, they remain related to each other because of their common ground. To describe the specific nature of this indirect relation between the conscious mind and the corresponding macroscopic physical environment, Atmanspacher and Fach (2013) suggested two types of interactions: structural and induced correlations (see also Atmanspacher, 2020).

Epistemic splits based on structural correlations (SCs) form dual realities through passive, conscious registrations of classical physical occurrences. SCs are considered reactive in nature, such as in conscious perceptions of ongoing physical events, since they follow the Born rule during the transfer from the pre-material quantum states from the *unus mundus* into the classical state and its conscious experience. Psychosomatic correlations exemplify such SCs. This type of entanglement correlation is considered stable and reproducible (Atmanspacher, 2020), likely due to its reactive and passive nature, since such manifestations do not demand the intentional agency of the individual (Maier et al., 2022).

By contrast, epistemic splits caused by induced correlations (ICs) distort the balance between mental and physical reality typical of SCs (Atmanspacher, 2020). They are at least rudimentarily based on some form of intentional agency. ICs include active, goal-based formations of physical realities that bias the Born rule along an observer's motivational state during observation. That is, autonomous individuals exert a mental impact on reality creation out of a quantum superposition (see also Maier et al., 2022). Micro-PK phenomena are a prototypical example of such correlations, whereby the observers' goals form corresponding realities. Pauli and Jung (see Atmanspacher et al., 2013) labeled such phenomena "synchronistic events" and emphasized the unifying principle of meaning in their occurrence. Owing to the autonomous, non-deterministic aspect of personal meaning, they also emphasized the elusive nature of these effects, which cannot be described with deterministic laws or demonstrated objectively. Consequently, in contrast to SCs, ICs are not considered exactly reproducible across studies. The reason for this is that the psyche, as a driving force of reality creation in this context, adds an active, subjective component not captured by purely passive, objective descriptions of physical laws (Maier et al., 2022; see also Bierman, 2001; Jahn & Dunne, 1997). Atmanspacher (2020) and von Lucadou et al.

(2007) provided formal models to support this view. According to them, ICs violate the "no-transmission theorem" (von Lucadou et al., 2007) and therefore cannot be stable and reproducible. Any effects based on ICs, including micro-PK, must decline over the course of further confirmation attempts. In sum, Pauli and Jung's UMM and the model of pragmatic information (MPI; von Lucadou et al., 2007) form the theoretical basis for this study's predictions regarding micro-PK.

An experimental micro-PK design should provide the following conditions that allow ICs to be studied scientifically: 1) The mind, as the causal source of an IC, must involve—among conscious thoughts—unconscious mental processes to affect physical systems (QRNG outcomes), since its impact is transferred through the pre-conscious and pre-material realm into the physical world. 2) A meaningful connection between the mental impulse and the quantum-based RNG outcome must be established. That is, the stimulus material to be selected by the QRNG must include subjectively relevant information. 3) Regarding the effect's replicability, an initial appearance in a first study demonstrating the existence of the effect might most likely be followed by a decline of the effect during replication attempts. This is the consequence of the effect's elusive nature, as described in the UMM and mathematically formalized in the MPI. The present study aimed to address the first two pre-conditions using a quasi-experimental setup and for now set aside the replicability issue. Nevertheless, the confirmatory value of the study was intended to be maximized by using preregistration and a Bayesian testing approach, setting the evidence criterion for the confirmation of an effect (H_1) to strong evidence ($BF_{10} \geq 10$).

The Central Role of Emotions in Intentional Agency

Because human intentions, ICs' core driving factors, must exert their impact on reality formation via pre-conscious processes, they are presumably shaped by their emotional content (Jakob et al., 2020). At their deepest level, intentional goals can be conceptualized as implicit approach or avoidance orientations, and their accompanying emotions are hope and fear, respectively (Elliot, 2008). For example, in an achievement setting, an individual might strive for a certain end state (e.g., passing an exam). In an approach orientation, the agent seeks a positive result (e.g., success in an exam), and the underlying emotion is hope for success. The actor implicitly expects to be able to succeed. Consequently, the hope that success will be achieved is the basic information encapsulated in this goal.

Conversely, avoidance orientation is characterized by a tendency to avoid failure (e.g., avoid failing an exam), and the underlying emotion is fear. The fear that failure will occur is the basic information encapsulated in this goal. These expectations of behavioral outcomes are pre-conscious and can differ from deliberate goals. Thus, although individuals overall wish to pass the exam, their implicit appraisal of the likely future scenario might contain completely opposite expectations and corresponding emotional consequences. Accordingly, hope will promote success-related realities, and fear will make failure-related realities more likely in this context (Elliot, 2008).

With regard to micro-PK, our assumption was that ICs bias micro-PK outcomes in line with an observer's motivational goals, whereby the affective part of the goal orientation, including implicit expectation, will primarily cause its outcome. This conjecture is based on the emotional transgression model (ETM; Jakob et al., 2020), which proposes that the emotionally coded information of an intention exclusively passes the border between the conscious, intentional mental realm from which an IC starts off and the pre-reality realm of the *unus mundus*. There, it increases the quantum probabilities of the affectively expected states. In this way, observers' expectation encoded in their emotions act like differential self-fulfilling prophecies in the context of micro-PK. During intentional quantum state observations, physical outputs whose content matches the observers' emotional states underlying their intentions will be more likely to emerge.

In the present study, we focused on avoidance-oriented goals driven by the emotional dynamic of fear. We aimed to identify individual differences based on certain personality traits (PTs) related to specific avoidance-oriented patterns. The corresponding fearful outcome expectations should induce a correlation with QRNG-selected stimuli (ICs) that express and confirm those fears.

Personality Traits

To assess individual differences in our participants' fear-based motivational patterns, we used PTs, which reflect specific implicit expectations of behavioral outcomes and tend to cause high emotional activation when triggered (Sachse, 2001). Dysfunctional personality tendencies are described in the DSM taxonomy and can be measured reliably and validly using the self-report VDS-30 questionnaire (Sulz, 2000).

We focused on Cluster C, consisting of three fearful and anxious personality types: dependent, avoidant, and obsessive-compulsive. These three are most common in a normal population. Individuals with dependent PT (DE-PT) exhibit significant fear of separation and strong need to feel secure, welcome, and guilt-free in rela-

tionships. They may neglect the need for autonomy by avoiding making decisions or being alone, always agreeing in conversations, and engaging in even unpleasant activities to satisfy those around them. The key process is the fearful expectation of separation from others.

Individuals with avoidant PT (AV-PT) also exhibit fear of loss of love and an intense need to be welcome and free of guilt. To avoid criticism and rejection by others, they act reluctantly and are distant in company or even move away from conversations and social activities. The crucial process is the fearful expectation of rejection or humiliation by others.

Individuals with obsessive-compulsive PT (OC-PT) appear less anxious than those scoring high on the other two traits described above (Sulz & Müller, 2000) but also exhibit a need to feel welcome and a fear of losing control. To avoid mistakes, they tend to hold onto details and are very persistent. This trait can cause problems in finishing tasks, making decisions, or delegating responsibility. Consequently, the key process is the fear that loss of control will occur.

Hypotheses

A micro-PK task was designed using target sentences as outcomes that mirrored the typical concerns for the Cluster C PTs, in addition to neutral sentences. The micro-PK task comprised three independent consecutive blocks of stimuli, with each block's stimulus material addressing one of the three PTs. We hypothesized that individuals with pronounced PT characteristics (expressed by high scores on the respective scale = target group) would influence the QRNG to produce outcomes that match their implicit expectations within each block. That is, we predicted the presentation of more meaningful PT-specific target stimuli than expected by chance for high scorers. Individuals with low scores on the respective PTs (control group) should show no or weaker deviations from chance.

We tested a specific preregistered hypothesis for each of Cluster C's three PTs:

A) DE-PT: Participants with high scores on the DE-PT performing a micro-PK task will observe more relevant target stimuli that address their fears (e.g., "Will we part in dispute?") than neutral stimuli (e.g., "We see a forest").

B) AV-PT: Participants with high scores on the AV-PT performing a micro-PK task will observe more relevant target stimuli that address their fears (e.g., "What should I say?") than neutral stimuli (e.g., "The shirt is white").

C) OC-PT: Participants with high scores on the OC-PT performing a micro-PK task will observe more relevant target stimuli that address their fears (e.g., "Did I miss a mistake in my work?") than neutral stimuli (e.g., "Digital watches are common").

We hypothesized that no strong evidence for H_1 would be found for the three control subsamples (low PT scorers), and we predicted on a confirmatory level in the preregistration that at least one of the target groups (high scorers) would reach a $BF_{10} \geq 10$ indicating strong Bayesian evidence for H_1 during data collection.

The researchers' a priori belief that the hypotheses would be supported can be classified on a scale from 5 = "strong belief" to 1 = "strong non-belief" as 4 = "moderate belief." This study was implemented and analyzed according to the preregistration uploaded to the Open Science Framework (https://osf.io/gw98t). The data and analyses can be found at the OSF project´s repository (https://osf.io/qxu3s). In addition, some exploratory analyses have been proposed in the preregistration that will not be reported in this article but a summary of the analyses and results can be found at OSF (https://osf.io/qxu3s).

Method

Consent

All research was conducted in accordance with the ethical requirements of the American Psychological Association (APA). The instructions did not reveal the study's purpose but assured anonymization and emphasized the participants' right to withdraw at any time. Voluntary participation was ensured by obtaining informed consent from all participants. The procedure and experiment were approved by the ethical board of the Department of Psychology at the LMU Munich.

Design

We selected a two-group design and conducted a quasi-experiment with high and low scorers on Cluster C's PTs as independent variables. Participants were divided based on their individual scores measured by the VDS-30 (Sulz, 2000), as detailed below. We applied specific micro-PK task blocks for the DE-, AV-, and OC-PT presented in randomized order across participants. Each micro-PK block comprised of 30 on-screen stimulus presentation trials. For each trial, a QRNG selected from be-

tween trait-related and neutral stimuli with a baseline 50/50 probability. The number of trait-related stimuli displayed in each block served as the dependent variable.

For each PT, the sample was divided into control and target subsamples based on individual scores (low vs. high scorers). This was done for each of the three scales measured by the VDS-30. We used the norms for N = 166 non-clinical individuals from Sulz et al. (2009) (DE-PT: M = 0.45, SD = 0.66; AV-PT: M = 0.58, SD = 0.84; OC-PT: M = 0.47, SD = 0.82) and data from our own pre-study for stimulus validation on N = 138 students (102 females, 36 males; age: M = 23.41 years, SD = 7.95; DE-PT: M = 1.01, SD = 0.46; AV-PT: M = 1.11, SD = 0.65; OC-PT: M = 1.13, SD = 0.50) to derive a cut-off value of mean ≥ 1.00 as high for each PT. We deemed this score to adequately represent both the average of the norms provided by Sulz et al. (2009) and our pre-sample's average. Because this criterion's selection implies a certain degree of freedom, the specified cut-off value was preregistered in advance.

Participants

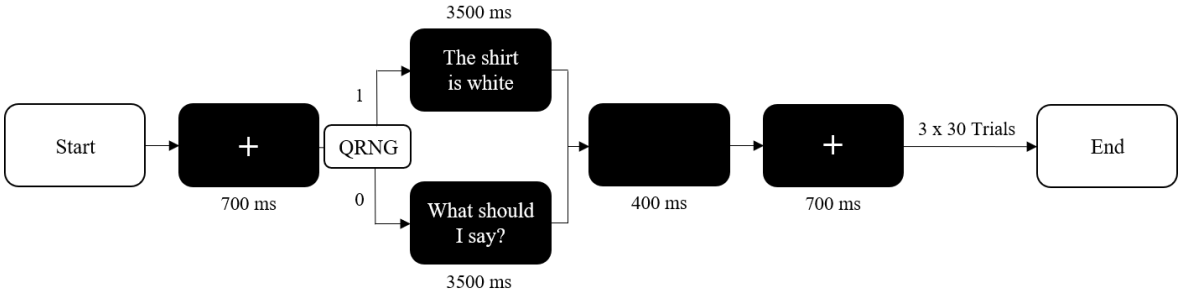
Overall, N = 2,403 participants were tested (56% females, 44% males, 0.2% unspecified; frequency of age categories: 18 – 27 years 62%, 28 – 37 years 17%, 38 – 47 years 8%, 48 – 57 years 9%, > 57 years 3%). Participants were recruited through the department's announcement board, handouts distributed in psychology classes, university Facebook groups, and the online recruiting platforms Mturk and Prolific. Students enrolled in the university's bachelor's psychology classes were awarded credits for participation. Out of all participants, n = 545 (23%) were recruited from Mturk, and n = 120 (4%) were recruited from Prolific and were paid $1.5 or 1.80 GBP respectively for their participation. Inclusion criteria were an age between 18 and 65 years and excellent knowledge of the German language. To ensure that participants paid attention to the stimuli on the screen during the three stimulus presentation blocks, we implemented an "attention check" later on in the course of data collection (at n = 1,105). During each block, a written request to press a "next" button appeared once at random occasions on the screen. Participant's response speed was assessed as an indirect indicator of compliance to the task. The attention check variable data were not part of the preregistration and will be used for exploratory analyses to check for any moderating effects.

Materials

Procedure. The participants were tested in an online study comprising two experiments in fixed order. The present study was the second. Together, both studies took around thirty minutes. Participants first confirmed their agreement on a digital consent form. The present study began by assessing demographic data and the three Cluster C PT scales. Participants subsequently progressed to the experiment, comprising of three blocks, with 30 trials each. The blocks' order was randomized across participants using a quantum based random number generator (QRNG; "Quantis" by idquantique). Each block used experimental stimuli (short sentences) closely semantically related to one of the three Cluster C PTs, in addition to neutral control stimuli. The instruction was to pay full attention to the different sentences being presented on the screen. Participants then passively observed three consecutive series of 30 trials (no manual response required). After each block of 30 trials, the program advised the test participants to maintain focus and press a button to confirm their readiness. The volunteers looked first at a fixation cue (700 msec), then at the stimulus (3500 msec), and finally at a black screen (inter-stimulus interval: 400 msec). Immediately before the stimulus presentation, the QRNG determined whether the next stimulus sentence shown on the screen would be out of the trait-related or neutral stimulus set. This process was repeated 30 times per block (see Figure 1). Three blocks with 30 trials each were presented, each using target sentence stimuli relating to one of the three PTs.

Figure 1

Overview of a Prototypical Trial Sequence Including Presentation Times.



Hardware and Software. This study was conducted online. The stimuli were presented on a black background with a size of 500 x 400 pixels. For QRNG-based stimulus generation, a presentation procedure was programmed in jsPsych (v 5.0.3; de Leeuw, 2015), which translated the QRNG's numerical output into either trait-related stimuli or neutral sentences during each trial. The QRNG used in this study produces quantum

states using photons sent through a semi-conductive mirror-like prism. The photon is equally likely to be deflected in one direction or another, producing a superposition of both states. The photon's location on either path with 50% probability is then transformed into a numerical score such as 0 or 1, depending on the track on which it was located (Quantis transforms 8 such bits into 1 byte). Thus, this procedure follows the structure of a double-slit quantum experiment. This hardware passed all serious tests of randomness, including the DIEHARD and NIST tests, and is among the most effective QRNGs (Turiel, 2007). This QRNG does not use a post-correction procedure during this standard process (confirmed via personal conversation with an idquantique representative). Consequently, a true quantum source for randomness was provided for each experimental trial.

Assessment of Personality Traits. The Cluster C PTs were measured using the VDS-30 (Sulz, 2000), which takes around ten minutes to administer: norms are given in Sulz et al. (2009). The standard sample varies in gender and age and represents 166 healthy individuals and 945 patients from psychotherapy settings. The questionnaire consists of ten items per PT, rated on a four-point scale. Participants indicate how adequately they feel described by the presented items (from 0 = "not" to 3 = "very"). The scales refer to Cluster C's three PTs (DE, AV, and OC). These scores were split into a target group (high scorers) and a control group (low scorers)—based on our preregistered cut-off (mean ≥ 1.00)—and used as independent factors. Internal consistency for the non-clinical sample can be considered good for all scales, with Cronbach's alpha ranging from $\alpha = .72$ to $\alpha = .86$. Retest reliability lies between $r = .70$ (good) and $r = .83$ (very good), with $r = .81$ (very good) on average (Sulz et al., 2009). The questionnaire's construct validity was confirmed by comparing it to the Munich Personality Test (Zerssen, 1993) and the Personality Traits and Disorders Inventory (Kuhl & Kazén, 1997) by Sulz et al. (1998, 2009).

Stimuli. For each of the three PTs, five target stimuli were created. Their content was related to the original items of the VDS-30 questionnaire (Sulz, 2000). The questions were designed to trigger inner doubts in individuals exhibiting the respective PT. For example, for the DE-PT, one experimental sentence was "Is she mad at me?" For each set of five experimental sentences, five neutral control sentences were selected from a greater set created out of neutrally rated words provided by Ben-David et al. (2011) or were self-designed in the same manner.

To ensure that each final set of experimental sentences closely matched the content of the respective PT, we conducted an online pre-study in which for each

PT, ten potential experimental and ten neutral sentences had to be rated on valence for each PT. For each stimulus, participants were asked how pleasant they perceived the sentences to be (1 = "very unpleasant" to 7 = "very pleasant"). The mean rating obtained for each sentence was then correlated with the respective PT of the raters assessed using the VDS-30 (see Tables 1 and 2). For this evaluation, the sample comprised N = 107 undergraduate students (88 females, 19 males; age: M = 22.30 years, SD = 6.83) in total.

As a first step, we collected data from n = 73 participants (58 females, 15 males; age: M = 23.10 years, SD = 7.67), assessing gender, age, and Cluster C's PTs. Sixty sentences—ten target questions and ten neutral sentences for each PT—were presented to each participant in randomized order. In addition, to avoid a general mood shift toward negative emotionality and confusion within participants while rating several sentences of negative valence, we randomly added 15 filler sentences of positive valence. As expected, all neutral sentences were barely rated as unpleasant, and their mean pleasantness scores showed no significant correlation with the individuals' PT scores (for further information see https://osf.io/qxu3s). For the DE- and AV-PT, four of the ten target sentences showed adequate trait-specific correlation (see Table 1). As the study design required at least five targets for each trait, one further target question was created without renewed evaluation for both PTs semantically closely related to one of the four already selected stimuli (DE-PT: "Is he disappointed in me?"; AV-PT: "Do the others see my nervousness?"). None of the given target sentences showed significant correlations with the OC-PT trait and, in many cases, the correlations with the other traits were significant. Thus, none of the 10 target sentences showed sufficiently high specificity for this PT.

We thus created ten new target sentences and tested them in a second pre-study for this PT only with n = 34 additional participants (30 females, 4 males; age: M = 20.59; SD = 4.16). The targets for the other two PTs were not included in this rating for economic reasons. In addition, five filler sentences, each comprising neutral and positive valences, were kept in the rating set. The results of this second pre-rating study showed that four out of ten target sentences correlated sufficiently with the participants' OC-PT scores and were thus selected for use in the micro-PK task (see Table 1). Again, one more missing target question ("Can I simply live from day to day?") was constructed without further evaluation based on an already-validated stimulus.

Table 1

Descriptive Results of the Stimulus Validation Study for the Selected Trait-Related Target Stimuli

Target Stimulus	*M*	*SD*	DE		AV		OC	
			r	*p*	*r*	*p*	*r*	*p*
DE (*n* = 73)								
"Is she mad at me?"	2.32	0.91	-.32**	<.01	-.12	.15	-.07	.29
"Will we part in dispute?"	2.11	0.92	-.26*	.01	-.08	.24	-.03	.39
"Will nobody help me?"	2.52	1.23	-.10	.20	.03	.40	-.10	.21
"Do I get back as much as I give?"	3.47	1.08	-.20*	.04	-.13	.14	-.12	.15
AV (*n* = 73)								
"Will I blush in front of everyone?"	2.78	1.90	-.08	.25	-.23*	.03	-.09	.22
"Can I be myself in front of others?"	4.45	1.56	-.14	.11	-.31*	<.01	-.09	.22
"What should I say?"	2.86	1.92	-.12	.17	-.27*	.01	-.18	.07
"Can I show self-confidence?"	4.38	1.39	-.12	.15	-.27*	.01	.01	.46
OC (*n* = 34)								
"Did I miss a mistake in my work?"	2.26	1.02	-.04,	.41	-.08	.32	-.11	.26
"Have I allowed myself too much free time?"	2.79	1.04	-.41**	<.01	-.40**	<.01	-.31*	.04
"Did I waste my time?"	1.91	0.75	-.06	.38	-.08	.34	-.26	.07
"Did I really work enough?"	2.35	0.92	-.23	.09	-.27	.06	-.32*	.03

Note. *M* = Mean pleasantness; *SD* = Standard Deviation; *r* = correlation (one-tailed). *p < .05, **p < .01.

Finally, fifteen of all neutrally rated sentences—five for each block—were randomly selected for application in the micro-PK tasks (see Table 2).

Table 2

Descriptive Results of the Stimulus Validation Study for the Final Neutral Stimuli

Neutral Stimulus	*M*	*SD*	*r*	*p*
DE (*n* = 73)				
"He is on deck"	4.15	0.72	-.01	.48
"There are magnets on the fridge"	4.29	0.66	-.01	.47
"We see a forest"	5.11	1.16	-.06	.31
"Some tablecloths are in the basket"	4.14	0.51	.01	.46
"Her book is under the bed"	4.04	0.77	-.05	.35
AV (*n* = 73)				
"The shirt is white"	4.27	0.71	.05	.35
"There is a pillow on the sofa"	4.64	0.98	.07	.29
"The spoon is on the table"	4.12	0.47	.08	.24
"The year has twelve months"	4.30	0.78	-.02	.42
"This is a trash can"	3.88	0.58	-.09	.22
OC (*n* = 73)				
"Red pipes are metal"	3.85	0.54	-.07	.28
"The room has many buckets"	3.67	0.69	.04	.38
"The cabinet has four drawers"	3.95	0.92	.01	.47
"The container has a blue lid"	3.99	0.31	-.02	.44
"Digital watches are common"	3.93	0.33	-.06	.30

Note. *M* = Mean pleasantness; *SD* = Standard Deviation; r = Pearson correlation (one-tailed). *$p < .05$, **$p < .01$.

Data Analysis

As recommended by Wagenmakers et al. (2011), the data were analyzed using Bayesian inference techniques with a preregistered strategy (for further details, see Jakob et al., 2020). The BF resembles the relative amount of evidence that the data provide for or against a postulated effect. For the three target groups with high scores on the PTs, each hypothesis was tested individually using a one-sided Bayesian one-sample *t*-test with the number of trait-related targets as the outcome variable, testing for more than 50% probability of occurrence.

For the control subsamples with low PT scores, we performed two-sided Bayesian one-sample *t*-tests against a probability of 50%. As 30 stimuli were presented for each trait, the expectation value under chance was 15. Based on previous findings (Jakob et al., 2020), we a priori decided on a narrow informed prior of δ ~ Cauchy (.05, .05). In the preregistration, an informed prior δ ~ Cauchy (0.5, 0.5) was erroneously mentioned. This was a typo. We address this error in the limitation section of our discussion and provide evidence there that this was a typo. Our a priori defined evidence criterion (stopping rule) was *BF* = 10, classified as strong evidence in any direction (H_0 or H_1). We predicted that a $BF_{10} \geq 10$ in favor of H_1 would occur during data collection for at least one of the micro-PK tasks (i.e., one of the three PTs' high groups). Since frequentist approaches are more common than Bayesian hypothesis testing in psychology, the *p*-scores of frequentist *t*-values are also provided. In addition to the BF stopping rule a maximum n of 1,000 participants was preregistered in case that at this sample size no clear trend towards BF = 10 in the data of at least one of the targets groups could be observed.

Results

Analyses of Total Deviations

As a basis for the preregistered sequential analyses reported later, six Bayesian one-sample *t*-tests were performed within the target and control subsamples for each PT, assessing whether the mean number of the target sentences was greater than 50% (target subsample) or differed from 50% (control subsample)[2]. We expect-

2 Results of the exploratory analyses for the subsample with the attention check (n = 1,059): Mean reaction time over all three tasks *M* = 7.30 s (*SD* = 59.96 s; Min = 0.70 s; Max = 2,884.50 s). Therefore, the average reaction time can be considered acceptable for an online study. The standard deviation of almost

ed to reach strong evidence for H_1 for at least one of the three target subsamples. No strong evidence for H_1 was expected for the control subsamples of individuals.

We first report the analyses of the target groups: A Bayesian one-tailed *t*-test analysis of the mean number of trait-related stimuli for the DE-PT-high scorers (n = 1,400) yielded a final BF_{10} = 10.41 (M = 15.18; SD = 2.68), indicating strong evidence for H_1 (cf.: $t(1399) = 2.51$; $p < .001$). The other two Bayesian one-sample *t*-test analyses showed a final BF_{01} = 2.01 (M = 15.06; SD = 2.71) for the AV-PT-high scorers (n = 1,308) (cf.: $t(1307) = 0.78$; $p = .22$) and a BF_{01} = 2.41 (M = 15.05; SD = 2.74) for the OC-PT-high scorers (n = 1,462) (cf.: $t(1461) = 0.68$; $p = .25$), representing anecdotal evidence for H_0. The sequential Bayesian analyses for these three tests can be seen in Figure 2.

Two-tailed Bayesian one-sample *t*-tests were performed for the three control subsamples. A two-sided approach was adopted since, for the control group, no deviation from 50% in either direction was expected. As postulated, no strong evidence for H1 was found, although the BFs did not confirm H0 either. The control subsample for the AV-PT-low scorers (n = 1,095) showed with BF_{10} = 3.27 (M = 15.17; SD = 2.74), indicating moderate evidence for H_1 (cf.: $t(1094) = 2.08$; $p = .04$). The Bayesian *t*-test analyses for the DE-PT-low scorers (n = 1,003) yielded a BF_{01} = 2.17 (M = 15.06; SD = 2.69), indicating anecdotal evidence for H_0 (cf.: $t(1002) = 0.73$; p =.47), and for the OC-PT-low scorers (n = 941), a BF_{01} = 2.84 (M = 14.88; SD = 2.64) was obtained, indicating anecdotal evidence for H_0 (cf.: $t(940) = -1.37$; $p = .17$).[3]

Further Preregistered Exploratory Analyses

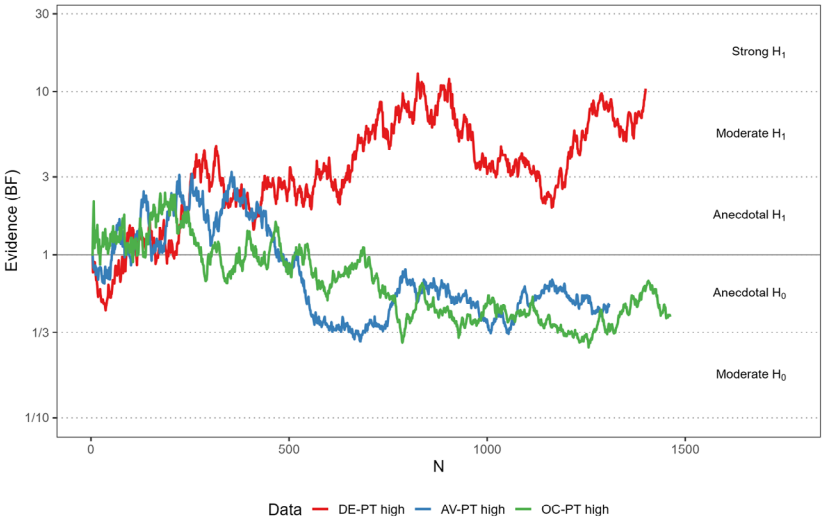
In the preregistration form, several further exploratory analyses regarding the influence of demographic variables (age, gender), the investigation of trait-specificity of the effect, Change of Evidence Analyses, and a test of a regression model with all PTs included at once as predictors were preregistered. These exploratory analyses have been performed but will not be reported in detail here. However, an overview of these results can be found at: https://osf.io/qxu3s.

one minute and the maximum response times of more than three quarters of an hour indicate that the sample included some inattentive participants.

3 Results for the erroneously preregistered prior δ ~ Cauchy (0.5, 0.5): DE-PT-high BF_{01} = 1.31, AV-PT-high BF_{01} = 30.19, OC-PT-high BF_{01} = 35.83; DE-PT-low BF_{01} = 29.04, AV-PT-low BF_{01} = 4.18, OC-PT-low BF_{01} = 16.37.

Figure 2

Sequential Bayesian One-Sample t-Test Analyses



Note. Red line: DE-PT-high, blue line: AV-PT-high, green line: OC-PT-high. Tests are one-tailed.

Discussion

This study aimed to provide empirical evidence of quantum-based psychophysical correlation models by investigating the micro-PK effects produced by QRNG. We tested hypotheses derived from Pauli and Jung's UMM and from von Lucadou's MPI. With reference to the UMM, we considered the activation of an individual's specific pre-conscious mental state while perceiving a QRNG outcome—meaningfully related to the observer's implicit goal orientations—as a type of IC. For specific fear-related motivational patterns, we predicted a non-local correlation between mind and matter, resulting in a higher likelihood of the appearance of a corresponding observation of a quantum outcome, that is, fear-related stimuli presentations. According to the ETM (see also Jakob et al., 2020), this bias can be explained by the individual´s pre-conscious outcome expectations. Only the affective content of an intention can pass the boundary from the conscious to the unconscious pre-reality realm and thus instigate changes in the quantum probabilities of the goal-related pre-material states. This motivational impact should bias the likelihood of a corresponding macroscopic manifestation. The observers' pre-conscious motive states were operationalized as Cluster C's PTs serving as an independent factor. These PTs mirror specific subjective theories about oneself and others and are strongly affectively laden. Moreover, they are grounded in unconscious fear-based expectations of need frustration. Therefore,

the ETM predicts a higher probability that these trait-specific concerns will be realized within the macroscopic material reality—that individuals tend to attract in life that which they are afraid of.

Consequently, we predicted strong evidence for deviations from randomness in a micro-PK task operating with fear-related (and neutral) stimulus presentations within subsamples of individuals expressing high PT levels but not for participants with low scores on these traits. Our primary prediction was confirmed: Within the target group for the DE-PT-high scorers, we found strong evidence ($BF > 10$) for our hypothesis (H_1). Participants scoring high on the DE-PT observed more fear-related target stimuli than chance. These findings can be considered a temporary confirmation of one of the tested hypotheses supporting the conjectures derived from the UMM and the MPI. Analyses of the micro-PK data of any other target and control groups found no strong evidence for the effects. Note that hypothesizing at least one of three tests performed with the target groups to show a result is a relatively weak postulate.

In sum, this study provided confirmatory evidence for a preregistered micro-PK effect for one out of three PTs. This evidence for a micro-PK effect within a subsample of participants expressing high extents of the DE-PT, if further substantiated in future studies, would support the assumption of a systematic influence of an implicitly motivated observer on quantum probabilities (von Lucadou et al., 2007; see also Mensky, 2014; Pradhan, 2012; Penrose & Hameroff, 2011; Stapp, 2007; Stanford, 1990). According to the ETM, individuals who fear losing security in relationships and others' support tend to attract such experiences at a macroscopic level. Thus, the pre-conscious, fearful expectation of abandonment might lead to a self-fulfilling prophecy. Such experiences might further confirm the individuals' expectations of behavioral outcomes, leading to a vicious circle of fear and fear-confirming realities that support the persistence of negative and psychopathologically relevant PTs. These findings corroborate earlier studies, including Jakob et al. (2020) or Maier and Dechamps (2018). It is unclear why no evidence for H_1 was found for the other two target groups. Design characteristics such as testing three different micro-PK effects within separate blocks in one study or sub-optimal target sentences' features might provide potential explanations.

One apparent limitation of this study is that the prior was incorrectly reported in the preregistration form. Instead of an informed prior with $\delta \sim$ Cauchy (.05, .05) used for the analyses reported above, $\delta \sim$ Cauchy (0.5, 0.5) was proposed. We would like to emphasize that this was due to a typo, and in fact, the $\delta \sim$ Cauchy (.05, .05) was meant. From our and others' previous micro-PK research, we knew that the overall effect size

for such effects is small, usually in the range of d = .1 or even lower. This is what we also expected for our study, as mentioned in the preregistration text. A small effect size of d = .1 translates into an informed prior of Cauchy (.05, .05) or an uninformed δ ~ Cauchy (0, .1). In our own micro-PK research, we almost exclusively use these two priors (e.g., Jakob et al., 2020; Dechamps et al., 2021). There was no reason to expect any different effect size for the study here. Given that a small effect size was expected and proposed for the present study, and a Cauchy (0.5, 0.5) would imply a medium effect size and given the fact that a Cauchy (.05, .05) was adequately and successfully applied in our past research, the scores mentioned in the preregistration should be interpreted as a typo. We are aware that the choice of a prior provides a degree of freedom in Bayesian analyses. Therefore, it is of the utmost importance to specify and potentially preregister the prior in advance. Admittedly, such a typo, although ruled out by the additional information presented above, diminishes the empirical strength of the data presented here.

Furthermore, we checked the results regularly during data collection, consistent with the Bayesian approach. However, this procedure might allow observer effects of the experimenters on the course of the study. The data analyst may be considered another observer on a level above the participants. An interim result with a certain degree of evidence for or against one's own hypotheses represents meaningful information for the observer, potentially leading to affectively laden expectations. As the analysis did not include masking, the analyst's unconscious intention could cause another IC, biasing the likelihood of a certain outcome along with the analyst's own belief in micro-PK (see Rabeyron, 2020). Our research group was characterized by a moderate belief, and this experimenter psi effect might have contributed to the results reported here. This additional or alternative explanation should be investigated in the future, perhaps by dispensing with intermediate analyses or deploying masked analyses. It should be mentioned in this context that the maximum N = 1,000 criterion as specified in the preregistration was ignored due to the fact that a clear trend in the data of the high-DP scorers was observed at n = 1,000. This conditional stopping rule was mentioned in the preregistration and is in line with the Bayesian testing approach.

Another limitation of the study was that the attention check was not implemented from the beginning but rather during the course of data collection. The subsample's results showed an acceptable mean reaction time of M = 3.19 s (SD = 7.08 s), indicating that most people were attentive during the online study. However, the variance in the speed of responses suggests that some participants were distracted (e.g., 8% participants showed reaction times larger than 10 seconds). Due to the reduced experimental control within online studies, it cannot be excluded that some participants were less attentive to the study content. Our theory assumes that conscious ob-

servation of stimuli is essential to establishing a meaningful connection between the individual's mind and the quantum process. It may, however, be sufficient if a certain portion of the stimuli is consciously perceived. Nevertheless, for further online studies on micro-PK, we suggest a change in the task so that the selected stimuli should not only be observed attentively without any further action but also rated respectively in their valence. In this way, the participants' continuous attention could also be guaranteed through online experiments.

Finally, the high correlation among Cluster C's three PTs prompts the question of whether a combined index averaged across all three personality scales would deliver a more adequate independent factor. Further exploratory analyses and, if necessary, renewed confirmatory investigations are planned.

This study's objective was to test intentional observer effects on deviations from quantum randomness during the process of measurement to provide evidence concerning quantum-based psychophysical correlation models (e.g., the UMM and the MPI). As predicted in the study's preregistration, in one of the three target groups (DEPT), we found strong evidence (BF_{10} > 10) for our hypothesis. By contrast, we observed no strong evidence for H_1 in the control groups. Overall, our findings align with earlier studies documenting evidence for the influence of unconscious observers' intentions on a QRNG (e.g., Jakob et al., 2020; Maier & Dechamps, 2018). In conclusion, an extension of quantum mechanics allowing intentional observers to contribute to the formation of macroscopic realities is worth considering as an alternative model for mind–matter interaction. This extension should also include ways to specify the degree to which the results can be objectified.

References

Alcock, J. E. (2003). Give the null hypothesis a chance: reasons to remain doubtful about the existence of psi. *Journal of Consciousness Studies*, *10(6-7)*, 29–50.

Atmanspacher, H. (2014). 20th century variants of dual-aspect thinking. *Mind and Matter*, *12(2)*, 245–288.

Atmanspacher, H. (2020). The Pauli–Jung Conjecture and its relatives: A formally augmented outline. *Open Philosophy*, *3*, 527–549. https://doi.org/10.1515/opphil-2020-0138

Atmanspacher, H, & Fach, W. (2013). A structural-phenomenological typology of mind-matter correlations. *Journal of Analytical Psychology, 58(2),* 219–244. https://doi.org/10.1111/1468-5922.12005

Atmanspacher, H., Primas, H., & Wertenschlag-Birkhäuser, E. (Eds.). (2013). *Der Pauli-Jung-Dialog und seine Bedeutung für die moderne Wissenschaft [The Pauli-Jung-Dialogue and its impact on modern science]*. Springer.

Atmanspacher, H., Römer, H., & Walach, H. (2002). Weak quantum theory: Complementarity and entanglement in physics and beyond. *Foundations of Physics*, *32*, 379–406. https://doi.org/10.1023/A:1014809312397

Ben-David, B. M., van Lieshout, P. H., & Leszcz, T. (2011). A resource of validated affective and neutral sentences to assess identification of emotion in spoken language after a brain injury. *Brain Injury*, *25(2)*, 206–220. https://doi.org/10.3109/02699052.2010.536197

Bierman, D. J. (2001). On the nature of anomalous phenomena: Another reality between the world of subjective consciousness and the objective world of physics. *The physical nature of consciousness,* 269–292. https://doi.org/10.1075/aicr.29.12bie

Born, M. (1926). Quantenmechanik der Stoßvorgänge [Quantum mechanics of collisions]. *Zeitschrift für Physik, 38,* 803–827. https://doi.org/10.1007/BF01397184

Bösch, H., Steinkamp, F., & Boller, E. (2006). Examining psychokinesis: the interaction of human intention with random number generators—a meta-analysis. *Psychological Bulletin*, *132(4)*, 497. https://doi.org/10.1037/0033-2909.132.4.497

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

Cardeña, E., Palmer, J., & Marcusson-Clavertz, D. (2015). *Parapsychology*: *A handbook for the 21st century*. McFarland & Co.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2(3)*, 200–219.

Dechamps, M. C., & Maier, M. A. (2019). How smokers change their world and how the world responds: Testing the oscillatory nature of micro-psychokinetic observer effects on addiction-related stimuli. *Journal of Scientific Exploration, 33(3)*, 406–434. https://doi.org/10.31275/2019/1513

Dechamps, M. C., Maier, M. A., Pflitsch, M., & Duggan, M. (2021). Observer dependent biases of quantum randomness: Effect stability and replicability. *Journal of Anomalous Experience and Cognition, 1(1–2).* https://doi.org/10.31156/jaex.23205

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*, 1–12. https://doi.org/10.3758/s13428-014-0458-y

Elliot, A. J. (2008). *Handbook of approach and avoidance motivation.* Taylor & Francis.

Filk, T., & Römer, H. (2011). Generalized Quantum Theory: Overview and latest developments. *Axiomathes, 21,* 211–220. https://doi.org/10.1007/s10516-010-9136-6

Greenstein, G., & Zajonc, A. G. (2006). *The quantum challenge*: *Modern research on the 544 foundations of quantum mechanics* (2nd ed.). Jones and Bartlett Learning.

Jahn, R. G., & Dunne, B. J. (1997). Science of the subjective. *Journal of Scientific Exploration, 11*(2), 201–224.

Jahn, R., Dunne, B., Bradish, G., Dobyns, Y., Lettieri, A., Nelson, R., ...Vaitl, D. (2000). Mind/machine interaction consortium: PortREG replication experiments. *Journal of Scientific Exploration, 14*, 499–555.

Jahn, R. G., Dunne, B. J., & Nelson, R. D. (1987). Engineering anomalies research. *Journal of Scientific Exploration, 1,* 21–50.

Jakob, M. J., Dechamps, M. C., & Maier, M. A. (2020). You attract what you are: The effect of unconscious needs on micro-Psychokinesis. *Journal of Parapsychology*, *84*, 227–253. http://doi.org/10.30891/jopar.2020.02.06

Kuhl, J., & Kazén, M. (1997). *PSSI Persönlichkeits-Stil- und Störungs-Inventar [Personality Style and Disorders Inventory].* Hogrefe.

Maier, M. A., & Dechamps, M. C. (2018). Observer effects on quantum randomness: Testing micro-psychokinetic effects of smokers on addiction-related stimuli. *Journal of Scientific Exploration*, *32(2).* http://doi.org/10.31275/2018.1250

Maier, M. A., Dechamps, M. C., & Rabeyron, T. (2022). Quantum measurement as pragmatic information transfer: Observer effects on (S) objective reality formation. *Journal of Anomalous Experience and Cognition, 2(1),* 16–48. https://doi.org/10.31156/jaex.23535

Mensky, M. B. (2014). Everett interpretation and quantum concept of consciousness. *NeuroQuantology*, *11*. https://doi.org/10.14704/nq.2013.11.1.635

Mossbridge, J., & Radin, D. (2021). Psi performance as a function of demographic and personality factors in smartphone-based tests: Using a "SEARCH" Approach. *Journal of Anomalous Experience and Cognition, 1*(1-2), 78–113. https://doi.org/10.31156/jaex.23419

Penrose, R., & Hameroff, S. (2011). Consciousness in the universe: Neuroscience, quantum spacetime geometry and Orch OR theory. *Journal of Cosmology, 14*, 1–17.

Pradhan, R. K. (2012). Psychophysical interpretation of quantum theory. *arXiv preprintarXiv:1206.6095.* https://doi.org/10.31156/jaex.23419

Rabeyron, T. (2020). Why most research findings about psi are false: The replicability crisis, the psi paradox and the myth of Sisyphus. *Frontiers in Psychology*, *11*, 2468. https://doi.org/10.3389/fpsyg.2020.562992

Radin, D. (2006). *Entangled minds*: *Extrasensory experiences in a quantum reality*. Paraview.

Radin, D. I., & Nelson, R. D. (1989). Evidence for consciousness-related anomalies in random physical systems. *Foundations of Physics*, *19(12)*, 1499–1514. https://doi.org/10.1007/BF00732509

Radin, D., Nelson, R., Dobyns, Y., & Houtkooper, J. (2006). Reexamining psychokinesis: Comment on the Bösch, Steinkamp and Boller (2006) meta-analysis. *Psychological Bulletin, 132(4)*, 529–532. https://psycnet.apa.org/doi/10.1037/0033-2909.132.4.529

Sachse, R. (2001). *Psychologische Psychotherapie der Persönlichkeitsstörungen [Psychotherapy of personality disorders]*. Hogrefe.

Schmidt, H. (1974). Comparison of PK action on two different random number generators. *Journal of Parapsychology, 38(1)*, 47–55.

Stanford, R. G. (1976). A study of motivational arousal and self-concept in psi-mediated instrumental response. *Journal of the American Society for Psychical Research, 70(2)*, 167–178.

Stanford, R. G. (1990). An experimentally testable model for spontaneous psi events: A review of related evidence and concepts from parapsychology and other sciences. In S. Krippner (Ed.), *Advances in parapsychological research* (Vol. 6, pp. 54–161). McFarland.

Stanford, R. G., Zenhausern, T. A., & Dwyer, M. A. (1975). Psychokinesis as psi-mediated instrumental response. *Journal of the American Society for Psychical Research, 69(2)*, 127–133.

Stapp, H. P. (2007). *Mindful universe. Quantum mechanics and the participating observer.* Springer.

Sulz, S. (2000). *Verhaltensdiagnostik und Fallkonzeption [Behavioral diagnostics and case conception].* CIP-Medien.

Sulz, S., Beste, E., Kerber, A. C., Rupp, E., Scheuerer, R., & Schmidt, A. (2009). Neue Beiträge zur Standarddiagnostik in Psychotherapie und Psychiatrie—Validität und Reliabilität der VDS90-Symptomliste und VDS30-Persönlichkeitsskalen [New contributions to standard diagnostics in psychotherapy and psychiatry—validity and reliability of the VDS90 symptom list and VDS30 personality scales]. *Psychotherapie in Psychiatrie, psychotherapeutischer Medizin und klinischer Psychologie*, *14*, 215–234.

Sulz, S., Gräff-Rudolph, U., & Jakob, C. (1998). Persönlichkeit und Persönlichkeitsstörung-Eine empirische Untersuchung der VDS-Persönlichkeitsskalen [Personality and personality disorder-An empirical investigation of the VDS Personality Scales.]. *Psychotherapie*, *3*, 46–56.

Sulz, S. K., & Müller, S. (2000). Bedürfnisse, Angst und Wut als Komponenten der Persönlichkeit [Needs, fear and anger as components of personality]. *Psychotherapie in Psychiatrie, Psychotherapeutischer Medizin und Klinischer Psychologie, 5(1),* 22–37.

Varvoglis, M., & Bancel, P. A. (2015). Micro-psychokinesis. In E. Cardeña, J. Palmer, & D. Marcusson-Clavertz (Eds.), *Parapsychology: A handbook for the 21st century* (pp. 266–281). McFarland.

von Lucadou, W., Römer, H., & Walach, H. (2007). Synchronistic phenomena as entanglement correlations in generalized quantum theory. *Journal of Consciousness Studies, 14*, 50–74.

von Lucadou, W. (2015, July). The correlation-matrix method (CMM): A new light upon the repeatability problem of parapsychology. In *Paper for the 58th Annual Convention of the Parapsychological Association and 39th SPR International Annual Conference, University of Greenwich*.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *Journal of Personality and Social Psychology, 100(3)*, 426–432. http://doi.org/10.1037/a0022790

Wilson, Colin (1985). *Afterlife: An investigation of the evidence for dife after death.* Harrap.

Zerssen, D. v. (1993). Normal and abnormal variants of premorbid personality in functional disorders. Conceptual and methodological issues. *Journal of Personality Disorders, 7(2),* 116–136. https://doi.org/10.1521/pedi.1993.7.2.116

Tester les Effets des Croyances Liées à la Personnalité sur le Micro-Psychokinèse

Marissa-Julia Jakob Moritz C. Dechamps Markus A. Maier

Résumé: Cette étude préliminaire examine les interactions entre l'esprit et la matière en testant les effets de l'observateur sur les résultats du générateur quantique de nombres aléatoires (QRNG), médiés par les intentions implicites. Les traits de personnalité des participants (PT) ont été évalués et des stimuli neutres ou liés à un objectif leur ont été présentés en fonction des résultats du QRNG. Des écarts par rapport au hasard ont été prédits, les participants ayant obtenu un score élevé à l'évaluation des traits de personnalité étant censés observer davantage de phrases liées aux traits de personnalité les concernant. Trois expériences de micropsychokinèse (micro-PK) ont été menées pour les participants présentant des traits de personnalité appartenant au groupe C: dépendants, évitants et obsessionnels-compulsifs. Les résultats ont révélé des preuves solides (facteur de Bayes > 10) d'un effet de micropsychokinèse dans le groupe des personnalités dépendantes, les personnes ayant obtenu un score élevé ayant observé plus de phrases traitant de leurs préoccupations que ce qui était attendu par hasard. Aucune preuve solide n'a été trouvée pour les autres groupes de traits de personnalité, ou chez les personnes ayant obtenu un score faible. Ces résultats suggèrent que l'observation intentionnelle biaise les résultats QRNG liés aux préoccupations implicites des individus, ce qui peut conduire à des prophéties auto-réalisatrices. Les implications de l'étude sont examinées dans le cadre du modèle *Unus Mundus* et du modèle de l'information pragmatique.

Translation into French by Antoine Bioy, Ph. D.

Zur Prüfung der Auswirkungen von persönlichkeitsbezogenen Überzeugungen auf Mikro-PK

Marissa-Julia Jakob Moritz C. Dechamps Markus A. Maier

Zusammenfassung: Diese vorabregistrierte Studie untersucht Geist-Materie-Interaktionen, indem sie Beobachtereffekte auf die Ergebnisse eines Quantenzufallszahlengenerators (QRNG) testet, die durch implizite Absichten vermittelt werden. Die Persönlichkeitsmerkmale (PTs) der Teilnehmer wurden eingeschätzt, und diese wurden mit zielorientierten oder neutralen Stimuli auf der Grundlage der QRNG-Ergebnisse präsentiert. Abweichungen vom Zufall wurden unter der Erwartung vorhergesagt, dass Teilnehmer mit hohen PT-Werten mehr PT-bezogene Sätze beobachteten. Drei Experimente zur Mikro-Psychokinese (Mikro-PK) wurden für die PTs von Cluster C durchgeführt: abhängig, vermeidend und zwanghaft. Die Ergebnisse ergaben eine starke Evidenz (Bayes-Faktor > 10) für einen Mikro-PK-Effekt in der Gruppe der abhängigen PTs, wobei Personen mit hoher Punktzahl mehr Sätze beobachteten, die ihre Probleme ansprachen, als zufällig erwartet. Für die anderen PT-Gruppen oder für die Personen mit niedrigen Punktzahlen wurde kein starker Nachweis gefunden. Diese Ergebnisse deuten darauf hin, dass die absichtliche Beobachtung die Ergebnisse von QRNG in Bezug auf die impliziten Bedenken von Personen verzerrt, was möglicherweise zu sich selbst erfüllenden Prophezei-

ungen führt. Die Implikationen der Studie werden im Rahmen des *Unus Mundus*-Modells und des Modells der Pragmatischen Information diskutiert.

Translation into German by Eberhard Bauer, Ph. D.

Testando os Efeitos das Crenças Relacionadas à Personalidade na Micro-PK

Marissa-Julia Jakob Moritz C. Dechamps Markus A. Maier

Resumo. Este estudo pré-registrado investiga as interações mente-matéria testando os efeitos do observador sobre os resultados do gerador quântico de números aleatórios (QRNG em inglês) mediados por intenções implícitas relacionadas à personalidade. Os traços de personalidade (PTs em inglês) dos participantes foram avaliados, e apresentados com estímulos relacionados a metas ou neutros, com base nos resultados do QRNG. Desvios em relação ao modelo foram previstos, com pontuadores altos de PT esperados, para se observar mais sentenças relacionadas a PT. Três experimentos de micro-psicocinese (micro-PK) foram conduzidos para os PTs do Cluster C: dependentes, evitativos e obsessivo-compulsivos. Os resultados revelaram forte evidência (Fator de Bayes > 10) para um efeito de micro-PK no grupo de PT dependentes, com pontuadores altos observando mais frases abordando suas preocupações do que o esperado ao acaso. Não foram encontradas evidências fortes para os outros grupos de PT ou pontuadores baixos. Essas descobertas sugerem que a observação intencional influencia os resultados do QRNG relacionados às preocupações implícitas dos indivíduos, potencialmente levando a profecias autorrealizáveis. As implicações do estudo são discutidas dentro do modelo *Unus Mundus* e do Modelo de Informação Pragmática.

Translation into Portuguese by Antônio Lima, Ph. D.

Una Evaluación de los Efectos de las Creencias Relacionadas con la Personalidad en Micro-PK

Marissa-Julia Jakob Moritz C. Dechamps Markus A. Maier

Resumen. Este estudio pre-registrado investigó las interacciones mente-materia evaluando los efectos del observador en los resultados de un generador cuántico de números aleatorios (QRNG), mediados por intenciones implícitas. Evaluamos los rasgos de personalidad de los participantes y les presentamos estímulos relacionados a objetivos o neutros basados en los resultados del QRNG. Predijimos desviaciones con respecto al azar, y esperabamos que los participantes con altas puntuaciones en PT observarían más frases relacionadas con PT. Realizamos tres experimentos de micropsicoquinesis (micro-PK) para los rasgos de personalidad del Grupo C: dependiente, evitativo, y obsesivo-compulsivo. Los resultados revelaron pruebas

sólidas (Factor de Bayes > 10) de un efecto micro-PK en el grupo de TP dependientes, en el que las personas con puntuaciones altas observaron más frases relacionadas con sus preocupaciones de lo que cabría esperar al azar. No encontramos pruebas sólidas para los otros grupos de TP ni para los de puntuación baja. Estos resultados sugieren que la observación intencionada sesga los resultados del QRNG relacionados con las preocupaciones implícitas de los individuos, lo que puede conducir a profecías autocumplidas. Discutimos las implicaciones del estudio dentro del modelo *Unus Mundus* y el Modelo de Información Pragmática.

Translation into Spanish by Etzel Cardeña, Ph. D.

Comment on Jakob, Deschamps & Maier: How to Read a Paper[1]

Peter A. Bancel

Institut Métapsychique International

Abstract: In this issue of JAEX, Jakob et al. (2024) report results from an experiment to test whether an implicit psychokinetic effect related to unconscious intention can alter the outcome of quantum events. The experiment is well conceived and executed. However, the claim of evidence to support the experimental hypothesis is undermined by errors of analysis and of omission. Some are not obvious and require a careful reading to discern. I explain the nature of the mistakes, where and how they arise, and the consequences for the inferences the authors' draw. I emphasize that a helpful way to detect such problems is by following the chain of inference in experimental reports. Editors, reviewers, and particularly authors can protect the integrity of the scientific literature by reading carefully in this way.

Keywords: Bayesian *t* test, optional stopping, multiple testing, psychokinesis, psi, mind-matter interaction, replication crisis.

Highlights

- An interesting paper by Jakob et al. (2024) reports supporting evidence for a micro-PK effect related to unintentional personality traits.
- The methods and procedures have errors and omissions that undermine the paper's conclusions.
- When reading scientific papers, it is helpful to examine carefully the chain of inferences that lead to research conclusions.

1 Address correspondence to: Peter A. Bancel, Ph. D., **contact@imiresearch.fr**

https://doi.org/10.31156/jaex.25420

I thank the editor-in-chief of JAEX for the opportunity to comment on the preceding paper, *Investigating the Effect of Cluster C Personality Styles on Micro-Psychokinesis*, for which I was a reviewer. In their paper, Jakob et al. (henceforth, *Jakob*) claim strong evidence for an unconscious (i.e., unintentional) psychokinesis effect (PK) related to personality traits. The claim of statistical evidence, and subsequent inferences to PK, are based on a Bayesian analysis of experimental data that I believe is misleading. I explain why the statistical inferences the authors make are not well-supported by their analyses. Along with this critique, I discuss the authors' contribution in the context of the so-called "replication crisis" in science to emphasize the importance and challenge of getting things right in the primary literature. This broader point is relevant for researchers, journal editors, and reviewers. It is a problem that although largely recognized and acknowledged, requires ongoing effort to address. I feel it is appropriate to address the issue in the pages of JAEX since these concerns, and the evolving response of the reform movement in statistics, are regularly highlighted in specialty journals in other fields of research.

I start with due praise for *Jakob*. The authors identify an important question in psi research, argue carefully for its connection to theory, conceive and execute a well-planned experiment, and are attentive to current best practices. In addition, the work is part of an ongoing research program at Ludwig-Maximilians-Universität (LMU) and it is laudable to make a sustained effort in a line of psi research. The paper is worth a read.

The research question they address is whether psi effects, and PK in particular, are associated with unconscious intention. This is important for its connection to theoretical models, such as those derived from the Pauli-Jung *Unus Mundus* model of synchronicity and others (see *Jakob* and references therein), but also because it investigates psi effects that potentially circumvent the inherent variability in psi agent effort. Our understanding of the circumstances of such implicit psi is cloudy, to say the least. The prospect that these effects could be reliable brightened with the now famous experiments of Bem (2011) and a subsequent meta-analysis of replications (Bem et al., 2016). Those hopes have dimmed since recent high-powered, multi-lab confirmatory studies have failed to find effects (Kekecs et al., 2023; Muhmenthaler et al., 2022; Schlitz et al., 2021). *Jakob* innovate around these questions and, in doing so, provide a complementary approach to the attempts at straightforward replication.

The design and execution of *Jakob* benefits from applying up-to-date practices: empirical verification of their personality trait metric; implementation of an online design to access a diverse participant base; use of Bayesian statistics that allow stopping rules to conserve resources; and, preregistration, which is now standard practice in parapsychology as elsewhere. Despite these advantages, methodological and pro-

cedural mistakes undermine their conclusions. Among these are: failure to account for multiple analyses; stipulating a primary hypothesis with an uncontrolled type I error rate; and not following preregistered procedure.

The problems in *Jakob* arise in part from the challenges to apply Bayesian methods in data analysis. Statistical analysis is tricky. Statisticians have lamented the pitfalls of Null Hypothesis Significance Testing (NHST) and (mis)use of *p*-values for a long time. Not surprisingly, these issues play a central role in the current "replication crisis" (Ioannidis, 2005) and they remain challenging to remedy (see for example, Wasserstein et al., 2019, and articles in the same issue of *The American Statistician*). It is thus natural that there is a learning curve for less familiar Bayesian techniques as well (for an accessible discussion see Schmalz et al., 2023). I have learned things in preparing this comment and appreciate, not for the first time, the value in reviewing statistical subtleties, so the comment aims to be helpful and tutorial. The purpose is to identify some pitfalls that may not at first be obvious, and thus aid the reader in a better assessment of statistical inferences.

To summarize, *Jakob* tested 2,403 participants with three consecutive psi tasks and divided the participants in groups with high or low tendencies relative to each of three personality traits (PTs; as determined from responses to a questionnaire). The separate psi tasks are PT-specific and each consists of 30 trials with random stimuli (sequentially presented word phrases) that are either neutral or relevant for the PT. An equi-probable quantum-based random bit generator (qRNG) selected whether neutral or relevant stimuli were presented. Under the Null hypothesis of no psi effect (H_0) the expected average count of relevant stimuli was thus 15 for the 30-trial tasks. The theoretical proposal is that an unconscious affinity for PT-relevant stimuli will bias the qRNG outcomes, thereby associating reality emergence (qRNG outcomes) with hidden intentional states (the PTs) via a PK effect.

The data were divided into 6 sets: a high or low PT group for each of the 3 PT-specific tasks. The hypotheses stated that data from the high PT groups would yield strong evidence of excess relevant stimuli (H1), whereas data from low PT groups would not. The criterion for evidence is a Bayes Factor (BF) > 10. A BF threshold of 10 is conventionally considered strong evidence in Bayesian analysis (Schönbrodt & Wagenmakers, 2018). Groups with low PT scores are controls and are hypothesized to yield BFs < 10, relative to either H0 or H1. Specifically, the registered primary hypothesis was that *one or more* of the 3 high PT groups would surpass the evidence criterion BF > 10, in either direction for or against the hypotheses. *Jakob* found one high PT dataset (labeled the

dependent or DE-group) with BF > 10 favoring H1, and that the 3 control datasets all had BF < 10.

Jakob conclude that the primary hypothesis was confirmed and that the hypothesis of no strong evidence for PK in the control groups (with low PT scores) was supported. The experiment is presented as confirmatory, which strengthens the statistical inference because all procedures are specified in a preregistration. An important feature is *Jakob*'s use of a Bayesian stopping rule to halt data accumulation. The preregistration states: "we will add participants until a) our desired level of evidence is reached at least for one of the three dependent variables (BF_{10} = 10, BF01 = 1/10) or b) or till a maximum number of participants (*N* = 1,000) without any trend towards 10 has been reached." To sum up, the paper concludes: "… this study provided confirmatory evidence for a pre-registered micro-PK effect for one out of three PTs."; "[for that] target group, strong evidence (BF_{10} > 10) was found. By contrast, no strong evidence for H1 was observed within the control groups."

To assess *Jakob*'s conclusions, it is useful to recall how experiments are done. Often, one begins with a chain of inference whereby a theoretical notion is translated into an experimental hypothesis that distinguishes between effects. An experiment is then designed and executed to produce data that can be tested according to some analytical procedure. The experiment only concludes when a reverse chain of inference leads from the data back to theory. The inferential return chain starts with a statistical inference about the data such as a claim that the data do not agree with H0. The inference may be hypothetico-deductive or Bayesian, or a mixture of these as is the case in *Jakob*. Further inferences in the chain tend to be qualitative, providing interpretative context for the statistical inference. They include experimental inferences to link the statistical claim to an effect, and theoretical inferences that link the effect to theory. In the case of *Jakob*, experimental inferences might involve reasoning that a statistical result is due to PK and not, say, instrumental error. Theoretical inferences might involve reasoning that the PK effect is implicit and associated with the participant PTs, but not due to an intention of the experimenter. Experiments can be confirmatory, like *Jakob*, or exploratory. The difference is the degree to which pre-stated procedures and inferences are followed, with confirmatory experiments adding force to statistical inferences and exploratory studies favoring discovery.

A number of problems undermine the chain of inference to *Jakob*'s conclusions. Most have to do with the statistical inference and it is instructive to see how these arise. They can be variously traced to statistical misunderstandings, procedural flaws, or definitional ambiguities. Some are clear-cut, others less so, but all weaken *Jakob*'s statistical claims and by extension the chain of inference to their conclusions. The point here is not to be overly critical, but rather to help researchers who do experiments (which is not easy!), and readers who follow the chain of inference in a paper. At issue in *Jakob* are: 1) Claims of evidence that are not adjusted for multiple testing; 2) lack of context for Bayesian *t* tests. 3) an undetermined type I error rate of the primary hypothesis; 4) failure to follow the pre-registered procedure; 5) inadequate treatment of the control group analysis.

In the interest of readability, I first give a succinct statement of the objections, and then follow with further comments.

Multiple Testing Invalidates the Inference to "Strong" Evidence.

Jakob's claim of strong evidence is overstated because the Bayesian *t* tests are not independent and therefore require adjustments for multiple testing. This is a compelling criticism of the paper because *Jakob* relies on the claim of strong evidence to support inference to a PK effect. Typically, test statistics are adjusted to control for the false positive error rate (FPE) when multiple tests are performed. Methods such as Bonferroni or False Discovery Rate corrections on *p*-values adjust for the fact that the probability of a type I error increases with the number of tests. However, opinions vary on how to apply corrections because frequentist theory does not determine when and how multiple testing should be treated. Within Bayesian statistics the situation is clear: no adjustment is needed if tests are independent (Sjölander & Vansteelandt, 2019).

To see this, note that the BF is proportional to the ratio of posterior probabilities of H0 and H1. From this, we can see how test independence affects the posterior probability for H1 when multiple tests are made. Imagine an experiment that tests two hypotheses, HA and HB, with data sets XA and XB. An example would be considering two of *Jakob*'s PT groups under their respective H1s. The posterior probability, P(HA, HB | XA, XB) is the (Bayesian) quantity of interest: it is the probability that the hypotheses are true given all the data that were collected. The posterior probability can be expanded using Bayes's rule:

$$P\left(H_A, H_B | X_A, X_B\right) = \frac{P\left(X_A, X_B | H_A, H_B\right) P\left(H_A, H_B\right)}{P\left(X_A, X_B\right)}$$

If the hypotheses and datasets are independent, the probabilities factorize:

$$P\left(H_A, H_B | X_A, X_B\right) = \frac{P\left(X_A | H_A\right) P\left(H_A\right) P\left(X_B | H_B\right) P\left(H_B\right)}{P\left(X_A\right) P\left(X_B\right)}$$

Summing over hypothesis parameters B (taking the marginal),

$$\int_B P\left(H_A, H_B | X_A, X_B\right) = \frac{P\left(X_A | H_A\right) P\left(H_A\right)}{P\left(X_A\right)} \int_B \frac{P\left(X_B | H_B\right) P\left(H_B\right)}{P\left(X_B\right)}$$

$$P\left(H_A | X_A, X_B\right) = \frac{P\left(X_A | H_A\right) P\left(H_A\right)}{P\left(X_A\right)}$$

$$P\left(H_A | X_A, X_B\right) = P\left(H_A | X_A\right)$$

where the last step is just from applying Bayes' rule. So, with independent tests, the probability of HA does not change when there is a second test. A consequence is that no adjustment for multiple testing is needed for the posterior probabilities or BFs (the proportionality factor between ratios of prior and posterior probabilities, given the data). This conclusion does not apply when tests are dependent, which holds if the hypotheses and/or the data sets are associated. In that case, the probabilities do not factorize and the BF of a test on A cannot be considered independently of data set B. In *Jakob*, it is easy to see that both the data sets and the hypotheses are associated. For hypotheses, *any plausible* theoretical association implies dependence, and *Jakob* is clear that the three personality traits they examine are associated. It is precisely because of the common characteristic of fear reactions that the PTs are chosen for study. The primary hypothesis that *one or more* of the high-PT groups should exhibit an effect, and the choice of identical prior probabilities, together express the assumed association implicitly.

The data sets for PT groups are also dependent. Each participant contributed to all three tests and the participant groupings into high and low PT are highly correlated (correlations of about .65). *Jakob* notes the group correlations, but overlooks that the data dependencies imply that BFs cannot be reported separately without taking the multiple tests into account. To see this better, consider the participant groups as treatments on the qRNG outcomes. Because the treatments are correlated, the data sets, in general, will contain dependencies. In Bayesian language one would say that

learning something about one high-PT group (from its data) provides information on the others (and their data).

All this is to say that *Jakob* make an invalid inference of strong evidence for an effect by considering one selected BF as an independent statistic. Instead, the BFs should be adjusted for the multiple tests on dependent hypotheses and data. There are ways to do this with hierarchical Bayesian models that incorporate hypothesis associations into higher dimensional priors (Berry & Hochberg, 1999; Gelman et al., 2012). In hierarchical models, adjustments for multiple tests are built in and model outputs can be taken at face value from a Bayesian perspective. But the models are quite involved and still do not address the problem of the dependencies in *Jakob*'s data. An alternative is a hybrid approach of frequentist and Bayesian reasoning (Sjölander & Vansteelandt, 2019). What remains is that the assertion that BF > 10 is strong evidence cannot be used, as such, to support claims further up the chain of inference.

Lack of Context for Bayesian t-Tests

It is important for Bayesian analysis to include sensitivity checks when reporting BFs (Kruschke, 2021). This is particularly important when using informed priors and for small effects, both of which pertain to *Jakob*. For instance, *Jakob* use a Cauchy(0.05, 0.05) prior that yields a final BF=10.41 for the DE-high group. Tight, informed priors such as this can be sensitive to parameter changes or to slight relaxations of the constraint of a point estimate for H0. For example, with a modified prior of Cauchy(0.1, 0.1) the BF reduces by nearly half to BF=5.78. When calculated BFs are not robust, authors should either qualify their inferences accordingly, or explain a priori why a tight choice of prior is preferred over neighboring ones. *Jakob* choose their prior because it was used in some previous papers, but that is not a reason to ignore the BF's sensitivity to parameter choices. The issue has special relevance in psi research where experimenter psi has been shown to influence statistical outcomes in cases where parameter selection is an avenue for psi effects (Bancel, 2017). *Jakob* mention the possibility of experimenter PK as a possible caveat to their conclusions, but parameter choice is another way experimenter psi could influence the results. Such influences aside, the sensitivity of the BFs to parameter choices is a factor of concern for statistical inference and this is not considered in *Jakob*.

Large, Unreported Type I Error of the Primary Hypothesis

Jakob's primary hypothesis is a test for BF > 10 among the 3 high PT groups. With optional stopping, the false positive error (FPE) rate of the test is roughly .15 (15%) which is far higher than the typical standard of .05 for confirmatory studies. *Jakob*'s large type I error is not reported or discussed. *Jakob* mentions in the Discussion that the primary hypothesis (at least one BF > 10) is "a relatively weak postulate", but they neither explain the weakness, nor present a FPE estimate, nor do they consider further the impact of FPE on the inference to "confirmatory evidence for a preregistered micro-PK effect." Both Bayesian and frequentist methods are subject to type I errors and FPE must be considered when making inferences about a test. When *N* is fixed, the criterion BF > 10 does a good job of controlling type I error, with a FPE of about .006 when *N* is comparable to that of the data in *Jakob*. However, with fully optional stopping, the FPE increases to ~ .05 for a single dependent variable and to ~ .15 when considering one "success" in three. If the stopping checks are performed only once per day, as *Jakob* specifies in preregistration, the FPE decreases slightly to perhaps ~.14. But we don't know the real check frequency and so the true FPE remains high and undetermined.

Although it is not statistically *incorrect* to use a test with high FPE, it is not advised, and whatever the FPE may be, it is necessary to make inferences with FPE estimates in mind. Further, the goal of any paper should be to provide key information transparently to allow readers to assess the inferential chain. To omit a discussion of FPE when it is large is an instance of inadequate reporting. Last, a strong recommendation is to include FPE analysis into the experimental design at the start. There are ways to adequately control FPE adapted to different scientific goals and resource limitations. These should be carefully considered when planning analysis of Bayes factors and setting stopping rules (Schönbrodt & Wagenmakers, 2018). *Jakob* have needlessly increased the FPE rate by checking the BF too early and frequently. Examples of some alternatives and details of the FPE estimates stated above are given in the Appendix.

Failure to Clarify and Follow Pre-Registered Procedures

Jakob ignores the pre-registered stopping rule. Data collection should have stopped at $N \approx 820$ when the high DE-PT group first yielded BF > 10 (in Figure 2 of their paper). This deviation from the preregistration adds an undetermined degree of experimenter freedom to the protocol. Part of the problem is due to a lack of clarity in the stopping rule. The rule states that data collection halts either at BF > 10, or when N =1,000 *unless a trend towards 10 has been reached*. The rule is ambiguous on several counts. Does the stop at BF > 10 apply only to high PT groups? If BF < 10, does the stop

occur when any group attains N = 1,000, or all groups, or high PT groups only? What is meant by "a trend towards 10" for continuing past N = 1000, and how is the trend determined? What then is the stopping rule if collection continues after N = 1,000? These ambiguities allow experimenter degrees of freedom to enter and it is difficult to assess their impact on inferences. Removing such hidden freedoms is precisely the reason for preregistration. *Jakob* undermines the value of preregistration by failing to state the rule clearly.

Ambiguities aside, the rule to stop at the first instance of BF > 10 is clear enough, but it was ignored. Figure 2 in *Jakob* shows BF > 10 at $N \approx 820$ for the DE-PT group, yet data collection continued. In an early version of the paper, collection in fact stopped when all 6 groups passed N = 1000. The result was BF=7.8 for the DE-PT group. Then, in the final version of the paper, data collection continued further until BF again reached 10 for the DE-PT group at N =1,400. The apparent justification for continuing was that a "trend towards 10 was observed" when BF was at 7.8 and N = 1,260. We are left to puzzle how this trend toward 10 was determined (there is in fact a trend away from 10 in Figure 2) and why the stop was not done at the earlier crossing of BF > 10 in the first place.

An argument can be made that these details are unimportant because the Bayes factor allows stopping for *any* reason as long as all the data are used to calculate BF. This is true when the objective is to report a BF for a single independent dataset. In that case, the BF supplies the relative update from prior to posterior probability. However, *Jakob* don't use the BF only in this way. The 3 BFs were used as binary variables input to a frequentist test for their primary hypothesis which predicts that one or more of the BFs > 10. A consequence of the failure to follow registration procedures is that the assessment of type I error for the primary hypothesis is muddled further. It is not even clear how to model the decision process that actually occurred, should we want to estimate the impact of decisions on the rate of FPE.

It is essential to follow preregistered procedures when claiming results are confirmatory. Should lapses occur, reports should at least explain the full history of data collection and discuss implications for inference. When explanations are post-hoc, confirmatory claims need to be qualified accordingly.

Inadequate Control Group Analysis

Jakob's claim that control groups show no strong evidence for an effect is misleading. This is because 1) no direct comparison between the target and control groups

was made and, 2) different tests were used for the groups.

A common procedural error in psychology (Makin & Orban de Xivry, 2019) and elsewhere (Nieuwenhuis et al., 2011) is to draw inferences from two groups without performing a direct test of difference. *Jakob* make this mistake and aggravate it by reporting a one-tailed test for target groups and a two-tailed test for control groups. No direct mean difference test is given. Reporting the difference between two separate tests is not equivalent to a direct difference test between groups, and this latter statistic is the relevant one for inferring that the groups differ. A direct test is needed because it is conceivable to have two groups for which a difference test is inconclusive, and yet individual tests alternately support H1 and H0, and thus a seeming difference (Palfi & Dienes, 2020). The proper inference is based on the direct test. Furthermore, *Jakob*'s two-tailed test on control data penalizes these BFs relative to the target groups.

A Reassessment of Jakob's Statistical Inference

We can reassess *Jakob*'s claims in light of the comments above. A fair paraphrasing of *Jakob*'s statistical inferences is: The primary hypothesis was substantiated by confirmatory evidence for one group; that group yielded strong evidence in favor of H1. Other groups, including all controls, did not provide evidence at that level. The hypothesis confirmation and the strong statistical evidence for the DE group supports inference to a PK effect.

A reassessment of these conclusions might read: Although our primary hypothesis was satisfied *sensu stricto*, it does not provide adequate evidential support for an effect because of a high false positive error rate. The FPE exceeds ~15% but a precise estimate is hampered by procedures that either deviated from preregistration or were stated unclearly. The procedural uncertainties also weaken claims that the results are confirmatory in the conventional sense. Furthermore, evidence for the group with BF > 10 must be assessed downwards because of test dependencies. A clearer statement of the strength of evidence unfortunately eludes us due to dependencies in the hypotheses and data groups. With our statistical results weakened and uncertain, we cannot claim adequate evidential support for hypothesis confirmation or for an inference to a PK effect.

The critique arrives at a statistical inference that is contrary to *Jakob*. The reassessment is not about adjusting *Jakob*'s criteria: if strong Bayesian evidence for an effect were provided it would be fine to present an argument for PK, for that would be the next step in the chain of inference. Rather, we find that upon careful reading, *Jakob*'s own criteria are either not met (evidence is not strong because of multiple testing),

or are inadequate (the primary 1-in-3 hypothesis has an unacceptable FPE rate and control comparisons are lacking). A consequence is that the chain of inference comes to a halt, and a precise statement of results cannot be made.

Discussion

How to Read a Paper

An experimental paper presents us with a loop of reasoning that goes from theory to data, and from data back to theory. This is mirrored in the typical structure of a journal article, with its sections of Introduction, Methods, Results, Discussion, and Conclusions. The loop of reasoning is notably Bayesian: we update our prior beliefs and knowledge of theory, given the data. However, the updating is not merely analytical as would be the case, say, in calculating a Bayes factor. For readers, it involves scrutinizing the strength – and our own comprehension – of each link in the chain of reasoning. Although perhaps obvious, it is useful to recall this approach when preparing to read a paper carefully, especially for reviewers and editors and of course for the authors themselves.

Applying a chain of reasoning examination to *Jakob* explains why the beginning of the paper is so satisfying: the link from theory to experiment to method is particularly well done and careful. We can also pinpoint where things go amiss. This happens at the stage of hypothesis formulation which mixes up confirmatory and exploratory intentions and, as emphasized above, fails to identify hypothesis dependencies. In hindsight. The confusion could have been avoided by simply comparing high and low PT data without the breakdown into separate PT groups, as this would obviate the dependencies of hypotheses. The data dependencies could be resolved as well by limiting the test to those participants with uniformly high or low scores on all three PT scales. Doing so would remove the correlation between test groups. Formulating the hypothesis this way allows a clean confirmatory test, and subsequent tests of individual PT groups then could be registered as exploratory. But, formulating a clean hypothesis this way does not give evidence for an effect. It yields a high-low difference t score of 0.84 with an estimated BF of about 1 under the *Jakob* prior. For all high PT data alone BF10 = 1.32 (M = 15.083; SD = 2.73) and $t(2699) = 1.58$; $p = .06$.

Aside from exposing logical weaknesses, the chain of reasoning approach also helps to identify errors of omission. Careful validation of the steps in statistical reasoning can expose gaps in information that is needed to move to the next link in the chain. The failure to give the FPE rate, or omitting a sensitivity check of the BF priors, are ex-

amples. Of course, it requires some statistical sophistication to spot these omissions, and it is the role of editors, reviewers, and authors to collaborate well and thoroughly so that such oversights are avoided.

In fact, considering peer review as a collaboration as opposed to, say, an adversarial process, is not a bad way to think about it. It highlights the mutual goal of all concerned, which is to assure the integrity of the primary literature. Therein lies the crux of the current crisis in science: the integrity of the primary literature has been compromised and once that happens, it is extremely complicated to remedy. This is true across the sciences, but perhaps even more so for smaller fields of research such as parapsychology with its limited readership. Mistaken conclusions that get published can persist as cited literature and thereby hinder progress. An example in *Jakob* is a citation (Jahn et al., 2000) to support evidence of decline effects. The citation's use is mistaken and has been refuted in popular, academic and peer-reviewed literature (Ehrlich, 2021; Varvoglis & Bancel, 2015, 2016). Nevertheless, it is cited erroneously in psi research to this day. *Jakob* risks this propagation of mistaken inferences with their present work. In their Conclusions, they state "Overall, our findings align with earlier studies documenting evidence for the influence of unconscious observers' intentions on a QRNG," citing previous papers by the group. But the apparent alignment is mistaken, as this critique has explained in some detail. So better to have an ounce of prevention than a pound of cure.

That being said, useful prescriptions such as preregistration, adopting new statistical methods, or commitments to publish null results are not sufficient. Papers need to transparently include full information about the chain of reasoning that leads to a conclusion (Wagenmakers et al., 2021). This allows readers to assess for themselves a paper's conclusions and also learn from the authors' experiences and missteps (for an example of a nicely done paper from the literature on meta-science, see Scheel et al., 2021). For those involved in the particular collaboration that is the process of publication, careful reading helps to maintain as reliable a scientific record as possible, and by extension, that the literature may provide us all with a good read.

References

Bancel, P. A. (2017). Searching for global consciousness: A 17-year exploration. *EXPLORE*, *13*(2), 94–101. https://doi.org/10.1016/j.explore.2016.12.003

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Bem, D., Tressoldi, P., Rabeyron, T., & Duggan, M. (2016). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events. *F1000Research*, *4*, 1188.

https://doi.org/10.12688/f1000research.7177.2

Berry, D. A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, *82*(1), 215–227. https://doi.org/10.1016/S0378-3758(99)00044-0

Ehrlich, R. (2021). *Eight preposterous propositions*. Princeton University Press. https://doi.org/10.1515/9780691228402

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jahn, R., Dunne, B., Bradish, G., Dobyns, Y., Lettieri, A., Nelson, R., …. & Walter, B. (2000). Mind/ Machine interaction consortium: portREG replication experiments. *Journal of Scientific Exploration*, *14*(4), 499–555.

Kekecs, Z., Palfi, B., Szaszi, B., Szecsi, P., Zrubka, M., Kovacs, M., … Aczel, B. (2023). Raising the value of research studies in psychological science by increasing the credibility of research reports: The transparent Psi project. *Royal Society Open Science*, *10*(2), 191375. https://doi.org/10.1098/rsos.191375

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour*, *5*(10), Article 10. https://doi.org/10.1038/s41562-021-01177-7

Makin, T. R., & Orban de Xivry, J.-J. (2019). Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife*, *8*, e48175. https://doi.org/10.7554/eLife.48175

Muhmenthaler, M. C., Dubravac, M., & Meier, B. (2022). The future failed: No evidence for precognition in a large scale replication attempt of Bem (2011). *Psychology of Consciousness: Theory, Research, and Practice*. https://doi.org/10.1037/cns0000342

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), Article 9. https://doi.org/10.1038/nn.2886

Palfi, B., & Dienes, Z. (2020). Why Bayesian "evidence for H1" in one condition and Bayesian "evidence for H0" in another condition does not mean good-enough Bayesian evidence for a difference between the conditions. *Advances in Methods and Practices in Psychological Science*, *3*(3), 300–308. https://doi.org/10.1177/2515245920913019

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007467. https://doi.org/10.1177/25152459211007467

Schlitz, M., Bem, D., Marcusson-Clavertz, D., Cardeña, E., Lyke, J., Grover, R., … & Delorme, A. (2021). Two replication studies of a time-reversed (psi) priming task and the role of expectancy in reaction times. *Journal of Scientific Exploration*, *35*(1), Article 1. https://doi.org/10.31275/20211903

Schmalz, X., Biurrun Manresa, J., & Zhang, L. (2023). What is a Bayes factor? *Psychological Meth-*

ods, *28*(3), 705–718. https://doi.org/10.1037/met0000421

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Sjölander, A., & Vansteelandt, S. (2019). Frequentist versus Bayesian approaches to multiple testing. *European Journal of Epidemiology*, *34*(9), 809–821. https://doi.org/10.1007/s10654-019-00517-2

Varvoglis, M., & Bancel, P. A. (2015). Micro-psychokinesis. In In E. Cardeña, J. Palmer, & D. Marcusson-Clavertz (Eds.), *Parapsychology: A handbook for the 21st century* (pp. 266–281). McFarland & Co.

Varvoglis, M., & Bancel, P. A. (2016). Micro-psychokinesis: Exceptional or universal? *Journal of Parapsychology*, *80*(1), 37–44.

Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š., … Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, *5*(11), Article 11. https://doi.org/10.1038/s41562-021-01211-8

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond " *p* < 0.05." *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Appendix

Calculation of False Positive Error (FPE) for Bayes factors in Jakob.

Monte Carlo calculations can be used to estimate the FPE rate. The BFs are marginal probabilities that are typically estimated via a Markov Chain Monte Carlo (MCMC) procedure and this can take considerable computer time. To speed the FPE estimates with stopping rules (which need millions of BF estimates), it is useful to have a closed expression to approximate BF from values *N* and simulation *t* scores, which have a monotonic relation to the BFs. The (somewhat tedious) details follow.

As a starting point, observe that for *p* values < .1 and a standard default prior of Cauchy(0, 0.71) (Wagenmakers, 2022):

$$BF\left(p, N\right) \sim \frac{1}{3p\sqrt{N}}$$

Using the closed form expression for the *p* value of a *t* score (in a normal approximation), we can try a general form for BF as:

$$BF(t,d) = \frac{9}{\sqrt{N}\left(1 - \mathrm{Erf}\left(\frac{t}{d}\right)\right)}$$

The scale factor *d* depends on both *N* and the prior, and the numerator value 9 is a convenience factor determined from fits. Next is to find an approximation for the scale factor *d*(*N*) under the Cauchy(0.05,0.05) prior used by *Jakob*. To do this, MCMC BFs are calculated for random data using a representative set of *N*s (using the statistical platform R, version 4.3.1). The *N* values used here are (200, 300, 500, 700, 900, 1000, 1200, 1400, 2000, 3000), which span the size of data sets in *Jakob*. By inspection, we find a fairly simple form for the scale factor:

$$d(N) = 1 + ae^{g/N} + be^{\frac{h}{\sqrt{N}}}$$

Values of *d*(*N*) that give a best fit to BF(*t*, *d*) in the vicinity of BF=10 are then determined for the *N* values above by iterating *d* and visually assessing the fit near BF=10 in a graphical plot. This takes a few minutes and is straightforward to do. The (*d*, *N*) value pairs are then fitted using *Mathematica's* NonlinearModelFit routine to the expression above. This yields parameter values for *d*(*N*) that are adapted to the Cauchy prior used in *Jakob*, as:

$$(a,b) = (-1.4831, +2.072) \qquad (g,h) = (-211.2, -4.4428)$$

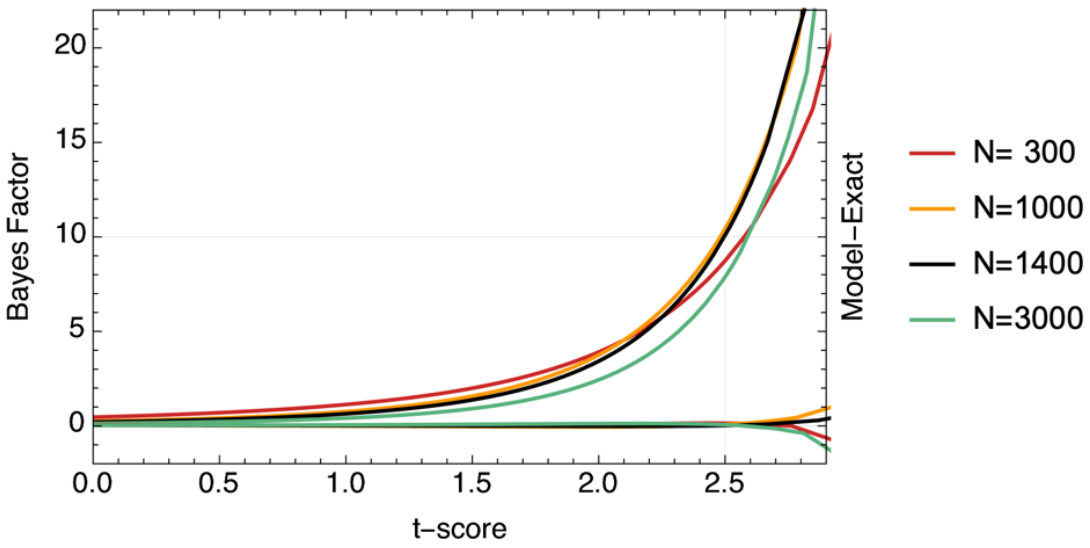
The BFs for the *Jakob* prior can thus be estimated quickly for a given *t* and *N* as:

$$BF(t,d) = \frac{9}{\sqrt{N}\left(1 - \mathrm{Erf}\left(\frac{t}{d(N)}\right)\right)}$$

Figure A.1 shows the accuracy of BF(*t*, *d*) for several sizes of data set. The *t* scores are values of Student *t* tests on simulated data with the 30-trial tasks, as in *Jakob*. The exact BFs (from full MCMC calculations done in R) versus *t* score are shown in Figure A.1 in the traces that curve upward. Note that *t* scores of around 2.5 correspond to BFs near 10, but that the BF value depends on N. The horizontal traces show the difference between exact BFs and the estimations using BF(*t*, *N*) outlined above. The estimates are quite good across the range of *N* and for *t* scores up to 2.6; beyond that, the deviations from exact BFs increase, but the fractional accuracy is still high for BFs well beyond 20. An interesting detail is that, for a fixed value of the *t* score, BFs are lowered at extremes of both small and large *N*. This is a consequence of Bayesian analysis when effects are small. For small effects, the Null is favored more for small samples, and BFs are more stringent with large samples (for a discussion, see Rouder et al., 2009).

Figure A.1

Analytical Estimates of BF as a Function of Sample N and T-Test Value



Note: The relation between BF and the one-sided t test for fixed sample N. The full MCMC calculations of BFs, for a selection of sample sizes (Ns) are shown as the four rising curves. The BFs rise sharply as t-test values rise above ~2. The horizontal traces show, for the same selection of Ns, the difference (Model-Exact) between calculated BFs and their values given by the approximate formula for BF(t, N) above. The negligible differences indicate that the formula is highly accurate over the selected range of Ns and t-scores.

Monte Carlo estimations of the FPE rate are shown in Table A.1. The FPE rate depends on the stopping rule and a variety of scenarios are simulated. Details are as follows. Null data arrays are generated using *Mathematica's* (version 12.3.1) RandomInteger function to obtain 10,000 surrogate data sets of *N*Final trials each. The sequential *t* scores for each array are calculated (that is, the score at entry *n* is the *t* value for trials 1 through *n*). The *t* scores are then converted to BFs using BF(*t*, *d*). The FPE rate is estimated by counting the fraction of arrays with a BF > 10 over a range of trials that begins at *N*Start and ends at *N*Final. Table A.1 shows the FPE for *N*Start set to 100 and 500, and for values of *N*Final of 300, 1000, 1400 and 3000. Depending on the stopping rule, data can be checked for the stop criterion continuously or more sparsely, at either fixed or opportunistic intervals. *Jakob* specified daily checking which, if followed, would impose a sampling rate of approximately 20%. The tests in Table A.1 comprise sampling rates of 100, 40, 20, 2 and 0 per cent. Errors for the FPE estimate are estimated in turn by a bootstrap resampling of the 10,000 MC arrays of each scenario. This gives an estimation of FPE for a single Bayesian *t* test BF as a function of total array length, the start check position and the sampling rate.

The relevant FPE estimates for *Jakob* are for at least one instance of BF > 10 in a set of 3 arrays. This roughly triples the FPE estimates for single tests and is shown in the first half of Table A.1. Note that the FPE is fairly insensitive to the sampling rate. Significant control of the FPE rate only occurs as the sampling becomes sparse. Setting the position of the first check, *N*Start, to larger values also provides a way to control the FPE. In the Table, the errors of estimates are tighter for the case of no stopping, and these are indicated separately in the last column.

Table A.1

False Positive Error for Bayes Factors > 10

3 Tests						
		FPE (± 0.018)				FPE (± 0.008)
N_{Start}	N_{Final}	Continuous	40%	20%	2%	No stopping
100	1000	0.141	0.133	0.130	0.092	0.017
500	1000	0.077	0.073	0.072	0.058	0.020
100	1400	0.157	0.152	0.147	0.109	0.018
500	1400	0.117	0.114	0.110	0.093	0.023
1 Test						
		FPE (±0.002)				FPE (± 0.001)
N_{Start}	N_{Final}	Continuous	40%	20%	2%	No stopping
100	1000	0.05	0.047	0.046	0.032	0.006
500	1000	0.026	0.025	0.024	0.02	0.007
100	1400	0.056	0.053	0.052	0.038	0.006
500	1400	0.04	0.039	0.038	0.032	0.008

Commentaire sur Jakob et al.: Comment Lire un Article

Peter A. Bancel

Résumé:Dans ce numéro de JAEX, Jakob et al (2024) rapportent les résultats d'une expérience visant à tester si un effet psychokinétique implicite lié à l'intention inconsciente peut modifier l'issue d'événements

quantiques. L'expérience est bien conçue et exécutée. Cependant, l'allégation de preuves à l'appui de l'hypothèse expérimentale est compromise par des erreurs d'analyse et d'omission. Certaines ne sont pas évidentes et nécessitent une lecture attentive pour être discernées. J'explique la nature des erreurs, où et comment elles se produisent, et les conséquences sur les conclusions faites par les auteurs. L'accent est missur le fait qu'un moyen utile de détecter ces problèmes consiste à suivre la chaîne d'inférence dans les rapports expérimentaux. Les rédacteurs, les reviewers, et surtout les auteurs peuvent protéger l'intégrité de la littérature scientifique en ayant une lecture attentive incluant les éléments ici développés.

Translation into French by Antoine Bioy, Ph. D.

Kommentar zu Jakob et al.: Wie man einen Artikel liest

Peter A. Bancel

Zusammenfassung: In dieser Ausgabe von JAEX berichten Jakob et al. (2024) über die Ergebnisse eines Experiments, in dem getestet wurde, ob ein impliziter psychokinetischer Effekt, der mit einer unbewussten Intention zusammenhängt, das Ergebnis von Quantenereignissen verändern kann. Das Experiment ist gut geplant und durchgeführt worden. Die Behauptung jedoch, es läge ein Beweis zugunsten der experimentelle Hyothese vor, wird jedoch durch fehlerhafte Analyse und durch Auslassungen unterminiert. Einige sind nicht offensichtlich und erfordern eine sorgfältige Lektüre, um sie zu erkennen. Ich erkläre die Art der Fehler, wo und wie sie auftreten, und die Konsequenzen für die Schlussfolgerungen, die die Autoren ziehen. Es wird betont, dass eine hilfreiche Methode, solche Probleme zu erkennen, darin besteht, auf die Abfolge der Schlussfolgerungen in den experimentellen Protokollen zu achten. Herausgeber, Gutachter und insbesondere Autoren können zum Schutz der Integrität der wissenschaftlichen Literatur beitragen, indem sie auf diese Weise sorgfältig lessen.

Translation into German by Eberhard Bauer, Ph. D.

Comentário sobre Jakob et al.: Como Ler um Artigo

Peter A. Bancel

Resumo: Neste número da JAEX, Jakob et al (2024) relatam resultados de um experimento que visava testar se um efeito psicocinético implícito relacionado à intenção inconsciente pode alterar o resultado de eventos quânticos. O experimento é bem concebido e executado. No entanto, a alegação de evidência para apoiar a hipótese experimental fica comprometida por erros de análise e omissão. Alguns não são óbvios e exigem uma leitura cuidadosa para serem reconhecidos. Explico a natureza dos erros, onde e como surgem, e as consequências para as inferências que os autores fazem. Enfatiza-se que uma maneira

útil de detectar tais problemas é seguindo a cadeia de inferências em relatórios experimentais. Editores, revisores e, especialmente, autores podem proteger a integridade da literatura científica lendo cuidadosamente dessa forma.

Translation into Portuguese by Antônio Lima, Ph. D.

Comentario Sobre Jakob et al.: Cómo Leer un Artículo

Peter A. Bancel

Resumen: En este número de *JAEX*, Jakob et al. (2024) reportan los resultados de un experimento para evaluar si un efecto psicokinético implícito relacionado con una intención inconsciente puede alterar el resultado de eventos cuánticos. El experimento está bien concebido y ejecutado. Sin embargo, la pretensión de apoyo a la hipótesis experimental se ve socavada por errores de análisis y de omisión. Algunos no son obvios y requieren una lectura cuidadosa para discernirlos. Explico la naturaleza de los errores, dónde y cómo se producen, y las consecuencias para las inferencias que extraen los autores. Subrayo que una forma útil de detectar estos problemas es seguir la cadena de inferencias en los informes experimentales. Editores, revisores y, sobre todo, autores pueden proteger la integridad de la literatura científica leyendo de este modo.

Translation into Spanish by Etzel Cardeña, Ph. D.

Response to Comment on Jakob et al.: How to Read a Paper

Marissa-Julia Jakob[1] Moritz C. Dechamps[1] Markus A. Maier

Ludwig-Maximilians-Universität, München

Abstract: We thank Peter Bancel for highlighting the potential strengths and weaknesses of our study and for the intensive examination of the applied methods in order to improve our understanding of the reported effect and support future attempts to scientifically study micro-PK effects. Many arguments put forward in the Comment deserve a deeper examination and we appreciate the discussion raised by Bancel, but we do not agree with the main points of his criticism. In the following, we briefly address the central arguments provided against our analysis strategy and our interpretation of the data and present our view on this matter.[1]

1. Evidence claims not adjusted for multiple testing

Bancel first discusses the issue of multiple testing and its impact on the validity of our claims about Bayesian evidence. One argument was that our claim of strong evidence is flawed because adjustments for multiple testing should have been performed due to the dependence of the tests conducted in the study that used three different micro-PK tasks within one sample.

As Bancel acknowledges, within Bayesian testing approaches adjustments for multiple testing are only mandatory when the tests performed are dependent (Sjölander & Vansteelandt, 2019). This is usually the case when some kind of interrelations between the three micro-PK measurements are present in the data. However, in our view, the fact that the same sample is tested across the three measurements does not automatically imply statistical dependence. We agree that evidence for an effect in all three hypotheses should be based on a common mechanism that is an unconscious belief about the self. However, we argue that the hypotheses can be interpreted

1 Shared first authorship.

Journal of Anomalous Experience and Cognition
2024, Vol. 4, No. 1, pp. 79-87

https://doi.org/10.31156/jaex.25782

as independent because they are based on three highly specific psychological beliefs about someone's reality that should only produce micro-PK effects when the stimuli used in the respective task match the content of the relevant belief. In addition, it is possible that the "dependent" personality trait is elicited significantly better than the others because of a better match between this belief and its corresponding stimuli, which would lead to strong evidence for only one of the tests, independent of the evidence obtained for the others. Since we did not know beforehand whether the stimulus material used in the three DVs sufficiently matched the beliefs tested in the study, we consequently formulated our main hypothesis less strictly: We predicted that we would find strong Bayesian evidence for H_1 for at least one of the three micro-PK tasks performed. In addition, we counterbalanced the order of the micro-PK tasks used in the study to control for any order effects between tasks.

Bancel goes on to argue that the data sets for the three PT groups are also dependent, because the groupings are highly correlated. Although this is true, we would like to point out that it is not the correlation of personality traits that is relevant here, but a possible systematic relation between the correlated PT scores and the three micro-PK tasks. From a theoretical point of view, statistical dependence in this context would imply that any (group-related) performance on one micro-PK task should have an impact on performance on any of the other micro-PK tasks included in this study. This assumed interrelation of QRNG outcomes would in itself constitute a micro-PK effect. The micro-PK performance in our study is based on QRNG outputs that involve a true random mechanism.

In any textbook of statistics, the core example of statistical independence is true random events. Regardless of whether different random events are produced by different individuals or by the same individual, they are always considered to be statistically independent. Thus, in theory, one should assume a priori that such measurements are statistically independent. Otherwise, the a priori assumption of statistical dependence between truly random events would not only a priori claim the existence of micro-PK, but would also, in the case of using quantum-based RNGs, question the validity of the Bell theorem and lead to violations of the locality assumption in macroscopic domains. This in turn would make it impossible to document objectively (for which locality would be required) the existence of any effects under study. In other words, the empirical documentation of micro-PK effects to accepted scientific standards would become impossible or even unnecessary, since the effect to be demonstrated is already defined a priori as existing.

Despite this theoretical paradox, which we encounter when the statistical dependence of true random events is postulated a priori, let us assume for a moment that Bancel's argument is valid. Our theoretical background is consistent with his view. We assume that micro-PK exists and that it affects our three different micro-PK results. Furthermore, since the three PT groups are empirically correlated, the three micro-PK measures may (or may not) also be related in some way. If they were related, they would have to be considered statistically dependent and multiplicity controls would have to be performed. Thus, Bancel's argument is empirical rather than theoretical. To address this empirical argument, we will next provide empirical tests of the statistical independence of our three micro-PK measures. If the three measures were statistically independent, then an above-chance score (hit) or a chance and below-chance score (miss) on one of the tests should be indicative of a hit or a miss on the other tests. We tested this assumption with three separate Pearson's chi-squared tests of independence. The results indicated that the likelihood of scoring a hit or a miss was not significantly associated with any outcome (hit or miss) on the other tasks (dependent vs. avoidant χ^2 (1, N = 2,403) = 0.43, p = .51; dependent vs. obsessive-compulsive χ^2 (1, N = 2,403) = 1.08, p = .30; avoidant vs. obsessive-compulsive χ^2 (1, N = 2,403) = 0.13, p = .72). These analyses indicate that hits or misses are completely randomly distributed across the three tasks and they empirically document that statistical independence can be assumed between the three micro-PK tasks. The power to detect even small violations (w = .1) of statistical independence was 99% in each χ^2 test.

Since statistical independence exists across the three micro-PK measurements multiplicity control for our Bayesian tests did not need to be performed (Sjölander & Vansteelandt, 2019) and the strong Bayesian evidence (BF_{10} > 10) reported for the micro-PK effect found within the dependent PT-group can still considered to be valid.

Finally, if one still prefers to maintain the statistical dependence assumption, one needs to perform multiplicity controls, and these would affect expectations about the prior model probabilities. One possibility is to adjust the model probabilities in such a way that not each (of the three) null hypotheses has a probability of ½ (leading to the a priori statement that the probability of finding no effect in all three tests is $½^3$= .125), but that the total prior probability of finding no differences in all three tests combined is ½ (null control method; Williams et al., 2016). Following de Jong (2019), this could be achieved by changing the prior model probability for the H_0 from 0.5 to $0.5^{¾}$ = 0.63 (see Westfall et al., 1997). It is important to note that this does not change the individual Bayes factors for each test, but changes the posterior probability of an effect when considering the entire study by a factor of 0.37/0.63=0.59 as seen in (1).

$$\frac{P(H_1|D}{P(H_0|D} = \frac{P(H_1)}{P(H_0)} \times \frac{P(D|H_1)}{P(D|H_0)} (1)$$

$$Posteriorodds = Priorodds \times BF$$

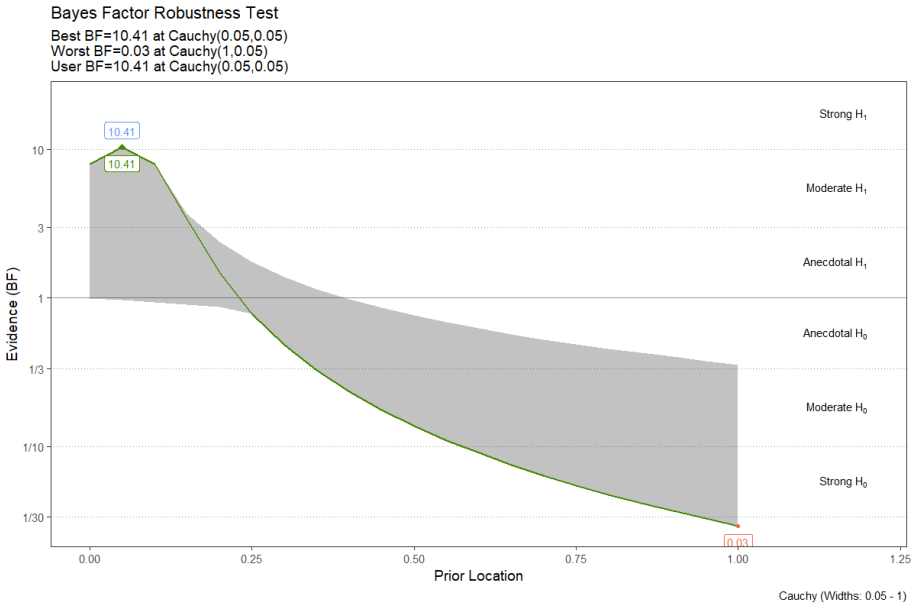
2. Lack of Context for Bayesian *t* Tests

Bancel emphasizes the importance of conducting sensitivity checks in Bayesian analysis, especially when using informed priors and dealing with small effects. He argues that an a priori justification for the prior based on the results of previous studies does not justify overlooking the sensitivity of the results to parameter choices, and links this issue to experimenter psi in psi research. We agree that the choice of the prior involves a degree of freedom for the experimenter. In Bayesian analysis, this is considered a strength, since it is advantageous to specify a test for the expected effect directly. It is therefore good practice to use these features of Bayesian statistics when they can be applied with some confidence. This is the case here, as we had a fairly certain expectation of the expected effect size for this type of micro-PK experiment based on previous studies and the literature. This deliberately translated expectation was pre-registered a priori, eliminating the degree of freedom in the analysis, and was ultimately confirmed by parts of the results of the study that showed an effect size of *d* = .07, which fits well with our a priori estimate.

Not surprisingly, a sensitivity analysis shows that this effect is not very robust to parameter changes (see robustness analysis in fig. 1 below). This is to be expected, since a very small effect can only be detected with a reasonable sample size if the test used is sensitive to that effect size. This is partly why we prefer Bayesian methods in this line of research, and is not a problem as long as the parameters are registered in advance.

Figure 1

Robustness analysis of the effect in the dependent sample. It can be seen that the BF has its highest value for the a priori chosen informed Cauchy parameters. The evidence transitions into pointing towards H0 for a very narrow prior (green line) at around Cauchy(0.25, 0.05). The gray area shows the BFs for different prior widths.



3. An Undetermined Type I Error Rate of the Primary Hypothesis

Bancel criticizes a lack of discussion of the study's false positive error rate (FPE) and its relation to a frequent testing of sequential evidence. He argues that the three tests are interdependent and therefore the FPE in this case is approximately 15%. We appreciate the effort put into the false positive error rate (FPE) analysis. Bancel cites Schönbrodt and Wagenmakers (2018), who provide a simple way to calculate the FPE in Bayesian designs and provide comparable but slightly lower results. This analysis shows an FPE of 3.8% for a sequential design with n-start = 100, n-max = 1,400, and a step size of 10 participants, which are conservative estimates. This shows that the alpha error for each test is well below the generally accepted threshold of .05.

Following the argument above that the data (the micro-PK trials corresponding to the specific personality trait) of the three tests are independent, it is not necessary to control for the experimental error rate (EER), which is typically done with corrections when testing multiple frequentist comparisons simultaneously. The power of each test is not affected by the outcome of the other tests, and the stopping criteria are based on each test individually reaching a specific BF threshold, not on a combination of tests. Therefore, the Type I error rate remains as reported for each hypothesis evaluated individually.

It can also be seen that the FPE changes only slightly with different step sizes (i.e., test intervals), and that the sequential Bayesian designs are usually quite robust to over-testing errors when a minimum N is used for the first test (e.g., changing the step-size to 50 reduces the FPE to 2.8%).

4. Failure to Follow the Pre-Registered Procedure

Bancel criticizes a deviation from the pre-registered stopping rule during data collection, leading to an undefined degree of experimenter freedom in the protocol. Specifically, the data collection was not stopped when the BF > 10 criterion was met for the first time.

We agree that some deviations from the pre-registered protocol were made. All deviations are explicitly mentioned in the original article. However, they do not compromise the evidence of the studies, since Bayesian evidence only becomes more precise as more data are added. According to the protocol, data collection should have stopped when the evidence criterion of BF > 10 was reached for the first time at n = 820. In the analysis process, we missed this exact point and observed a decrease in evidence shortly thereafter. Therefore, we decided to collect more data until a conclusive result was reached, as suggested by Schönbrodt et al. (2017). In Bayesian statistics, it is always possible to continue collecting data while updating the data analysis, as Bayes factors are consistent because they converge either to zero (if H_0 is true) or to infinity (if an effect is present) for one-tailed designs (Bayarri & Berger, 2004; Morey & Rouder, 2011). Schönbrodt et al. (2017) showed that sequential Bayesian designs have a lower long-term rate of misleading evidence than frequentist procedures with 5% Type I and 20% Type II error rates, while most errors occur at small sample sizes. Unlike p-values, the interpretation of Bayes factors does not depend on stopping rules (Rouder, 2014). Therefore, adding data to a Bayesian analysis is never an issue, and stopping criteria are rather a means to design studies efficiently. This was the case here, and the stopping rule of N = 1,000 participants in the absence of evidence was implemented for economic reasons. With the narrowly informed prior we chose, obtaining strong evidence for H_0 would require a very large sample size. At the time we uploaded the pre-registration, our resources were limited to 1,000 participants. As the process progressed, more resources became available to continue data collection beyond N = 1,000.

Furthermore, we never intended to perform a frequentist overall test of the three micro-PK tasks. Instead, we view them as three independent experiments, each with its own independent and dependent variables. This was emphasized in the pre-reg-

istration form, which allowed us to stop data collection as soon as any of the three experiments reached our statistical evidence criterion. It is up to the research community to continue data collection if they are interested.

5. Inadequate Treatment of the Control Group Analysis

Finally, Bancel notes the lack of a direct comparison between the target and control groups and criticizes the use of one-tailed tests for the former and two-tailed tests for the latter, which makes the tests inappropriate for comparison. As explained in the pre-registration form, we deliberately chose a one-sample *t*-test design against the expected value under chance rather than a direct group comparison for theoretical reasons. Our main point is not that the two groups are different, but that there is a micro-PK effect due to intentional observation within the PT-high group.

The group split can be seen as a sample pre-selection of individuals with pronounced traits (PT-high groups) that favor the micro-PK effect. A direct comparison of the groups is not meaningful, because the chosen splitting criterion is based on a continuous measure (VDS-30 questionnaire). Therefore, it is possible that the control group also shows (weaker) micro-PK effects in the same or in the opposite direction as the experimental group, since they can also be considered as motivated observers. Consequently, a two-tailed *t*-test was performed for the control group, since a non-random, less-than-chance result does not fit H_0 either. In contrast, the hypotheses for the experimental group were formulated as one-tailed, following the predictions of the Emotional Transgression Model as our theoretical background. Note that the group differences are not explained by the one-tailed vs. two-tailed setting variation, which can be checked with the data set and analysis scripts provided at OSF.

In summary, some of the criticisms in the comment are valid, but we disagree with the conclusion that they lead to nonconfirmatory results and reduce the significance of the strong evidence for the micro-PK effect found in one of the three experiments. Note that the analyses of all PT-high data suggested by the author of the comment yielded $p = .06$, which just barely exceeds the convention of .05. As noted in the Discussion section of our paper, further analyses of the combined score of the three tasks will be included in another paper we are currently working on. In addition, we disagree that the citation of "Jahn et al. (2000)" was misused in our paper. We mentioned this study as an example of an initial micro-PK effect that could not be replicated. However, our claim was not that the replications failed due to a decline effect. Moreover, we believe that such minor complaints should be part of a review process rather than raised in a comment.

References

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science, 19(1),* 58-80. https://doi.org/10.1214/088342304000000116

de Jong, T. (2019). A Bayesian approach to the correction for multiplicity. https://doi.org/10.31234/osf.io/s56mk

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16*, 406–419. https://doi.org/10.1037/a0024377

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*, 301- 308. https://doi.org/10.3758/s13423-014-0595-4

Schönbrodt, F.D., Wagenmakers, E.J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychon Bull Rev 25*, 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22(2),* 322. https://doi.org/10.1037/met0000061

Sjölander, A., & Vansteelandt, S. (2019). Frequentist versus Bayesian approaches to multiple testing. *European Journal of Epidemiology, 34(9),* 809–821. https://doi.org/10.1007/s10654-019-00517-2

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika, 84(2),* 419–427. https://doi.org/10.1093/biomet/84.2.419

Williams, P., Heathcote, A., Nesbitt, K., & Eidels, A. (2016). Post-error recklessness and the hot hand. *Judgment and Decision Making, 11(2),* 174–184. https://doi.org/10.1017/S1930297500007282

Response au Commentaire sur Jakob et al.: How to Read a Paper (Comment lire un article)

Marissa-Julia Jakob Moritz C. Dechamps Markus A. Maier

Résumé: Nous remercions Peter Bancel d'avoir souligné les forces et les faiblesses potentielles de notre étude et d'avoir examiné en profondeur les méthodes appliquées afin d'améliorer notre compréhension de l'effet rapporté et de soutenir les futures tentatives d'étude scientifique des effets des micro-PK. De nombreux arguments avancés dans le commentaire méritent un examen plus approfondi et nous apprécions la discussion soulevée par Bancel, mais nous ne sommes pas d'accord avec les principaux points de sa critique. Dans ce qui suit, nous abordons brièvement les principaux arguments avancés à l'encontre de notre stratégie d'analyse et de notre interprétation des données, et nous présentons notre point de vue sur la question.

Translation into French by Antoine Bioy, Ph. D.

Antwort auf den Kommentar zur Jakob et al.: Wie man eine Arbeit liest

Marissa-Julia Jakob Moritz C. Dechamps Markus A. Maier

Wir danken Peter Bancel für das Aufzeigen möglicher Stärken und Schwächen unserer Studie und für die intensive Überprüfung mit den angewandten Methoden, um unser Verständnis des berichteten Effekts zu verbessern und zukünftige Versuche, Mikro-PK-Effekte wissenschaftlich zu untersuchen, zu unterstützen. Viele der in dem Kommentar vorgebrachten Argumente verdienen eine eingehendere Prüfung, und wir schätzen die von Bancel geführte Diskussion, stimmen aber nicht mit den Hauptpunkten seiner Kritik überein. Im Folgenden gehen wir kurz auf die zentralen Argumente ein, die gegen unsere Analysestrategie und unsere Interpretation der Daten vorgebracht wurden, und legen unsere Meinung dazu dar.

Translation into German by Eberhard Bauer, Ph. D.

Resposta ao Comentário sobre Jakob et al.: Como Ler um Artigo

Marissa-Julia Jakob Moritz C. Dechamps Markus A. Maier

Resumo: Agradecemos a Peter Bancel por destacar os potenciais aspectos positivos e fragilidades de nosso estudo e pela rigorosa análise dos métodos aplicados, visando melhorar nossa compreensão acerca do efeito relatado e apoiar futuras tentativas de estudar cientificamente os efeitos micro-PK. Muitos argumentos apresentados no Comentário merecem uma análise mais profunda, e agradecemos a discussão levantada por Bancel, porém, não concordamos com os principais pontos de sua crítica. A seguir, abordamos brevemente os argumentos centrais sugeridos contra nossa estratégia de análise e nossa interpretação dos dados e apresentamos nossa visão sobre o tema.

Translation into Portuguese by Antônio Lima, Ph. D.

Respuesta al Comentario sobre Jakob et al.: Cómo Leer un Artículo

Marissa-Julia Jakob Moritz C. Dechamps Markus A. Maier

Resumen: Damos las gracias a Peter Bancel por señalar los posibles puntos fuertes y débiles de nuestro estudio y por el examen intensivo de los métodos aplicados con el fin de mejorar nuestra comprensión del efecto mencionado y apoyar futuros intentos de estudiar científicamente los efectos micro-PK. Muchos de los argumentos expuestos en el comentario merecen un examen más profundo y agradecemos el debate planteado por Bancel, pero no estamos de acuerdo con los puntos principales de su crítica. Abordamos brevemente los argumentos centrales aportados contra nuestra estrategia de análisis y nuestra interpretación de los datos y presentamos nuestra opinión al respecto.

Translation into Spanish by Etzel Cardeña, Ph. D.

2024, Vol. 4, No. 1, pp. 88-121

Journal of Anomalous Experience and Cognition (JAEX)

State, Trait, and Target Parameters Associated with Accuracy in Two Online Tests of Precognitive Remote Viewing[1]

Julia Mossbridge
University of San Diego

Kirsten Cameron
TILT

Mark Boccuzzi
Windbridge Institute

Abstract: *Objective*. To better characterize the relations between accuracy on precognitive remote viewing (PRV) tasks and potentially relevant trait, state, and target parameters, we gathered PRV data in two online experiments and examined accuracy relative to: sex-at-birth, gender, age, anxiety, unconditional love, and target interestingness. *Method*. In experiment 1 we used a forced-choice, uncontrolled-time, self-judged PRV task for which 682 unpaid participants contributed a total of 5,432 trials. Experiment 2 used a free-response, controlled-time, independently judged PRV task for which 307 paid participants each contributed a single trial. In neither case were the participants pre-screened for precognition ability. *Results*. In experiment 1 (forced-choice PRV task), there was no significant target precognition and no effect of age on PRV performance, but we found a complex effect of sex-at-birth. We also found that targets most likely to be correctly predicted were also more likely to be judged as interesting compared to targets most likely to be incorrectly predicted; a pre-registered analysis confirmed this effect. In experiment 2 (free-response PRV task) we found significant target precognition, no effect of age on performance, and a weak and indirect effect of gender. Feelings of unconditional love and anxiety were both correlated with higher accuracy in experiment 2. Again, target interestingness was positively related to accuracy. *Conclusion*. These results suggest that accuracy on PRV tasks is related to the emotional state of participants and target interestingness, and that task char-

1 Address correspondence to Julia Mossbridge, Ph. D., **jmossbridge@gmail.com**. The authors thank the Bial Foundation for funding the first author (Bial grants 2014_260, 97_16, 369/20). We are also grateful to Theresa Cheung who funded the website that allowed us to gather data for the first analysis set. We are grateful to The Windbridge Institute, for building and maintaining the website and software to gather data for both analysis sets. See also ***Letter to the Editor*** in this issue.

https://doi.org/10.31156/jaex.24743

acteristics mitigate overall performance. We provide recommendations for future research based on these observations.

Keywords: precognition, remote viewing, anomalous cognition, parapsychology, gender differences, sex differences, love

Highlights

- To better understand precognition, we examined how age, sex-at-birth, gender, two different mood states, and target interestingness related to accuracy on two PRV tasks.
- A forced-choice PRV task revealed no target precognition, but a free-response PRV task revealed target precognition at a rate higher than chance.
- Target interestingness was significantly related to greater accuracy in both tasks.
- Greater feelings of unconditional love were significantly related to greater accuracy in the free-response PRV task.

Precognition, the scientific name for cases in which an individual knows, senses, or acts upon accurate information about a seemingly unpredictable future event without using ordinary inference or sensory cues, is one of the most robust psi phenomena (for recent reviews see Mossbridge, 2023; Mossbridge & Radin, 2018a, b). Evidence for precognition has been provided by controlled laboratory experiments, in which participants must implicitly or explicitly predict the future outcome of a truly random process (e.g., Bem, 2011; Bem et al., 2015; Mossbridge et al., 2012; Watt et al., 2020). Rigorous laboratory examinations of precognition have indicated it is statistically reliable, as per a report released as part of a U.S. government intelligence program (Mumford et al., 1995) and related work (e.g., May, 2014).

Decades later, only a small portion of the scientific community concurs that precognition has been verified, though scientists have continued to find evidence supporting precognition. Controlled experiments examining behavioral precognition use one of two task types: 1) implicit tasks, in which participants are not asked to directly or consciously predict the future but instead are asked to act in ways that are later associated with future events (e. g., button presses), and 2) explicit tasks such as precognitive remote viewing (PRV), in which participants are asked to consciously make a

prediction about an unknown future event (for reviews: Mossbridge, 2023; Mossbridge & Radin, 2018a,b; rebuttals: Houran et al., 2018; Schwarzkopf, 2018; response to rebuttals: Mossbridge & Radin, 2018b). In hopes of better understanding precognition, and initiating more nuanced methods of investigation into precognitive remote viewing in particular, in this paper we examine several participant and target parameters that may affect performance on PRV tasks.

Briefly, during a PRV task participants are asked to describe a target that will be randomly selected and presented to them only after they have completed their description, which is called a *transcript*. There are two general types of PRV task: forced-choice and free-response. In a forced-choice PRV task, the participant themselves will select which of two or more potential targets best matches their transcript before the target is selected. In a free-response PRV task, the participant will submit their transcript itself to the experimenter before the target is selected. Once a PRV experiment is complete, the experimenter determines the relation (if any) between the transcript and the target by asking independent judges to compare the transcript with the actual target as well as one or more non-target stimuli, with the judges unaware of which stimulus was the real target (Nelson et al., 1996; Targ, 2019; Utts, 1995).

PRV has been employed in several non-experimental applications, including predicting the outcomes of sporting events and fluctuations in stock prices (Katz et al., 2019, 2021; Kolodziejzyk, 2013; Smith et al., 2014; Tait, 2019) and answering important research questions about the future of humanity and the planet (Mossbridge & Vivanco, 2022; Schwartz, 2021). Looking towards future PRV applications, artificially intelligent systems could be used to better select participants, transcripts, and targets with the aim of creating a four-dimensional map of the future that would better inform decisions in the present based on additional information about surprising or concerning effects in the future (Mossbridge, 2023b). Thus, PRV may eventually have a large impact on humanity, but we do not currently have a good understanding of the parameters that influence performance on PRV tasks or the mechanisms governing precognition (Evrard & Ventola, 2018). This slow progress may be caused by taboos against scientific investigation of apparently nonphysical phenomena, resulting in the reticence of public science organizations to fund precognition research and of mainstream psychology and neuroscience researchers to become involved (Cardeña, 2014, 2018; Wargo, 2015). Progress has also been slow because existing research has revealed that participant and target parameters are related to precognitive accuracy in complex ways, making the most basic questions about precognition difficult to address simply.

What is known about the parameters affecting PRV performance? Thus far it is clear that two personality traits, openness to experience and extraversion, are often associated with performance on individually-tested extrasensory perception or *psi* skills such as telepathy, clairvoyance and precognition (Hitchman et al., 2012; Honorton et al., 1998; Palmer & Carpenter, 1998), but the direction of this relation may depend on the particular task performed to assess a given skill (Mossbridge, 2023; Mossbridge & Radin, 2021). Some gender effects on performance in forced-choice precognition tasks have been reported (Bierman & Scholte, 2002; Radin & Lobach, 2007; Zdrenka & Wilson, 2017), but they have not been rigorously examined in free-response tasks, and age effects have been under-examined. This may be because several of the early Star Gate program remote viewers were women who provided data on par with sessions from talented men (notably Hella Hammid and Charlene Cavanaugh Shufelt; Smith, 2005), lending no anecdotal support for a gender difference in free-response PRV tasks. Here in each of two experiments, we examine the potential relations between age, gender, sex-at-birth, or reproductive hormone status and performance on forced-choice and free-response PRV tasks.

Feelings related to self-transcendence may also support PRV accuracy. This hypothesis arises from several lines of evidence. First, the non-anxious state of *mindfulness* – a state of awareness of oneself and one's own surroundings while being filled with compassion – seems to be related to improved performance on forced-choice precognition tasks (Roney-Dougal et al., 2008; Roney-Dougal & Solfvin, 2011; Varvoglis et al., 2019). Meanwhile, ganzfeld stimulation, which can be considered to induce many of the aspects of mindfulness, compared favorably to non-stimulated sessions in a crossover design in terms of performance on a free-response PRV task (Roe et al., 2020). The experience of being in a positive or expansive mood has also been correlated with better accuracy in a largely precognitive ESP study conducted with ganzfeld stimulation (Carpenter, 2005), and was seen to promote stronger relations between state and trait parameters and precognitive scoring (Carpenter et al., 2021). Relatedly, one of our previous examinations of free-response PRV indicated that participants experiencing the self-transcendent state of unconditional love performed better at PRV in one condition (Mossbridge et al., 2021b). Taken together, these results motivated our examinations of the relations between free-response PRV accuracy, anxiety, and unconditional love in the second experiment reported here.

The final parameter we examined was target interestingness. Anecdotal reports suggest that interesting targets may improve remote viewers' motivation to perform the relatively laborious and attention-requiring free-response PRV process even though viewers do not know what the target is ahead of time. Relatedly, two studies

have provided evidence that targets rated as more intriguing or awe-inspiring ("numinous" targets) are more likely to be correctly identified in free-response remote viewing (RV) and free-response PRV tasks (Krippner et al., 2019; Schwartz, 2007). However, to our knowledge there have been no peer-reviewed studies examining whether target interestingness is related to performance on forced-choice PRV tasks.

It is important to note that this report follows our discovery process from exploration to confirmation/disconfirmation. As described by the "SEARCH" strategy used in a previous examination of performance on online psi tasks (Mossbridge & Radin, 2021), we took an exploratory approach in analyzing the first portion of the data from experiment 1, then pre-registered a confirmatory analysis for the only effect we felt was robust enough to warrant further study (the effect of target interestingness). Before conducting experiment 2, we made four formal predictions, one a confirmatory prediction, and reported them to a granting agency. A fifth confirmatory prediction was added after the analysis of experiment 1 was complete. We thus made no attempts to correct for multiple comparisons during the exploratory phases or to make ourselves appear prescient in retrospect during the confirmation/replication phases.

Methods: Experiment 1

Data Separation by Date

Data from the experimental website at http://ThePremonitionCode.com/tester were obtained in two downloaded batches. The first consisted of data recorded from the launch date of the website (October 1, 2018) to midnight GMT on April 30, 2019, the second consisted of data recorded from 12:01 am GMT on May 1, 2019 to midnight GMT on April 30, 2020. These cutoff times were used because they corresponded relatively well to the data download times used in a related experiment published previously (Mossbridge & Radin, 2021).

Procedure

Visitors to the website were asked whether they would like to be part of an experiment. If so, they were shown an online consent form. People could use the website for precognitive remote viewing (PRV) practice without signing the consent form. After signing the consent form, participants could continue to use the practice and test features on the website.

Researcher Information

JM interacted with participants by responding to questions via email and sending out a newsletter to anyone who used the website (regardless of consent status). Her belief that the experiment would produce positive results was at the strongest level (5 on an integer scale from 1 [very little belief] to 5 [highest belief]). MB interacted with participants indirectly via periodic database queries as requested by JM or participants and his belief in a positive outcome of the experiment was also 5 on the same scale.

Participants

All procedures and protocols for human subjects' research were approved before use by the Institutional Review Board of The Windbridge Institute (WIIRB# 2018-MI-0930). We included as participants those who signed the electronic informed consent form and performed at least one trial on the site during the periods included within the two data batches, reported that they were 18 or over, and were not among the authors of this paper. Participants could optionally self-identify age, gender, and sex via a text field; not all participants provided this information.

We recorded data from a total of 470 participants in the first batch (age M = 43.9 years, SD = 13.2 years). Of those who reported sex: 190 were females (178 females were cis-gendered, 8 did not report a gender, 1 was gender fluid, 2 were non-binary, and 1 female was mostly female but at times male). With respect to males (n = 109), 100 males were cis-gendered and 9 did not report a gender.

We recorded data from 212 total participants in the second batch, 37 of whom had contributed data to the first batch. Of those 175 participants who registered only during the time frame of the second batch (age M = 42.5, SD = 11.6), 147 reported their sex: 98 were female and 49 were male. Of the females, 94 were cis-gendered, 1 did not report a gender, 1 was male gendered, 1 was fluid, and 1 was bigendered. Of the self-reported males, 41 were cis-gendered, 4 did not report gender, 1 was female gendered, 1 was nonbinary, 1 was gender fluid, and 1 was androgynous.

For the sake of the statistical analyses, we assumed all of the participants were unique individuals, but it was possible that different userIDs may have originated from the same person creating two different userIDs. However, each registered userID was verified with an email verification process, so we assume that using multiple IDs, if present at all, was rare.

Forced-choice Precognitive Remote Viewing (PRV) Task

The website task was designed to measure conscious precognition performance via a forced-choice precognitive remote-viewing (PRV) task (Figure 1). Users could choose to participate in the "practice" or "testing" versions of the task at any time and as often as they liked. The only differences between the practice and testing versions were the names of the tasks displayed to the users and the fact that participating in "testing" meant a user could compete for a position in the website's "Hall of Fame." Data from both tasks were combined except when noted in *Results.*

Participants were asked to perform the six steps of "controlled precognition" (i. e., precognitive remote viewing [PRV] with a forced-choice response), as outlined on the website as well as in an accompanying book (Cheung & Mossbridge, 2018). These steps were designed to help individuals feel that they could sense information about a visual target to be presented in the future. We had no way of determining whether users actually performed the steps before step 5. In step 5, they were presented with two categorical graphs (Figure 1, top) and asked to determine which of them represented the contents of the upcoming visual target. The software forced the users to choose between the two graphs at that point to continue.

Each graph showed the presence or absence of eight potential descriptor elements in the upcoming image (e. g., water or fluid, landform, human-made structure, animal or human, vegetation or food, space or sky, energy, words/symbols/ideas). We designed the graphs to represent a randomly selected but pre-paired set of 500 to 589 images (during the time period of the first batch, the original 500 targets were updated with 92 additional targets [Feb. 6, 2019], creating image pairing issues that resulted in the removal of 3 targets [April 1, 2019], resulting in 589 target images for the remainder of the time period of the first batch and during the entire time period of the second batch). The target images, called Profile A (described in the left or top graph) and Profile B (described in the right or bottom graph) were chosen to be mutually exclusive in their contents. For example, if the graph representing Profile A indicated that one possible target contained animals or humans, the graph representing Profile B would indicate a paired possible target that did not contain these same elements (Figure 1, top).

After the user had selected one of the graphs, the software would randomly choose one of the two images represented by the graphs to display to the participant (Figure 1, bottom). This was the target image that the users would have been intending to describe if they had been following the steps of controlled precognition. After view-

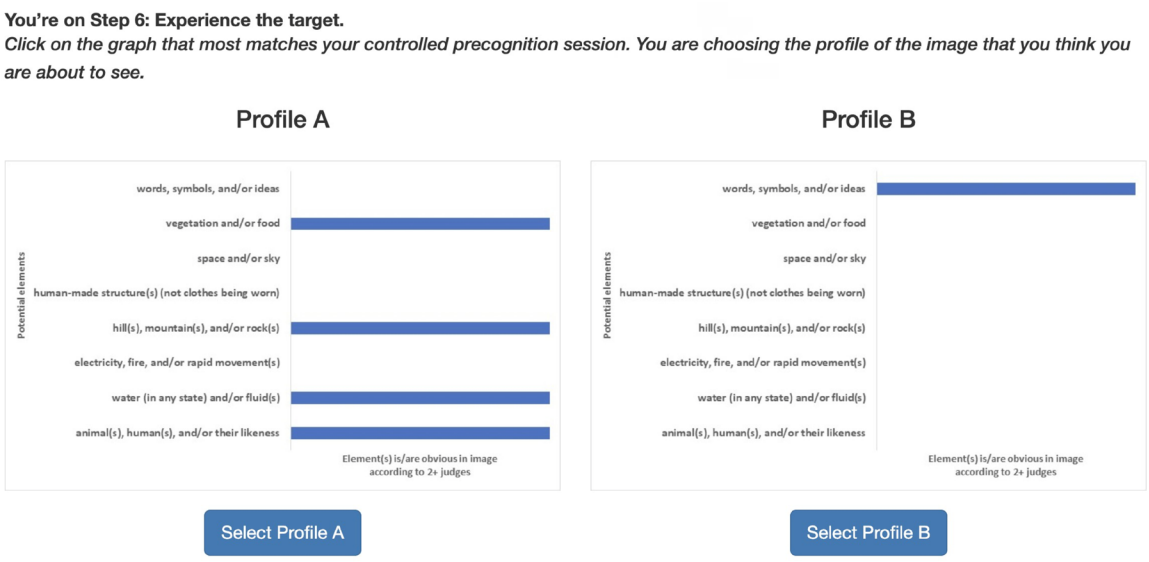


Figure 1

Screenshots showing the last two steps of the forced-choice PRV task (experiment 1). (Top) Participants chose between two graphs representing elements in each of two potential targets. (Bottom) After the graph was selected a final target was randomly selected and displayed.

ing the target image, users were invited to see their performance statistics. By design, there was no quick way to launch another trial except by going back to the starting screen. This enforced a relatively long (seconds-to-minutes) delay between trials in an attempt to emphasize the importance of a single trial and make participants more likely to take their time when they performed the forced-choice PRV task.

Random Number Generation

The website used a random number generator that drew from a truly random source. The tester portion of the site was developed using PHP 7.1, MySQL, HTML, javascript, and Bootstrap and was hosted by Tiger Technologies (Berkeley, CA). Random numbers used by the application for target selection were generated using calls to the PHP random_int() function, which returns cryptographically secure random numbers through the use of the Linux getrandom(2) system call. The server hosting the site files uses /dev/urandom which "gathers environmental noise from device drivers and other sources into an entropy pool" (https://linux.die.net/man/4/urandom; Tiger Technologies, personal communication, 2019) and passes all National Institute of Standards and Technology (NIST) tests of randomness (Cieslarová, 2018).

Data Analysis

Overall Approach. Forced-choice PRV data were analyzed in Microsoft Excel and Matlab 2018b. The threshold for statistical significance was set at $p = .05$, two-tailed. *Raw data are available upon request.*

Our analyses of the forced-choice PRV data comprised two parts. First, we treated participants who contributed to both batches (17% of the participants) as independent individuals contributing to the separate batches. Thus, we first analyzed the batches independently to examine overall accuracy, the relation between targets that were more likely to be correct vs. incorrect, performance over time, and the relation between the targets' level of "interestingness" and accuracy. After examining these factors, we combined scores across batches such that each user with a unique login had one accuracy score averaged across all trials they had performed within the data download periods. Within this larger combined dataset, which we believe was more appropriate for an individual differences analysis, we examined whether age or sex-at-birth influenced accuracy.

Interestingness Analysis. To determine whether the "interestingness" of targets was related to their likelihood of being correctly selected in the forced-choice PRV

task, we examined only targets presented to the participants in the second position on their device's screen. This is because in the first batch we found a strong bias toward choosing targets presented in the first position (see *Results*), and we wanted our analysis to primarily include targets that were selected for reasons other than screen position bias.

The first step in this analysis was to find targets that were consistently correctly selected – or incorrectly ignored (i. e., they were the correct target, but the other target was consistently selected). For each of the two data batches, we first calculated the ratio of incorrect to correct trials for each target presented in the second screen position. Then we identified the *n* targets that had a ratio of incorrect to correct trials of 0.4 or less (i. e., top-*n*-correct targets, *n* = 8 for the first batch, *n* = 5 for the second batch). Most-likely incorrect targets had to have an incorrect-to-correct ratio of 2.2 or greater; this more stringent standard was selected because targets presented in the second position were more often incorrect than correct due to bias towards selecting targets in the first position. We chose the *n* targets with the highest incorrect-to-correct ratios and called these the top-*n*-incorrect targets (Figures A-D for top-*n*-correct/-incorrect target sets for both batches; their features can be retrieved from https://tinyurl.com/MossbridgeEtAl2024Appendix). Once we identified these two sets of targets, we created *n* unique *interestingness questionnaires* designed to assess the "interestingness" of the targets. Across all *n* questionnaires, each top-*n*-correct target was presented alongside each of top-*n*-incorrect target in every possible pair combination (e.g., the first most incorrect target was presented with the first most correct target in one questionnaire, and presented with the second most correct target in another questionnaire, etc.).

To determine the interestingness of the targets, interestingness questionnaires were presented one at a time to different groups of paid workers on the Amazon Mechanical Turk website, with each group consisting of 10 paid workers. Their task was to look at each of the pairings of the images in the questionnaire they were viewing and ask themselves which of the two images was most interesting/intriguing, or whether they could not determine the answer. Workers were not told about the context of the original experiment or that it had anything to do with precognition, and each worker only saw one questionnaire. Each target was given *n* independent "interestingness" scores based on the data from each of the *n* questionnaires. Interestingness scores were calculated according to the proportion of workers who completed a given questionnaire who chose a target as most interesting (e.g., 1 = all workers thought it was the most interesting target; 0 = none of them thought it was most interesting; 0.5 = half thought it was most interesting).

We compared interestingness scores derived from the Amazon Mechanical Turk workers' rankings between the top-*n*-correct and top-*n*-incorrect target sets using *t*-tests. We compared these rankings in two ways: 1) across sets of responses derived from *n***n* questionnaires per each of the two target sets (the "across-responses" method), and 2) between averages of responses from the *n* questionnaires representing each target (the "averages" method). These analyses were pre-registered prior to downloading data in the second batch (as in Mossbridge & Radin, 2021), so both methods of analysis are reported here. Following pre-registration we noticed a large variation in rankings for each target depending on which other target they were paired with in the questionnaires, so we believe the across-responses method is probably most appropriate in that it captures the larger-than-expected intra-worker variability. Regardless, we report both analyses. Finally, we note that the pre-registration was completed prior to discovering that in the original analysis of the first batch we had failed to remove data from the website developer and the experimenter, so the numerical results presented here do not match the pre-registration document approved prior to discovering this oversight.

Results: Experiment 1

Overall Accuracy

Performance on the forced-choice precognitive remote-viewing (PRV) task for the 3,003 trials in the first data batch was at chance for both practice and test trials, with participants performing slightly but not significantly better on practice than test trials (practice trials: proportion correct 0.522, binomial test $p < .10$; test trials: proportion correct 0.50, binomial test $p > .99$; $\chi^2_{(1, N = 3003)} = .43$, $p > .51$). For the 2,429 total trials in the second batch, performance was significantly worse than chance for test trials, while performance on practice trials was no different from chance, following the same trend as in the first batch (practice trials: proportion correct 0.49, binomial test $p > .79$; test trials: proportion correct 0.45, binomial test $p < .01$). This difference between trial types was significant in the second batch ($\chi^2_{(1, N = 2429)} = 3.9$; $p < .05$; practice trials more accurate than test trials). The worse-than-chance performance on test trials, while small, indicated an expectation-opposing effect. In all further analyses, practice and test trials were not differentiated.

Data from the first batch revealed that participants exhibited a massive bias toward selecting Profile A (proportion of Profile A trials: 0.58, binomial test $p < 3x10^{-}$

[16]), a bias larger than any precognition effect we would expect to see. This bias was much reduced in the second batch (proportion of Profile A trials: 0.51, binomial test $p > .17$). This bias reduction could have resulted from a change in behavior resulting from a newsletter sent out by the first author after analyzing the first batch, in which subscribers to the website were informed about the overall bias we were seeing (see *Discussion*).

Target Interestingness

In the first batch, there were 8 targets presented as Profile B that were most likely to be correctly chosen as the target (also called top-8-correct), and in the second there were 5 (also called top-5-correct; see *Methods*: *Experiment 1;* and https://tinyurl.com/MossbridgeEtAl2024Appendix). In both batches, independent Amazon mTurk judges who were not informed about the purpose of the experiment rated the likely-correct images as more interesting when compared to the likely-incorrect images (first batch: $t_{12\ 6} = 3.49$, $p < .0007$ for across-responses, $t_{14} = 1.90$, $p < .08$ for averages, Figure 2a; second batch: $t_{48} = 5.03$, $p < .000008$ for across-responses, $t_{8} = 3.59$, $p < .008$ for averages, Figure 2b). Note that for the second batch, these analyses were pre-registered and thus the original interestingness effects were replicated with confirmatory analyses. Supporting these results, linear regression of the ratio of incorrect-to-correct trials for each target in the top-*n*-correct and top-*n*-incorrect data sets on interestingness ratings revealed significant or near-significant negative correlations (first batch: $r_{14} = -0.45$, $p < .08$; second batch: $r_{8} = -0.70$, $p < .03$), again indicating greater accuracy with increasing interestingness.

Further scrutiny of this effect seemed warranted because the images in the top-*n*-correct data sets contained different elements from those in the top-*n*-incorrect data sets. The design of the task allowed participants to see the elements of each potential target via the categorical graphs presented before the presentation of the actual target (Figure 1). Thus, it seems possible that these elements themselves were considered differentially interesting – causing participants to avoid selecting uninteresting targets, potentially producing the interestingness effect. To determine whether this kind of explicit foreknowledge explained the relation between interestingness and accuracy, we examined the interestingness scores of the five top-*n*-correct/-incorrect targets in both batches again, this time matched for content. First, we removed all targets containing text only and then performed the interestingness analyses again.

The results revealed a dominance of more interesting targets in the top-*n*-correct data sets, a significant difference for the second batch (first batch: $t_{102} = 1.51$, $p > .113$

2024, Vol. 4, No. 1, pp. 88–121

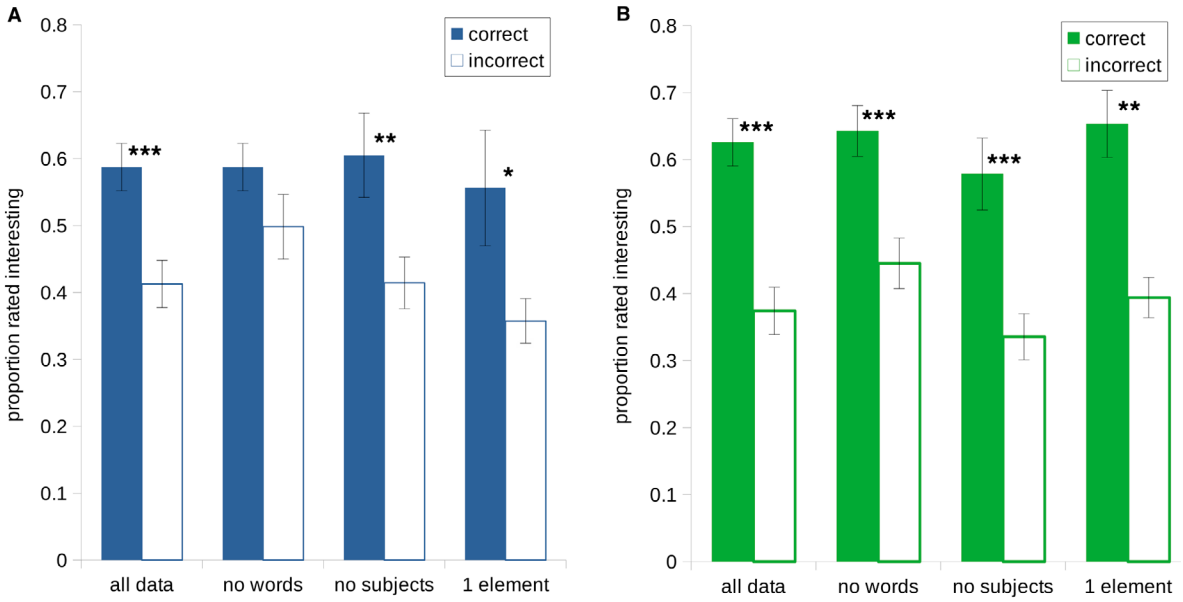


Figure 2

*Average interestingness scores across across all questionnaires for first (A) and second (B) data batches for the forced-choice PRV tass (experiment 1). Error bars indicate +/- 1 SEM between rankings. * p < .05, ** p < .01, *** p < .001*

for across-responses, $t_{11} = .85$, $p > .41$ for averages; second batch: $t_{33} = 3.62$, $p < .001$ for across-responses, $t_5 = 3.26$, $p < .03$ for averages). Next, we examined only targets excluding human or animal subjects and performed the analyses again. Again, the results overall revealed the same difference in interestingness between the top-*n*-correct and -incorrect targets (first batch: $t_{78} = 2.65$, $p < .01$ for across-responses, $t_8 = 1.32$, $p > .22$ for averages; second batch: $t_{33} = 3.99$, $p < .0004$ for across-responses, $t_5 = 3.16$, $p < .03$ for averages). Perhaps most importantly, we realized it was also possible that the number of elements in an image might explain the differences in interestingness scores between the top-*n*-correct and -incorrect targets, but comparing across targets that had only 1 element in the target revealed the same significant effect (first batch: $t_{70} = 2.57$, $p < .01$ for across-responses, $t_7 = 1.34$, $p > .22$ for averages; second batch: $t_{18} = 4.34$, $p < .004$ for across-responses, the averages method could not be calculated as only one target in the second batch had a single element). Noting that this last analysis was especially underpowered, we could not entirely put to rest the possibility that a number-of-elements bias caused the interestingness effect.

Performance Across Trials

To examine how performance might vary across trials among participants who performed multiple trials, we first calculated the average proportion of correct trials for those who performed at least five trials. For the first batch, mean proportion correct over the first five trials was significantly greater than chance as was performance on trials 3 and 4 (Figure 3; trial 3: $p < .002$, trial 4: $p < .05$ [binomial tests]; overall t-test on average first-batch performance versus chance: $t_{122} = 3.95$, $p < .0002$). However, there were no significant effects in the second batch for those who performed at least five trials (overall t-test on average second-batch performance versus chance: $p > .52$).

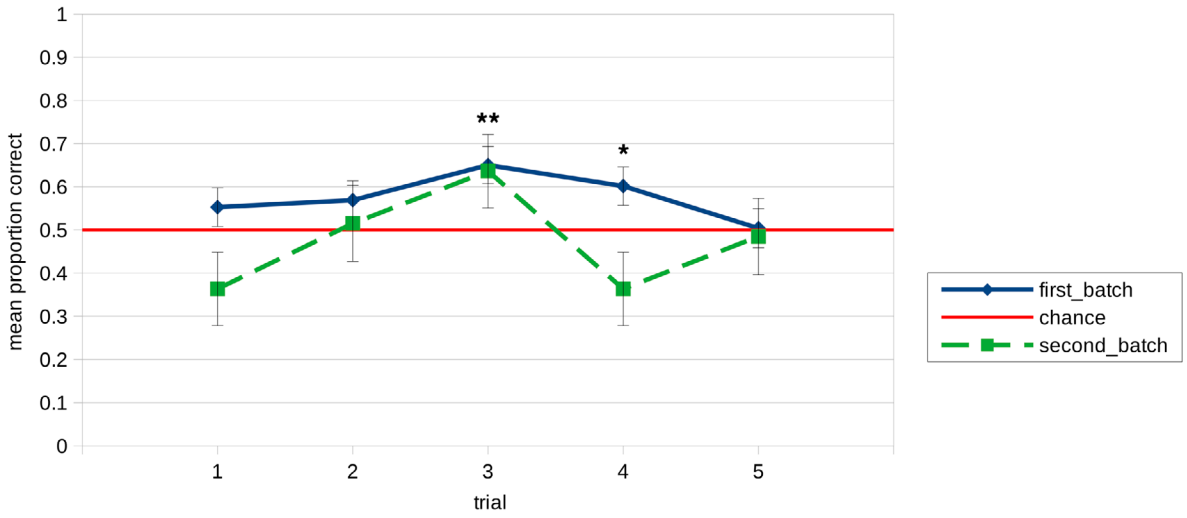


Figure 3

*Mean proportion correct for the forced-choice PRV task (experiment 1) versus trial number (solid blue line = first batch; dashed green line = second batch). Red line indicates chance performance. Error bars show +/- 1 SEM. * p < .05, ** p < .01.*

Influence of Age and Sex at Birth

To determine whether performance on the task was affected by age or sex at birth, we calculated the mean proportion of trials that were correct for those participants who reported either their age or sex at birth (unique users who reported sex: $n = 446$; reported age: $n = 445$, users who reported sex but not age = 2; users who reported age but not sex = 1). A linear regression between age and mean proportion correct revealed no significant relation, $r_{443} = -.003$, $p > .95$, so we examined any potential relation between sex at birth and mean proportion correct. We ignored reported gender in these analyses and instead focused on sex-at-birth, as only 33 individuals

did not report their sex-at-birth as being the same as their gender, and these individuals were split across multiple categories, providing too small a data set to perform individual difference analyses. We found that those who reported female as their sex-at-birth had mean proportion correct scores worse than chance (Figure 4a; $t_{287} = 2.68$, $p < .008$), while mean proportion correct scores from males at birth were no different from chance, $t_{156} = 0.63$, $p = .53$. These data show a significant difference favoring men (Figure 4a, first column; $t_{444} = 2.10$, $p < .04$).

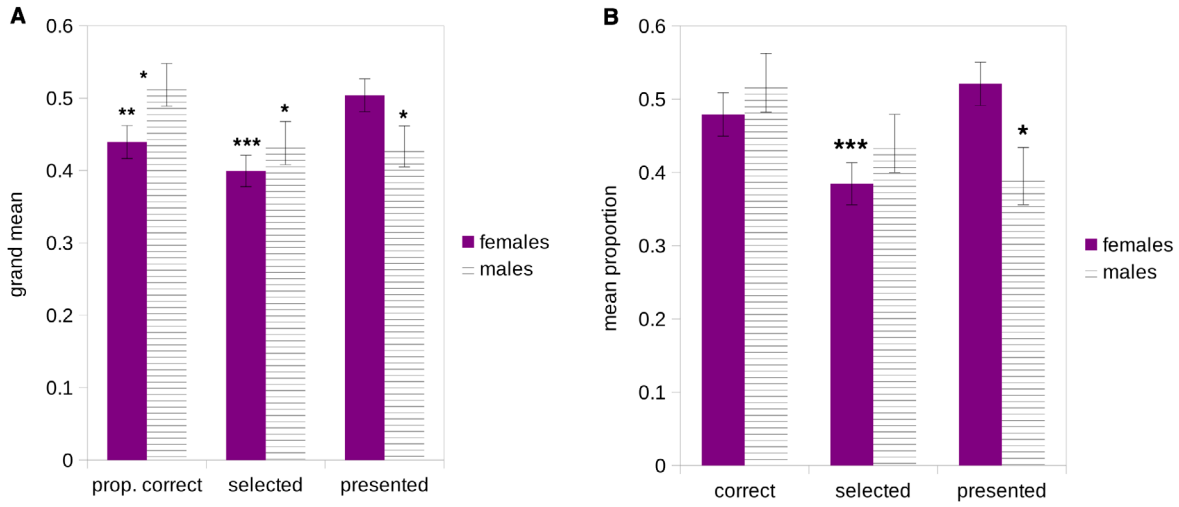


Figure 4

*Grand means of all trials performed by each participant (A) and mean proportion for first trials only from each participant (B) compared between participants who reported their sex at birth as female (purple solid) or male (horizontal hatching), combined across the two data batches. "Selected" = proportion of trials on which participants selected profile B; "Presented" = proportion of trials on which the target was profile B. *p < .05, ** p < .01, ***p < . 001. Error bars show +/- 1 SEM.*

To further understand this effect, we examined the relation between sex, target selection, and target presentation. There were no significant differences between males and females in terms of the mean proportion of trials on which each sex selected profile B as their choice (Figure 4a second column; grand means for females: 40%, males: 44%, ns, $p > .30$). However, we found a near-significant difference between the sexes in the mean proportion of time the software itself presented profile A versus profile B as the target ($t_{444} = 1.90$, $p < .058$). In trials performed by females, profile B was selected by the software as the target for a mean of 50% of the time, whereas in those

performed by males, profile B was selected by the software 43% of the time, lower than chance (t_{157} = 2.36, p < .02; Figure 4a, third column).

Discussion: Experiment 1

Experiment 1 was performed using a website designed for free and uncontrolled practice on the PRV task, a situation that resulted in a wide range of participation (e.g., each participant performed anywhere from 1 to 30+ trials within a time frame of 1 to 364 days). Further, a newsletter was sent to subscribers to the website, a subset of which were participants, to remind participants in Experiment 1 that the software was not biased even though it appeared that people preferred to select Profile A, which is likely why the Profile A bias disappeared in the second batch. Despite these uncontrolled aspects of experiment 1, it revealed several interesting effects that require further examination.

First, there were two weak "expectation-opposing effects" countering the presumed intention of the participants, a result apparently common for online forced-choice precognition tasks (Mossbridge & Radin, 2021). Although there was no significant target precognition in this task, there was, if anything, worse performance on test as compared to practice trials. This could suggest that performance anxiety might have hurt accuracy when participants knew that their results would be used to gauge their inclusion in the "Hall of Fame," which was the only difference between the tasks. Second, women were less accurate than men on this task overall and women's performance was, on average, worse than chance. Men's better performance on the forced-choice PRV task may have resulted from the random number generator presenting the target in the Profile A position more often for men, matching the Profile A bias. However, these results would need confirmatory replication to support a true gender difference.

Via a pre-registered confirmatory analysis, we found that performance in the second batch of trials on the forced-choice PRV task reflected the same dependence on independently-rated target interestingness as seen in the first batch (Figure 2). This effect could be caused by three different forms of participant knowledge about target content: words (text), subjects, and more than one element, but the results from control analyses did not support these ideas. The pre-registered confirmatory analysis of data in the second batch and the extra controls supported the idea that target interestingness may influence accuracy, though an additional experiment would be needed to draw firm conclusions.

Rationale and Approach for Experiment 2

The major effect found in experiment 1 was that greater target interestingness seemed to be related to increased accuracy. Our concerns about this result were addressed via experiment 2, in which participants did not have any information about the potential target prior to viewing it. A weaker effect found in experiment 1 was that gender seemed to be related to PRV performance in a complex way. We knew that any attempt to further examine these effects or our additional hypotheses related to participant mood would require a more controlled experiment. Further, we were interested in whether we would obtain significant target precognition using a free-response precognitive remote viewing task that more closely resembled the tasks used in applied precognition projects.

In experiment 2, our dependent variable was performance on an online two-minute, one-trial, free-response precognitive remote viewing task (the free-response PRV task) that had previously been shown to produce large effect sizes (Mossbridge et al., 2021b). Our independent variables were: anxiety score, feelings of unconditional love score, gender, exogenous reproductive hormonal status, and target interestingness. We tested five predictions, based on the results of experiment 1 and previous work in the field. All predictions were submitted to the granting agency before conducting the experiment, except where noted:

1. Free-response PRV accuracy will be better among participants who report higher versus lower feelings of unconditional love (confirmatory prediction based on Mossbridge et al., 2021b).
2. Free-response PRV accuracy will be better among participants who report lower versus higher feelings of anxiety (exploratory prediction).
3. Regardless of directionality, accuracy will be most closely related to higher feelings of unconditional love in women and to lower feelings of anxiety in men (exploratory prediction).
4. Women taking reproductive hormones will demonstrate above-chance accuracy and out-perform women not taking reproductive hormones (exploratory prediction). Note: In the plan originally submitted to the granting agency, we set out to test this hypothesis with pregnant women but IRB constraints required us to change the population to women taking reproductive hormones.
5. Targets rated as more interesting by independent judges will be more likely to be correctly described as future targets in the free-response PRV task (confirmatory prediction).

Methods: Experiment 2

Participants

All participants were screened, recruited, and paid via Amazon Mechanical Turk (mTurk). The screening and the experiment itself were only open to mTurk workers who read consent forms for each task and consented to be in the study (per Windbridge IRB 2020-BF/TILT-1208). Based on the pre-screening task, we accepted workers who lived in the United States, had the ability to perform the pencil-and-paper task required, were between the ages of 18 and 40, considered their gender as either men or women, and were not pregnant. Note that we excluded gender nonbinary people from this study because we needed enough participants to make a comparison between genders and we had a limited budget, not because of any lack of worthiness of asking similar questions about gender nonbinary people. Participants also answered questions about whether they were taking reproductive hormones, had given birth to a child in the last 6 months, and were breastfeeding. In all, 307 "non-bots" (presumed human participants, see below) completed the free-response PRV task (151 men, 125 women not taking hormones, 31 women taking hormones). We initially planned to recruit 150 men, 125 women not taking hormones, and 30 women taking hormones, but the experiment ended just after we reached these goals due to difficulties replacing apparent bots with actual respondents. Even so, the numbers remain similar to our goals. We identified 18 apparent bots; these were incomplete attempts at using non-human or AI means to complete the task quickly and get paid without doing the task (see Data analysis, below). When analyzing the data, we removed data from these apparent bots and replaced them with data from human participants.

Procedure and Free-response PRV Task

Before screened and consented participants performed the free-response PRV task, we asked them to complete a questionnaire with seven questions about their feelings of anxiety (Spitzer et al., 2006) and four questions about their feelings of unconditional love (Mossbridge et al., 2021a,b). We then presented participants with custom web software to assess free-response PRV ability (https://doi.myprecog.com/tilt01/). In this task, participants first watched a brief video about the task and then were asked to imagine they were opening a door that contained a future target behind that door. They were allowed two minutes (with an onscreen countdown) to describe what they imagined in words and/or images on a single sheet of paper. Finally, they were asked to take a photo of their paper (called a "transcript") and upload it. After the upload of

the transcript was complete, it could not be altered unless the experimenter asked for a re-upload because the original image was crumpled or blurry (two cases). In those cases, the original transcript was compared with the re-uploaded transcript and no discrepancies were found.

After transcript upload, the software used a true random-number generator drawing from network traffic (the same type as in experiment 1) to select one of 89 potential targets. This target was shown to the participant. In all, through the course of the experiment 86 targets (all but three of the targets) were shown at least once. To ensure that the participant observed the target with some level of conscious attention, the participant was then required to write a brief description of the target in an open text field.

Researcher Information

JM interacted with participants by creating the advertisement for the experiment, addressing rare questions via email from participants, and ensuring participants were paid for their work. Her belief that the experiment would produce positive results was at the strongest level (5 out of 5, on the same scale used in experiment 1). KC did not interact with participants except indirectly by analyzing their transcripts to detect bots, and her belief in the outcome of the experiment was neutral (3). MB interacted with participants indirectly via database queries as requested from JM and his belief in a positive outcome of the experiment was at the strongest level (5).

Free-response PRV Transcript Judging

We paid two experienced remote viewing judges, who did not know about the purpose of the experiment, to assess each participant's transcript as compared to two images: the actual future target the participant was shown and a non-target comparison image. The judges were not told which was the non-target or the actual target, and the transcript, target, and non-target images were presented to the judges in a randomized order sorted across trials via a hashed version of the participant ID (which did not reflect any independent or dependent variable). In anticipation of a potential position bias, the target was presented with equiprobability on the left or the right of the spreadsheet columns. Judges were not given each other's identities or contact information and thus did their judging separately.

The judges' task was to identify, for each transcript, which of the two images was most like it. With the two judges' responses for each participant, we calculated three types of judgments and scored them as follows: 1 (both judges agreed that the target image matched the transcript; chance expectation = .25), 0 (both judges agreed that the non-target image matched the transcript; chance expectation = .25), and -1 (the judges disagreed on which image was most similar to the transcript; chance expectation = .50). In this way, our scoring system discounted sketches that did not match either the target or non-target image well enough to gain judge agreement. Note that this is a more conservative scoring method than a version that just considers the proportion of judges that selected the actual target as matching a transcript.

Approximately 14 months after the original judging process, the same judges performed a control judging process on the same transcripts. In this process, the judges were given the transcripts in the original order, but the two images the judges were asked to compare with each transcript were not the original target and non-target. Instead, neither image was the target image – both were comparison images. The judges were told that we were doing a double-check on the judging and that they were to follow the same rules as in the first judging process. They were not told that the true target was not shown to them for any participant's trial. After the data were received from the judges, the judges were debriefed as to the purpose of this second judging process. The control judging process was scored according to the originally correct identity of the target (i.e., the first or second image presented to the judges) for each trial, effectively "yoking" the scoring to the original order of the targets shown to the judges more than a year before.

Anxiety and Unconditional Love Questionnaire

The anxiety portion of the questionnaire (Spitzer et al., 2006) presented the questions below, with possible responses of: not at all (0 points), several days (1 point), over half the days (2 points), and nearly every day (3 points). "Over the last 2 weeks, how often have you been bothered by the following problems? 1. Feeling nervous, anxious, or on edge. 2. Not being able to stop or control worrying. 3. Worrying too much about different things. 4. Trouble relaxing. 5. Being so restless that it's hard to sit still. 6. Becoming easily annoyed or irritable. 7. Feeling afraid as if something awful might happen." Summing the points from all questions produced the anxiety score.

The unconditional love portion of the questionnaire (Mossbridge et al., 2021a,b) first presented a definition of unconditional love followed by four questions with six possible responses: not applicable (not scored), never (1 point), a few times (2 points),

sometimes (3 points), often (4 points), a great deal (5 points). The definition was: "Unconditional love is the heartfelt benevolent desire that everyone and everything – ourselves, others, and all that exists in the universe – reaches their greatest possible fulfillment, whatever that may prove to be. This love is freely given, with no consideration of merit, with no strings attached, with no expectation of return, and it is a love that motivates supportive action in the one who loves." This definition was presented along with the questions, which were, "Given the definition you just read, to what extent do you feel unconditional love... 1. toward yourself? 2. toward other humans? 3. toward animals? 4. toward the computer on which you are completing this survey?" The calculated unconditional love score was the point value of the last question only, for reasons described below.

As expected, scores on the anxiety questionnaire were collinear across all seven questions and thus responses from the seven questions were summed to provide a single anxiety score (mean Pearson r = .67, range: .52-.89; all 307 participants included). The unconditional love measure had been previously validated in a laboratory experiment (Mossbridge et al., 2021a) and been successfully combined in a subsequent experiment with other state measures (Mossbridge et al., 2021b), so we expected its characteristics to remain the same for this experiment when used in combination with the anxiety questions. However, the four questions related to unconditional love revealed inconsistent relations with the anxiety questions. Across all participants, scores on the first three questions were essentially collinear with and negatively related to the anxiety questions and significantly positively correlated to each other, while scores on the last factor – unconditional love for the device on which the survey was being completed -- were not related to any other unconditional love or anxiety measure.

Thus, we used this final item alone as our measure for feelings of unconditional love, under the assumption that the collinearity of the other factors made this single question about unconditional love the only question that uniquely revealed feelings of unconditional love rather than the inverse of anxiety in this online context. Specifically, the average correlation of scores on the first three unconditional love questions with scores on the "device" question was r = .06, range: .03-.07, while the average correlation on the first three unconditional love questions with one another was r = .30, range: .15-.42. The average correlation of scores on the first three unconditional love questions with the average anxiety scores was r = .21, range: -.47 -.02, and the correlation of scores on the "device" question with the average anxiety scores was r = .05.

Target Set

The target image set was a subset of the targets at remoteviewed.com that were used in a previous experiment (Mossbridge et al., 2021b), following the removal of several targets that could be interpreted as violent or destructive. The non-target image presented with each target to the judges was also from the same target set and was determined for each target prior to collecting data. This comparison-target determination was based on pairings designed to accentuate differences between the contents, moods, and meanings of the images (as in Mossbridge et al., 2021b).

Data Analysis

Uploaded transcripts from all participants were judged as to whether they were from a bot according to three tests: 1) Does the writing in the transcript look like a type-set font? 2) Is the same transcript uploaded by more than one participant? 3) Does the transcript make no sense according to the instructions (instead of a photo of a piece of paper with a sketch and/or words on it, was it a photograph taken from the internet)? If any test gave a YES response from either of the two examiners (JM and KC), it was considered a bot and rejected.

To understand whether participants were fully attending to their answers on the survey questions, we examined consistency across survey responses. Seventy-six participants gave the same answer for all seven anxiety questions within the survey. This seemed unlikely to reflect attentive responding to the survey, especially given that the average amount of time spent completing the survey was significantly lower for these participants than those who responded with more variety. Thus, while we analyzed data from all 307 participants for any prediction not related to the anxiety and unconditional love survey, we removed these 76 "survey-inattentive" participants from our analyses for the three predictions related to that survey (predictions 1-3), because we could not be sure that their responses reflected their emotional states accurately.

To test the confirmatory prediction from experiment 1 that targets rated by independent judges as more interesting were also more likely to be accurately described in the PRV task, we asked other Amazon mTurk workers who had not participated in this experiment to rate all 86 targets shown to participants according to how intriguing/interesting they were. We presented two images at a time in the same pairs that were presented to the judges. The workers' task was to choose which image was most interesting or if this could not be determined (this was rare). We then created an in-

terestingness score based on the sum of all a given target's determinations as most interesting divided by the total number of determinations excluding "could not be determined" responses. To investigate any relation between interestingness and accuracy, we calculated each target's "accuracy score" as the sum of scores for each target across participants who viewed that target. Note that by calculating the sum rather than the average, we weighted more heavily the scores from targets presented more often and that scored more consistently (e.g., a target presented 3 times and correct all 3 times will have a score of 3, while one presented 1 time and correct that time will have a score of 1). All statistical analyses used null hypothesis significance testing with alpha set at 0.05 and two-tailed tests where appropriate.

Results: Experiment 2

On this task, significant precognition of the target image was evident. To examine precognition, we used binomial tests to compare actual performance to chance expectation (Figure 5). Overall, there were significantly more instances of matches to the target images than expected by chance (35% vs 25% chance expectation [CE], $p < .0002$, $h = .22$) and more instances of matches to the non-target images than expected by chance (42% vs. 25% CE, $p < .00001$, $h = .36$). There were fewer judge disagreements than expected by chance (23% vs. 50% CE, $p < .00001$, $h = -.57$). The control judging process, in which judges were never shown the actual target displayed to the participant, produced a significantly different pattern. Specifically, the number of matches to the yoked "targets" was indistinguishable from chance (26% vs 25% CE, $p > .69$, $h = .03$), while the number of matches to the yoked "non-targets" was almost identical to the results from the original method and therefore greater than chance (41% vs 25% CE, $p < .00001$, $h = .34$). The number of no-agreement trials remained significantly lower than chance (33% vs 50% CE, $p < .00001$, $h = .35$). A Chi-square analysis revealed that the overall score distributions differed between the original and control judging processes ($\chi^2_{(1, N = 614)} = 8.5$, $p < .015$).

The data supported prediction 1, showing a positive relation between free-response PRV accuracy and feelings of unconditional love, but not prediction 2, that anxiety is negatively related to accuracy. We performed a median split with an excluded median on both unconditional love and anxiety scores for survey-attentive participants. Comparing accuracy between these extremes revealed that greater feelings of unconditional love were associated with better accuracy on the task ($t_{192} = 2.50$, $p < .01$; $\chi^2_{(2, N = 194)} = 10.0$, $p < .007$). In contrast, greater feelings of anxiety were marginally associated with *greater* accuracy on the task, weakly opposing prediction 2 ($t_{217} = 1.990$, $p < .048$; $\chi^2_{(2, N = 219)} = 4.2$, $p > .12$).

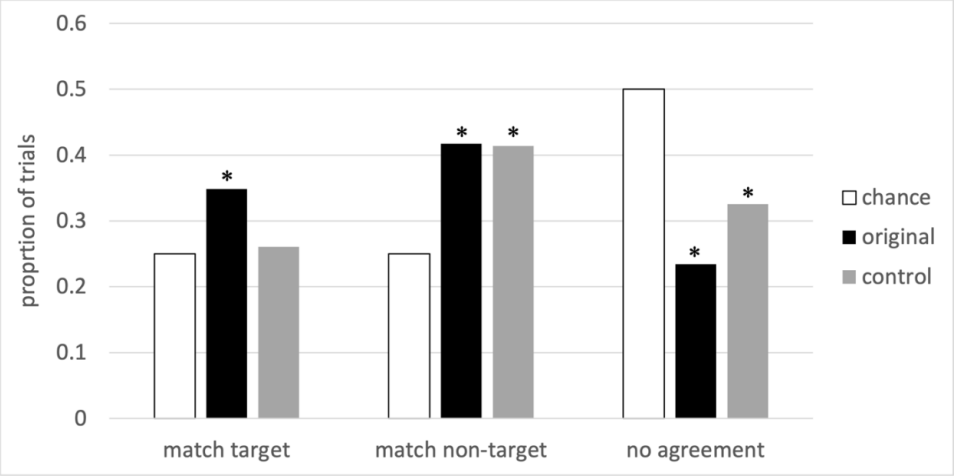


Figure 5

*Proportion of transcripts that matched the target (score = !), matched the non-target (score = 0), or on which there was no agreement (score = -1). Parameter is chance expectation (empty), original judgment method (black) and control judging method (gray). * p < .05.*

We found no gender difference in the relation between unconditional love and accuracy, but did find a gender difference in accuracy related to anxiety (prediction 3). Unconditional love was quantitatively associated with better accuracy in both men and women in the survey-attentive group, but there were no significant effects when data from each gender were analyzed separately (women: $t_{44} = 1.56$ $p > .12$; $\chi^2_{(2, N = 46)} = 1.7$, $p > .42$; men: $t_{55} = 1.59$, $p > .12$; $\chi^2_{(2, N = 57)} = 2.8$, $p > .42$). In contrast, anxiety in the survey-attentive group was associated with *better* free-response PRV accuracy in women according to the results of a Chi-squared test ($t_{114} = 1.93$, $p < .057$; $\chi^2_{(2, N = 116)} = 8.0$, $p < .02$), but not men ($t_{93} = .755$, $p > .45$; $\chi^2_{(2, N = 95)} = 2.80$, $p < .10$). Men with high anxiety scores showed a different pattern of performance than women across the three score types ($\chi^2_{(2, N = 104)} = 8.18$, $p < .02$), with the difference being most obvious for instances in which both judges agreed that transcripts matched the targets (score of 1; women: low anxiety = 12 trials, high anxiety = 32 trials; men: low anxiety = 21 trials, high anxiety = 8 trials).

Any relation between free-response PRV accuracy and exogenous reproductive hormones (prediction 4) was not evident in our data. Women taking reproductive hormones, including birth control pills, had above-chance free-response PRV accuracy as predicted, but did not significantly out-perform women not taking hormones ($t_{154} = .82$, $p > .41$; $\chi^2_{(2, N = 156)} = .68$, $p > .71$).

Finally, confirming prediction 5, we found that targets more likely to be correctly described were also considered more interesting by independent judges. Interesting-

ness scores for each target were determined by mTurk workers who were not included among our participants and not informed about the experiment or its hypotheses. These scores correlated positively with target accuracy (Figure 6; r_{84} = .21, p < .05). Extreme inaccuracy scores (target accuracy score = -3) were associated with two targets and extreme accuracy scores (target accuracy score = 3) were associated with seven targets. The extreme-scoring targets (Figures E, F) can be retrieved from https://tinyurl.com/MossbridgeEtAl2024Appendix.

Figure 6

Results from free-response PRV task (experiment 2) showing the relation between target accuracy and interestingness. Each dot is a target (N = 86); the purple line shows the linear regression.

Discussion: Experiment 2

In experiment 2, we imposed a time constraint on responses and required a free-response protocol using independent judges to assess PRV accuracy, and we found overall significant target precognition. Judges were more likely than expected by chance to agree on their match of a transcript to either of the two images shown to them on each trial. This result could indicate successful precognition of the target and non-target images, or the result could indicate poor precognition across the board. In the latter case, the argument would be that the transcripts were so non-reflective of the future target that any similarity between a given transcript and each of the two comparison images was agreed upon by the judges because it was the only possible connection between the transcript and either of the images. We used the control judging process to disentangle these possibilities.

If the original effects were entirely due to good precognition, then the control judging process in which the targets were never shown to the judges with the correct transcripts would reveal results indistinguishable from chance. However, if the original effects were entirely due to poor precognition, then the control judging process would reveal results indistinguishable from the original pattern. The control judging process revealed a cross between these two, in which the proportion of agreed-upon target matches was indistinguishable from chance, but the proportion of agreed-upon non-target matches was indistinguishable from the original effect. Further, the results from the two judging methods were significantly different overall. Our interpretation of these results is that the original results revealed good precognition of the target image. However, it remains difficult to tease out whether the non-target images shown to the judges in both scoring processes were also precognized at a rate above chance. Perhaps this interpretational ambiguity is a good argument for single-comparison judging methods in PRV experiments, like the calculation of a figure-of-merit for each transcript (May, 2014).

In terms of our trait-, state- and target-interestingness predictions, the results from experiment 2 confirmed our prediction that feelings of unconditional love would be positively associated with free-response PRV accuracy (prediction 1). Meanwhile, the results provided contrasting evidence for our exploratory prediction that feelings of anxiety would be negatively associated with free-response PRV accuracy (prediction 2), instead showing that feelings of anxiety were at least partially associated with *greater* accuracy, particularly for women. Our data did not support the exact prediction that higher feelings of unconditional love would be related to higher accuracy mostly for women and higher anxiety would be related to lower accuracy mostly for men (prediction 3), but the data did support a gender interaction with anxiety. While women taking exogenous reproductive hormones performed significantly better than chance on the free-response PRV task, this effect was no different from that of women not taking hormones, so prediction 4 was not supported. Finally, prediction 5 was confirmed. Targets rated by independent judges as more interesting were also more likely to be consistently accurately described in the participants' free-response PRV transcripts.

The fact that free-response PRV accuracy correlated with feelings of unconditional love confirmed previous exploratory analyses (Mossbridge et al., 2021b), but accuracy was weakly correlated with greater feelings of anxiety. The anxiety result seems counterintuitive, but the relation was weak and therefore potentially spurious; greater effort in understanding this result would be necessary if it were to be replicated. In contrast, the unconditional love result is consistent with an established relation

between positive or self-transcendent feelings and better performance on psi tasks. What we conclude from the present results is that our measure of unconditional love seems to have a gender-independent, positive relation with accuracy on the free-response PRV task we administered.

Overall Discussion and Recommendations for Future Research

One key finding is that an online free-response PRV task can produce significant results without pre-screening for trained remote viewers. Further, both experiments suggest that PRV accuracy in both forced-choice and free-response tasks is affected by target interestingness. The finding that interestingness was associated with target accuracy scores in three analyses across two experiments suggests that whatever psi capacity is being used to perform PRV tasks, it is behaving similarly to visual working memory and visual attention. Specifically, images that are more interesting, memorable, or salient are recalled more easily (Fine & Minnery, 2009) and our attention is drawn to them over other targets (Wang & Theeuwes, 2020).

In contrast to the forced-choice PRV task used in experiment 1 where we found an accuracy difference for sex-at-birth, for the free-response PRV task in experiment 2 there was no gender difference in accuracy except a weak and indirect relation with anxiety. Although sex-at-birth and gender are not the same thing and there were major differences between the two experiments, it is worth considering that our results echo gender effects found in forced-choice precognition tasks (Bierman & Scholte, 2002; Lobach, 2009; Mossbridge, 2017; Mossbridge, Tressoldi & Utts, 2012; Radin & Lobach, 2007; Wittmann et al., 2021; Mossbridge & Radin, 2021) as well as the *lack* of such effects reported anecdotally among free-response PRV trainers and students.

Based on the present experiments, our recommendations for future researchers studying precognition and related phenomena are:

1. Avoid using a pool for which some participants can choose to communicate with the experimenter and gain insight into the experiment. This could bias or de-bias their behavior in ways inconsistent across participants, increasing noise in the data and making it more difficult to observe effects.
2. Use a time-controlled, single-trial, online free-response precognitive remote viewing task to examine research questions about precognition. The effect sizes are robust, overall gender effects are not apparent, and it is not too difficult to find participants if you have the funding to pay them. Make sure you have a system for eliminating bots from your sample.

3. For your judging process, consider a single-target judge-free method (May, 2014) to avoid ambiguities related to non-target precognition.
4. Try to hold target interestingness steady across the target pool, and consider measuring unconditional love or other self-transcendent states in your participants.
5. Build on your results by creating a useful AI-based predictive model of real-world events at least partially informed by precognitive remote viewing transcripts weighted according to each viewer's particular skills, characteristics, and vulnerabilities.

References

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407-425. doi:10.1037/a0021524

Bem, D., Tressoldi, P., Rabeyron, T., & Duggan, M. (2015). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events (version 2). *F1000 Research, 4*, 1188. doi:10.12688/f1000research.7177.1

Bierman, D., and Scholte, H. (2002). A fMRI brain imaging study of presentiment. *Journal of the International Society of Life Information Sciences 20*(2), 380–388.

Cardeña, E. (2014). A call for an open, informed study of all aspects of consciousness. *Frontiers in Human Neuroscience*, *8*, 17. doi:10.3389/fnhum.2014.00017

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

Carpenter, J. C. (2005). Implicit measures of participants' experiences in the Ganzfeld: Confirmation of previous relationships in a new sample. In, *Proceedings of Presented Papers*: *The Parapsychological Association 48th Annual Convention* (pp 36-45).

Carpenter, J., Simmonds-Moore, C., Moore, S., & Carpenter, F. (2021). ESP contributes to the unconscious formation of preferences. *Journal of Parapsychology*, *85*(1), 28-53. doi10.30891/jopar.2021.01.06

Cieslarová, R. (2018). *Analysis of the Linux random number generator in virtualized environment.* Masaryk University Faculty of Informatics.

Evrard, R., & Ventola, A. M. (2018). [Whole issue]. *Mindfield*: *The Bulletin of the Parapsychological Association, 10*(3), 86-127.

Fine, M. S., & Minnery, B. S. (2009). Visual salience affects performance in a working memory task. *Journal of Neuroscience*, *29*(25), 8016-8021. Ddoi10.1523/JNEUROSCI.5503-08.2009

Hitchman, G. A., Roe, C. A., & Sherwood, S. J. (2012). A reexamination of nonintentional precognition with openness to experience, creativity, psi beliefs, and luck beliefs as predictors of success. *Journal of Parapsychology*, *76*(1), 109-145.

Honorton, C., Ferrari, D. G., & Bem, D. J. (1998). Extraversion and ESP performance: A meta-analysis and a new confirmation. *Journal of Parapsychology 62*(3), 255-276.

Houran, J., Lange, R., & Hooper, D. (2018). Cross-examining the case for precognition: Comment on Mossbridge and Radin (2018). *Psychology of Consciousness*: *Theory, Research, and Practice, 5*(1), 98-109. doi:10.1037/cns0000126

Katz, D., Smith, N., Bulgatz, M., Graff, D. & 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.

Katz, D.L., Grgic, I., Tressoldi, P., & Fendley, T. (2021). Associative remote viewing projects: Assessing rater reliability and factors affecting successful predictions. *Journal of the Society for Psychical Research, 85*(2), 65-90.

Kolodziejzyk, G. (2013). Greg Kolodziejzyk's 13-year associative remote viewing experiment results. *Journal of Parapsychology, 76*, 349-368.

Krippner, S., Saunders, D. T., Morgan, A., & Quan, A. (2019). Remote viewing of concealed target pictures under light and dark conditions. *Explore*, *15*(1), 27-37. doi:10.1016/j.explore.2018.07.001

Lobach, E. (2009). Presentiment research: Past, present and future. In C.A. Roe, L. Coly, & W. Kramer (Eds) *Utrecht II*: *Charting the future of parapsychology* (pp. 22-45). Parapsychology Foundation.

May, E. (2014). Advances in anomalous cognition analysis: A judge-free and accurate confidence-calling technique. In E. May & S. B. Marwaha (Eds.), *Anomalous cognition*: *Remote viewing research and theory* (pp. 80–88). McFarland.

Miao, C., Humphrey, R. doi., & Qian, S. (2018). The relationship between emotional intelligence and trait mindfulness: A meta-analytic review. *Personality and Individual Differences*, *135*, 101-107.

Mossbridge, J. A. (2023a). Precognition at the boundaries: An empirical review and theoretical discussion. *Journal of Anomalous Experience and Cognition*, *3*(1), 5-41. doi:10.31156/jaex.24216

Mossbridge, J. A. (2023b). Sam Knight details how not to build a premonitions bureau. *Journal of Anomalous Experience and Cognition, 3*(2), 369-375. doi: 10.31156/jaex.25150

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

Mossbridge, J. A., & Radin, D. (2018b). Plausibility, statistical interpretations, physical mechanisms and a new outlook: Response to commentaries on a precognition review. *Psychology of Consciousness*: *Theory, Research, and Practice, 5*(1), 110-116. doi:10.1037/cns0000152

Mossbridge, J. A., Nisam, M., & Crabtree, A. (2021b). Can hypnotic suggestion induce feelings of unconditional love and supernormal performance? *Spirituality in Clinical Practice*, *8*(1), 30-50. doi:10.1037/scp0000239

Mossbridge, J., & Radin, D. (2021). Psi performance as a function of demographic and personality factors in smartphone-based tests: Using a "SEARCH" approach. *Journal of Anomalous Experience and Cognition*, *1*(1-2), 78-113. doi:10.31156/jaex.23419

Mossbridge, J., Johnson, K., Washburn, P., Williams, A., & Sapiro, M. (2021a). Smartphone time machine: Tech-supported improvements in time perspective and wellbeing measures. *Frontiers in Psychology*, 5012. doi:10.3389/fpsyg.2021.744209

Mossbridge, J., Tressoldi, P. E., & Utts, J. (2012). Predictive physiological anticipation preceding seemingly unpredictable stimuli: a meta-analysis. *Frontiers in Psychology*, *3*, 390. doi:10.3389/fpsyg.2012.00390

Mossbridge, J.A. & Vivanco, J. (2022). Can we heal the earth with intuition? *Aperture 34,* 44-66.

Mumford, M. D., Rose, A. M., & Goshin, D. A. (1995). *An evaluation of remote viewing: Research and applications*. American Institutes for Research.

Nelson, R. D., Dunne, B. J., Dobyns, Y. H., & Jahn, R. G. (1996). Precognitive remote perception: Replication of remote viewing. *Journal of Scientific Exploration*, *10*(1), 109-110.

Palmer, J., & Carpenter, J. C. (1998). Comments on the extraversion-ESP meta-analysis by Honorton, Ferrari and Bem. *Journal of Parapsychology 62*(3), 277-282.

Radin, D., & Lobach, E. (2007). Toward understanding the placebo effect: Investigating a possible retrocausal factor. *Journal of Alternative and Complementary Medicine, 13*, 733–739. doi:10.1089/acm.2006.6243

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

Roney-Dougal, S. M., & Solfvin, J. (2011). Exploring the relationship between Tibetan meditation attainment and precognition. *Journal of Scientific Exploration, 25*(1), 29-46.

Roney-Dougal, S. M., Solfvin, J., & Fox, J. (2008). An exploration of degree of meditation attainment in relation to psychic awareness with Tibetan Buddhists. *Journal of Scientific Exploration, 22*(2), 161-78.

Schwartz, S. A. (2007). *Opening to the infinite: The art and science of nonlocal awareness*. Nemoseen Media.

Schwartz, S. A. (2021). 2060 Visions of the future. *Venture Inward, 38,* 36-40.

Schwarzkopf, D. S. (2018). On the plausibility of scientific hypotheses: Commentary on Mossbridge and Radin (2018). P*sychology of Consciousness: Theory, Research, and Practice, 5*(1), 94-97. doi:10.1037/cns0000125

Smith, C. C., Laham, D., & Moddel, J. (2014). Stock market prediction using associative remote viewing by inexperienced remote viewers. *Journal of Scientific Exploration, 28*(1), 7-16.

Smith, P. (2005). *Reading the enemy's mind: Inside Star Gate: America's psychic espionage program*. Macmillan.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Lowe, B. (2006). A brief measure for assessing generalized anxiety disorder. *Archives of Internal Medicine, 166,* 1092-1097. doi:10.1001/archinte.166.10.1092

Tait, A. (2019, Sept. 29). Psychic future: What's next for the 'precog economy'? *The Guardian*. https://doi.theguardian.com/global/2019/sep/29/psychic-future-what-next-for-the-precognition-economy

Targ, R. (2019). What do we know about psi? The first decade of remote-viewing research and operations at Stanford Research Institute. *Journal of Scientific Exploration, 33*(3), 569-592. doi:10.31275/2019/1669

Utts, J. (1995). An assessment of the evidence for psychic functioning. *Journal of Parapsychology, 59*, 289-320.

Varvoglis, M., Bancel, P. A., Bailly, J. P., Boban, J., & si Ahmed, D. (2019). The selfield: Optimizing precognition research 4. *Journal of Parapsychology*, *83*(1), 13-24. doi:10.30891/jopar.2019.01.02

Wargo, E. (2015). Time's taboos: Dirty thoughts on systems, syntropy, and psi. *Syntropy 1,* 57-67.

Wang, B., & Theeuwes, J. (2020). Salience determines attentional orienting in visual selection. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(10), 1051. doi:10.1037/xhp0000796

Watt, C., Dawson, E., Tullo, A., Pooley, A., & Rice, H. (2020). Testing precognition and alterations of consciousness with selected participants in the ganzfeld. *Journal of Parapsychology*, *84*(1), 21-37. doi:10.30891/jopar2020.01.05

Wittmann, M., Scheck, F., Feldmann, J., Glaesmann, A., Mossbridge, J., & Bem, D. (2021). The German version of a retroactive priming task shows mixed effects. *Psychology of Consciousness: Theory, Research, and Practice*. Advance online publication. doi: 10.1037/cns0000296

Zdrenka, M., & Wilson, M. S. (2017). Individual difference correlates of psi performance in forced-choice precognition experiments: A meta-analysis (1945-2016). *Journal of Parapsychology, 81*(1), 9-32.

Paramètres d'Etat, de Trait et de Cible Associés à la Performance dans Deux Tests en Ligne de Vision à Distance Précognitive

Julia Mossbridge Kirsten Cameron Mark Boccuzzi

Résumé: *Objectif*:Pour mieux caractériser les relations entre la précision dans les tâches de visualisation à distance précognitive (PRV) et les paramètres potentiellement pertinents de trait, d'état et de cible, nous avons recueilli des données PRV dans deux expériences en ligne et examiné la performance selon : le sexe de naissance, le genre, l'âge, l'anxiété, l'amour inconditionnel, et l'intérêt pour la cible. *Méthode*: L'expérience 1 a utilisé un choix forcé, un temps non contrôlé, une tâche de PRV auto-jugée pour laquelle 682 participants non rémunérés ont contribué à un total de 5432 essais. L'expérience 2 a utilisé une tâche de PRV à réponse libre, à temps contrôlé et à jugement indépendant, pour laquelle 307 participants rémunérés ont contribué à un seul essai chacun. Dans les deux cas, les participants n'ont pas été présélectionnés pour leur capacité de précognition. *Résultats*: Dans l'expérience 1 (tâche de PRV à choix forcé), il n'y a pas eu de précognition significative de la cible, ni d'effet de l'âge sur la performance PRV, mais nous avons trouvé un effet complexe concernant le sexe de naissance. Nous avons également constaté que les cibles les plus susceptibles d'être correctement prédites étaient également plus susceptibles d'être jugées intéressantes que les cibles les plus susceptibles d'être incorrectement prédites; une analyse préliminaire a confirmé cet effet. Dans l'expérience 2 (tâche PRV à réponse libre), nous avons constaté une précognition significative des cibles,

aucun effet de l'âge sur la performance, et un effet faible et indirect du sexe de naissance. Les sentiments d'amour inconditionnel et d'anxiété étaient tous deux corrélés à une plus grande précision dans l'expérience 2. Une fois de plus, l'intérêt de la cible était positivement lié à la précision. *Conclusion*: Ces résultats suggèrent que la performance dans les tâches de PRV est liée à l'état émotionnel des participants et à leur intérêt pour la cible, et que les caractéristiques de la tâche atténuent la performance globale. Sur la base de ces observations, nous formulons des recommandations pour de futures recherches.

Translation into French by Antoine Bioy, Ph. D.

Zustands-, Eigenschafts- und Zielparameter in Verbindung mit der Genauigkeit in zwei Online-Tests für präkognitives Remote Viewing

Julia Mossbridge Kirsten Cameron Mark Boccuzzi

Zusammenfassung: Zielsetzung: Um die Beziehungen zwischen der Genauigkeit bei präkognitiven Remote-Viewing-Aufgaben (PRV) und potenziell relevanten Merkmals-, Zustands- und Zielparametern besser zu charakterisieren, sammelten wir PRV-Daten in zwei Online-Experimenten und untersuchten die Genauigkeit in Abhängigkeit von: Geschlecht bei der Geburt, Geschlecht, Alter, Angst, bedingungslose Liebe und Interessantheit des Ziels. Methode: In Experiment 1 wurde eine Forced-Choice-Aufgabe mit unkontrollierter Zeit und selbsteingeschätzter PRV-Aufgabe verwendet, an der sich 682 unbezahlte Teilnehmer mit insgesamt 5.432 Einzelversuchen beteiligten. In Experiment 2 wurde eine zeitlich kontrollierte, unabhängig bewertete PRV-Aufgabe mit freier Antwortmöglichkeit verwendet, zu der 307 bezahlte Teilnehmer jeweils einen einzigen Versuch beisteuerten. In beiden Experimenten wurde keine Vorauswahl der Teilnehmer in Bezug auf ihre Präkognitionsfähigkeit getroffen. Ergebnisse: In Experiment 1 (forced-choice bei PRV-Aufgabe) gab es keine signifikante Zielpräkognition und keinen Effekt des Alters auf die PRV-Leistung, aber wir fanden einen komplexen Effekt des Geschlechts bei der Geburt. Wir fanden außerdem heraus, dass Targets, die am wahrscheinlichsten richtig vorhergesagt wurden, auch eher als interessant eingestuft wurden als Targets, die am wahrscheinlichsten falsch vorhergesagt wurden; eine vorab registrierte Analyse bestätigte diesen Effekt. In Experiment 2 (PRV-Aufgabe mit freier Antwortmöglichkeit) fanden wir eine signifikante Präkognition des Targets, keinen Einfluss des Alters auf die Leistung und einen schwachen und indirekten Einfluss des Geschlechts. Gefühle von bedingungsloser Liebe und Angst waren in Experiment 2 beide mit höherer Genauigkeit korreliert. Auch hier war das Interesse am Ziel positiv mit der Genauigkeit verbunden. Schlussfolgerung: Diese Ergebnisse deuten darauf hin, dass die Genauigkeit bei PRV-Aufgaben mit dem emotionalen Zustand der Teilnehmer und dem Interesse am Ziel zusammenhängt und dass die Merkmale der Aufgaben die Gesamtleistung abschwächen. Auf der Grundlage dieser Beobachtungen geben wir Empfehlungen für die zukünftige Forschung.

Translation into German by Eberhard Bauer, Ph. D.

Parâmetros de Estado, Traço e Alvo Associados à Precisão em Dois Testes Online de Visualização Remota Precognitiva

Julia Mossbridge Kirsten Cameron Mark Boccuzzi

Resumo: Objetivo: Para caracterizar melhor as relações entre a precisão em tarefas de visualização remota precognitiva (PRV em inglês) e parâmetros de traço, estado e alvo potencialmente relevantes, coletamos dados de PRV em dois experimentos online e examinamos sua precisão em relação a: sexo de nascimento, gênero, idade, ansiedade, amor incondicional e atratividade do alvo. Método: O Experimento 1 utilizou uma tarefa de PRV de escolha-forçada, sem controle de tempo, autoavaliada, para a qual 682 participantes não-remunerados contribuíram com um total de 5.432 tentativas. O Experimento 2 utilizou uma tarefa de PRV de resposta-livre, com tempo controlado, avaliada de forma independente, para a qual 307 participantes remunerados contribuíram com uma única tentativa cada. Em ambos os casos, os participantes não foram pré-selecionados quanto à habilidade de precognição. Resultados: No Experimento 1 (tarefa de PRV de escolha-forçada), não houve precognição significativa acerca do alvo e nenhum efeito de idade no desempenho do PRV, mas encontramos um efeito complexo relacionado ao sexo de nascimento. Também descobrimos que os alvos mais propensos a serem previstos corretamente eram mais propensos igualmente a serem julgados como interessantes em comparação com os alvos mais propensos a serem previstos incorretamente; uma análise pré-registrada confirmou esse efeito. No Experimento 2 (tarefa de PRV de resposta-livre), encontramos uma precognição significativa acerca do alvo, nenhum efeito da idade no desempenho e um efeito fraco e indireto do gênero. Sentimentos de amor incondicional e ansiedade estavam ambos correlacionados com uma maior precisão no Experimento 2. Novamente, a atratividade do alvo estava positivamente relacionada à precisão. Conclusão: Esses resultados sugerem que a precisão em tarefas de PRV está relacionada ao estado emocional dos participantes e a atratividade do alvo, e que as características da tarefa mitigam seu desempenho geral. Concluímos com recomendações para pesquisas futuras baseadas nessas observações.

Translation into Portuguese by Antônio Lima, Ph. D.

Parámetros de Estado, Rasgo, y Objetivo Asociados a Acertar en Dos Pruebas en la Red de Visión Remota Precognitiva

Julia Mossbridge Kirsten Cameron Mark Boccuzzi

Resumen: *Objetivo*: Para mejor caracterizar las relaciones entre acertar en las pruebas de visión remota precognitiva (VPR) y los parámetros de rasgo, estado, y objetivos potencialmente relevantes, recopilamos datos de VPR en dos experimentos en la red y examinamos la precisión en relación con: sexo al nacer, género, edad, ansiedad, amor incondicional, e interés en el objetivo. *Método*: En el Experimento 1 utilizamos una tarea de PRV de elección forzada, de tiempo no controlado y autoevaluada en la que 682 participantes

no remunerados contribuyeron 5,432 respuestas. En el Experimento 2 utilizamos una tarea PRV de respuesta libre, tiempo controlado, y juicio independiente en la que 307 participantes remunerados contribuyeron una sola respuesta cada uno. En ninguno de los dos casos preseleccionamos la capacidad de precognición de los participantes. *Resultados*: En el experimento 1 (tarea de PRV de elección forzada), no hubo precognición significativa de objetivos ni efecto de la edad en el rendimiento de PRV, pero encontramos un efecto complejo del sexo al nacer. También encontramos que los objetivos con mayor probabilidad de ser predichos correctamente eran más propensos a ser juzgados como interesantes en comparación con los objetivos con mayor probabilidad de ser predichos incorrectamente; un análisis pre-registrado confirmó este efecto. En el experimento 2 (tarea de PRV de respuesta libre) encontramos una precognición significativa del objetivo, ningún efecto de la edad sobre el rendimiento, y un efecto débil e indirecto del sexo. En el experimento 2, los sentimientos de amor incondicional y ansiedad correlacionaron con una mayor precisión. De nuevo, el interés por el objetivo se relacionó positivamente con la precisión. *Conclusiones*: Los resultados sugieren que acertar en las tareas de PRV se relaciona con el estado emocional de los participantes y el interés por el objetivo, y las características de la tarea afectan el rendimiento general. Basándonos en estas observaciones, ofrecemos recomendaciones para futuras investigaciones.

Translation into Spanish by Etzel Cardeña, Ph. D.

2024, Vol. 4, No. 1, pp. 122-141

Journal of Anomalous Experience and Cognition (JAEX)

A Comparison of Four New Automated Telephone Telepathy Tests[1]

Rupert Sheldrake
Schumacher College

Tom Stedall
The Schumacher Institute

Abstract: *Objective*. To develop user-friendly automated telephone telepathy tests. *Method*. In one kind of test, three participants who knew each other were linked together continuously in a conference call format. In each trial, the receiver was selected at random. The other two participants were muted and one was selected at random as the caller and asked to think about the receiver before being connected to that person. The receiver was asked to identify who was on the line, and then the caller and receiver were linked up and could talk. In the second type of test, trials were spaced out over longer time periods and callers and receivers went about their normal lives in between trials. *Results*. In none of the "conference call" tests was the hit rate significantly different from the chance level of 50%. In the second type of test, with a total of 266 trials, the hit rate was 57% (p = .01). *Conclusion*. The failure of our "conference call" tests to show any significant telepathic effects could have been because all three participants were continuously engaged with the test, which may have confounded any telepathic influences. Tests in which non-callers were not engaged with the experiment gave better results. We suggest developing an intuition training application that would work along with people's regular calls and messages. Such an app could be more user-friendly and enable participants to practice their intuitive skills, as well as enabling talented participants to be identified for more rigorous testing.

Keywords: telephone telepathy, automated tests, intuition training, psi

---

1 Address correspondence to Rupert Sheldrake, Ph. D., 20 Willow Road, London NW3 1TJ, UK, **rupert@rsheldrake.org**. We are grateful for financial support from the Planet Heritage Foundation, Naples, Florida through the Gaia Foundation, London, and from the Watson Family Foundation. We thank Pam Smart for her help in recruiting participants, Merlin Sheldrake for his comments on a draft of this paper, and Cosmo Sheldrake for the music used in Experiments 2 and 3.

https://doi.org/10.31156/jaex.25250

Highlights

- We explored a range of automated tests for telephone telepathy in an attempt to develop a user-friendly system.
- Tests involving a "conference call" format in which all three participants remained connected to the system showed no significant telepathic effects.
- Tests in which trials were conducted less frequently and in which non-callers were unaware that trials were taking place gave significant positive results, $p = .01$.
- We suggest developing an intuition training app that would work with calls or messages that are happening anyway.

Seemingly telepathic experiences in connection with telephone calls are very common. Surveys in Europe and the Americas have shown that most people say they have thought of someone for no apparent reason, then that person called, or that they know who is calling before looking at the caller ID or answering the phone (Sheldrake, 2003). Committed skeptics dismiss all these experiences on the grounds that they are a product of chance coincidence and/or selective memory, and some argue that in any case telepathic experiences are impossible in principle, and therefore any evidence for them can be dismissed in advance (Pinker, 2021; Reber & Alcock, 2020). Despite these negative opinions, based on prior assumptions rather than empirical data, there is empirical evidence that telephone telepathy does in fact occur. In experimental tests, participants had 2, 3 or 4 potential callers, and for each trial the callers were selected at random; when the callers called the participants, the participants indicated who they felt was calling before answering the phone. If they had simply been guessing then the expectation of being right by chance would have been 25% with 4 potential callers, 33% with 3 and 50% with 2. In studies so far, the overall hit rates have significantly exceeded chance levels (Lobach & Bierman, 2004; Sheldrake & Smart, 2003a, 2003b, Sheldrake et al., 2004; Sheldrake et al., 2015). Similar studies testing for telepathy in connection with online messages (Sheldrake & Beharee, 2009; Sheldrake & Lambert 2007), SMS messages (Sheldrake et al., 2009), and emails (Sheldrake & Avraamides, 2009; Sheldrake & Smart, 2005) have likewise given significant positive results.

In some of these experimental studies, participants were recruited on the basis of their claim that they often experienced telephone telepathy in real life situations, and some were selected for further testing on the basis of high hit rates in initial tests. In initial tests with 63 self-identified *sensitive* participants, the average hit rate in a to-

tal of 571 telephone telepathy trials was 40%, very significantly above the chance rate of 25% ($p < 1 \times 10^{-15}$; Sheldrake & Smart, 2003a). These initial tests were not filmed, which left open the possibility that some participants might have cheated.

In a second series of tests, carried out with four selected participants under filmed conditions that greatly reduced the possibility of cheating, the hit rate was 45% compared with the chance level of 25% ($p < 1 \times 10^{-12}$). The highest-scoring participant had a hit rate of 57%, with 16 trials correct out of 28, $p = .0003$ (Sheldrake & Smart, 2003b). These findings indicate that some people are much more sensitive than others, a conclusion independently supported by the findings of Schmidt et al. (2009). In their initial telephone telepathy tests, most of their participants did not score significantly better than chance, but in further videotaped tests one woman scored 50% compared with 25% expected by chance. In an additional 60 filmed trials, she was right 24 times (40%, $p = .007$). Thus, some participants were strikingly better than others in these tests and were able to score reliably and repeatedly above chance under rigorous test conditions.

In automated tests for telephone telepathy conducted with mobile phones with unselected participants, hit rates were less impressive than with selected participants, but still very significant. In tests with three potential callers, out of 2,080 trials there were 869 hits (42%) compared with the 33% chance level; and with two potential callers 411 hits out of 745 trials (55%) compared with 50% by chance (Sheldrake et al., 2015). In these automated tests, potential callers, selected at random, were sent a text message asking them to call their receiver through the automated system. When they made the call, they were not connected directly to the receivers; instead, the receivers were sent a text message asking them to indicate which of their potential callers was on the line, after which the line opened up and they could talk to the caller. Although in most cases this system worked well, a major problem was that, for a variety of reasons, some of the potential callers did not respond when asked to, and the tests therefore came to a premature end.

Here, we describe four new exploratory automated telephone telepathy experiments, the first three of which were designed to avoid the problem of missed calls by keeping all potential callers connected with the system throughout the test. Our aim was to make the tests work more reliably by preventing them from ending prematurely because of unreliable callers. Our fourth experiment was similar to the automated telephone telepathy tests in previous studies but used more sophisticated software than was available when the first automated tests were devised.

Unsupervised tests of these kinds are potentially open to cheating and cannot provide conclusive evidence for telepathy by themselves, but they enable people to try to improve their telepathic abilities by practice, with the ability to monitor their own progress. Such tests could also enable experimenters to identify unusually sensitive participants for further research under more rigorous conditions.

Methods

Test System

All these tests were conducted on mobile phones and were designed to be carried out by participants and callers who could be in any location within the same country – the UK or US – although, of course, they were asked not to be in the same room or otherwise nearby in such a way that participants could see or hear what callers were doing.

The test platform was the programmable communication system Twilio, which allows the automation of calls and texts, and the construction of call environments. The application was written in TwiML (Twilio Markup Language), PHP and MySQL. Random numbers were generated using the *mt_rand()* function in PHP, which utilizes the Mersenne Twister Random Number Generator. The generator was seeded from user time, location, and various system variables. This randomization determined the choice of caller.

Tests were designed to work in a given country, and the participants chose their country in which the test would be performed test during registration. Country dialing codes were automatically added to the numbers provided, and the test did not work in countries other than the one chosen. Where applicable, available test times were given in the local time of the chosen country. Each version had its own database.

Experiments 1, 2, and 3

In Experiments 1, 2, and 3, when the initiator started the test, all three participants were called by the automated system and asked to confirm their participation. If all responded and agreed to participate, they were connected together in a conference call. They were given instructions for the test by a voice message, and the first trial began. All three participants remained continuously connected together throughout the test, during which six trials took place. In each trial one of the participants was

randomly selected as a caller and one as a receiver; the remaining participant was a non-receiver. The receiver was then asked to indicate which of the two potential callers was thinking about them, choosing by key press from the two names, which were presented in alphabetical order. The receiver and caller were then connected so they could speak to each other for 15 seconds. After each trial ended, all three participants could speak to each other for 45 seconds before the next trial began. The tests ended when all six trials were completed, or if they could not run to completion for any reason (e.g., participants disconnecting). At the end of the test, summary scores for the group were given. Tests with guesses recorded in all six trials were marked as *complete* while those with guesses recorded in less than six trials were marked as *incomplete*. The whole test took approximately 10 minutes. The first three tests differed as follows:

Experiment 1: This test was designed to obtain as much information as possible, involving all three participants in every trial. The caller was asked to think about the receiver, while the receiver and non-receiver were told they were or were not being thought about respectively, and asked to indicate who they thought the caller was from the two other names presented in alphabetical order. Thus, the non-receiver was a kind of control. Once the receiver had responded, s/he was connected with the caller. Once the non-receiver had responded, s/he was played hold music. All three participants were then connected back together, so they could talk to each other until the next trial began. The test ended when all six trials were completed. In every trial the three roles were assigned afresh at random so that, on average, each person was a receiver in two trials, a caller in two trials and a non-receiver in two trials.

Out of a total of 1,047 trials, 1,031 were by participants in the UK and 16 in the US. Complete tests were those in which guesses were recorded in all 6 trials; incomplete tests were those with guesses recorded in less than 6 trials.

Experiment 2: This version was similar to Experiment 1, except that the non-receiver was informed that s/he was a non-receiver, and then was played music while the trial proceeded until being reconnected with the other two participants after the trial ended. All the results in Experiment 2 came from participants in the UK and Canada.

Experiment 3: Experiment 3 was the same as Experiment 2 except that the same person, chosen at random, served as the receiver in all six trials. In Experiment 3, 430 guesses were from participants in the UK, and 17 from the US and Canada.

Experiment 4: This test procedure differed from the previous three in that not all

three participants were on the phone at the same time. The person who initiated the test was the receiver in all trials. In each trial, at randomly chosen times during preselected time windows, one of the two callers, chosen at random, was telephoned and asked by a voice message asked to press 1 to take part. He or she was then asked to think about the receiver and played music. The receiver was also called, and told that one of the two callers was on the line, and asked to guess which by pressing 1 for the first caller and 2 for the second. The names of callers were alphabetically ordered, so nothing could be inferred from which was designated 1 or 2. Once the receiver had made her choice, he or she was connected to their caller, after which the two could talk for as long as they wanted. It was only at this point that receivers knew if they were correct or not. The test continued until all six trials had been attempted. As previously, tests were marked as *complete* if six trials were completed, otherwise they were treated as *incomplete*. Tests could be completed in as little as an hour or over several days. For most of this time the participants were not on the telephone. Receivers necessarily took part in all trials, but callers took part in an average of three trials each lasting only a minute or so. When the trials were not taking place, participants pursued other activities and went about their normal lives.

This general procedure was similar to that used by Sheldrake et al. (2015) with the difference that the previous procedure involved sending text messages to the randomly chosen callers asking them to call the receiver through a special telephone number, which was that of the test system. In this new test, the system built with Twilio enabled the message to the callers and their calls to the receiver to be made seamlessly.

Instructions

The instruction messages for the Experiment 1 were as follows; in this example the person registering the test was Rupert and the other two participants were Tom and Pam. Experiments 2 and 3 used similar messages, but with the non-receiver informed that they were not being thought about and would listen to music until they were reconnected with the other participants.

Message to all participants when test begun:

"Welcome to Rupert's telepathy test Rupert/Tom/Pam. To participate in the test, please press one… You will now be connected to the other participants.

Hello and thanks for taking part. When the first trial begins, one of you will be selected at random to be the caller, and will be asked to think about one of the other

participants. The other two will both be told whether or not they are being thought about, and will be asked who they feel the caller is. After they have registered their responses, the caller will be connected to the person he or she has been thinking about. The third person will then be reconnected, and all three of you can talk to each other for a while, until the next trial begins. There will be 6 trials altogether… The test will commence in five seconds.

Message to Rupert:

You will think about another participant. Please think about Tom. You will be connected to this person shortly.

Message to Tom:

You are being thought about by one of the other participants. Please take your time to try to guess who this is. You will have about a minute. Please make your choice when you are ready. [Choice is made here]. Your guess of Rupert has been registered. You will now be connected to the person thinking about you.

Messages to Pam:

In this trial you are not being thought about. But please take your time to try to guess who is thinking about someone. You will have about a minute. Please make your choice when you are ready. [Choice is made here] Your guess of Rupert has been registered. You will be reconnected to the other participants soon. You will now be connected back with all the participants, and the next round will commence.

Final message to all participants:

Thank you for taking part. In guessing who was calling, your group scored 3 hits out of 6 when they were being called by someone and 3 hits out of 6 when they were not being called by anyone. The test is now complete. Thanks again for taking part and goodbye.

The simpler instructions for Experiment 4 were as follows:

Message to Rupert

Welcome to Rupert Sheldrake's telepathy test, Rupert. Please press one to continue. Please think about Tom. When you are ready, press one and we will try to connect you to them."

Message to Tom

Welcome to Rupert Sheldrake's telepathy test, Tom. One of your friends is on the line. Please try to guess who it is, and you will be connected to them. Please enter your choice when you are ready. We will now connect you. You can both hang up when you are ready.

Registration

During test registration, the initiating participant provided their name and phone number in a web interface on www.sheldrake.org. Optionally, they also provided an email address, and their gender (male or female) and their approximate age (20-30, 30-40, etc.). The initiating participant also provided names and phone numbers for two friends, and optionally their genders and approximate ages. For the first three tests, the test began immediately after initiation and the roles were determined at random immediately before each trial. For the fourth test, the initiating participant selected time windows in which trials could be attempted. These were one-hour periods available from the next complete hour after registration; for example, if the volunteer registered at 9.30 am, the first slot was at 10.00 am. Slots were available during the next two days also, from 9.00 am until 9.00 pm. Slots on the day of registration were likewise available until 9.00 pm. Once the volunteer had provided all these details, they could initiate the test. Times were chosen for six trials within the chosen time slots, and the caller for each trial was also chosen (the initiator was always the receiver). This information was not revealed to the participants but was predetermined at this point.

Participants

Participants were recruited through Rupert Sheldrake's website www.sheldrake.org, newsletter, Facebook website, and also newsletter, sent to people who had subscribed to it, and at his talks. It was hard to persuade people to take part, largely because of the difficulty of their having to recruit two other people to be available at the same time (Experiments 1-3) or continuously available during the test period (Experiment 4). For this reason, we did not set in advance a target number of participants for each experiment, although we aimed for a number exceeding a minimum of 200 trials, and simply recruited as many as we could within a given time period. We did not ask participants about their prior beliefs in telephone telepathy, but most participants were probably aware of the phenomenon and took part because they believed it was possible. As experimenters, we thought that telephone telepathy was a real phenom-

enon, based on prior experimental results, but were neutral (3 on a 1 to 5 scale, where 5=strong belief and 1=strong non-belief) about the effectiveness of the procedures in Experiments 1-3, which had never been tried before. Experiment 4 was similar to previous experiments that had given positive results and so had a moderate belief (level 4) that this experiment would also give positive results. This research was conducted with ethical principles and informed consent as described in www.publichealthnotes.com/research-ethics-definition-principles-and-advantages/. This project was preregistered with the Open Science Foundation (OSF) on November 12, 2015 with registration https://osf.io/72b3g/

Storage of Data

All test and trial data were stored in the system database, either during test registration, or during the execution of a trial. MySQL views were used to make the data easily readable for individual trials and tests, as well as provide running totals of the guess success rate. These tables were available in a password protected web interface. We also archived data by emailing copies of it to a third party every week. The data are publicly available on the OSF website at https://osf.io/72b3g/

Analysis

Statistical analyses were carried out by binomial test with the expected probability of a correct response by chance of .5 or 50%, one tailed, and with chi squared analysis for proportions.

Results

In all tests, it was possible that a guess was not registered, either by the participant not pressing a key, pressing the wrong key, or mis-pressing the intended key. The number of recorded guesses is thus generally lower than the number of trials conducted.

Experiment 1

In this test there was a total of 1,089 trials, with 536/1047 (51%) correct responses by receivers, not significantly different from the chance level (p = .23). There were 145 completed tests (870 trials) with all receiver guesses recorded, with 456 hits (52%)

(Table 1). The non-receivers served as controls: they knew that they were non-receivers and were asked to guess who the caller was even though s/he was not thinking about them. Their responses could therefore be expected to be close to chance, which in fact they were, with 509 hits out of 1,035 responses (49%, p = .31).

Table 1.

Hit Rates with Receivers and Non-receivers in Experiment 1

	Receiver			**Non-receiver**		
Tests	Trials	Hits	Hits %	Trials	Hits	Hits %
Complete	870	456	52	822	410	50
Incomplete	177	80	45	213	99	47
Total	1047	536	51	1035	509	49

Experiment 2

After finding a hit rate close to chance level in Experiment 1, we explored the possibility that our design might have created a telepathic confusion for the receivers. The non-receivers were trying to detect which of the other two participants might be the caller and were therefore thinking about these participants, one of whom was the receiver. Consequently, the receiver may have detected influences both from the actual caller and also from the non-receiver, confounding any possible telepathic effect. To avoid this confusion, we ran a new version, Experiment 2, in which non-receivers were again informed that they were not the receivers and were played music during the trials in which they were not participating. In this test the hit rate was still not significantly above the chance level: out of 231 trials there were 118 hits (51%, p = .35). The results from complete and incomplete tests are shown in Table 2.

Table 2.

Hit Rates in Experiments 2 and 3

	Experiment 2			**Experiment 3**		
Test	Trials	Hits	Hits %	Trials	Hits	Hits %
Complete	180	93	52	414	216	52
Incomplete	51	25	49	33	15	46
Total	231	118	51	447	231	52

Experiment 3

We next asked whether the non-significant result in Experiment 2 might have occurred because the receiver was changed from trial to trial at random, possibly confusing the receivers and perhaps also reducing the possibility that they could improve through practice. Thus, in Experiment 3, the same participant was the receiver in all six trials within a test, which was otherwise conducted in the same way as Experiment 2. Once again the results were not significantly above chance, with 231 hits out of 447 trials (52%, p = .22). The results from complete and incomplete tests are also shown in Table 2.

Experiment 4

We next wondered why the results of Experiments 1-3, so close to chance, differed from previous telephone telepathy tests, which gave above-chance results. One possible reason was that our experimental design meant that all participants were continuously linked throughout the tests in a conference call format and were talking to each other between trials. This closeness may have meant that they were telepathically entangled even when we wanted the non-receivers to switch their attention to the music they were hearing and not to think about the other participants. We therefore ran another test that was similar to the experimental designs in previous studies, in which the callers and receivers were detached from each other between trials and could take part in other activities. In particular, each caller only occasionally took part in a trial, with gaps of up to several hours in between. Moreover, when one caller was

taking part, the other caller was unaware that a trial was taking place. Under these circumstances, there were 152 hits out of 266 trials (57%), a hit rate similar to that in previous automated telephone telepathy tests, and significant, p = .01. In the complete tests the hit rate was slightly below chance, but in the incomplete tests the average hit rate was 60% (Table 3). However, this difference between hit rates was not significant, χ^2 = 2.69, p = .10. by a 2x2 chi-squared test.

Table 3.

Hit Rates in Complete and Incomplete Tests in Experiment 4.

Test	Trials	Hits	Hits %
Complete	66	32	49
Incomplete	200	120	60
Total	266	152	57

For these tests to be completed, both callers and receivers had to be free to answer their phones six times, and therefore this low proportion of completed tests is not surprising and was indeed expected. The main reason we developed the procedures in Experiments 1-3 was to reduce the proportion of incomplete tests, and in that sense these procedures worked. In Experiment 1, the proportion of trials in complete as opposed to incomplete tests was 81%; in Experiment 2, 78% and in experiment 3, 93%. By contrast, in Experiment 4 this proportion was 25%.

One reason that we tried to increase the proportion of complete tests was to avoid the problem of "optional stopping," whereby participants who were not scoring above chance could simply have stopped doing the test, while those who were scoring above chance might have continued, even though their positive scores might have been a result of random guessing. Thus, optional stopping could, in principle, introduce a bias into the results, favoring artifactual above-chance scores. In Experiments 1-3, the hit rates in complete tests were at chance levels, and hence optional stopping, if it occurred, did not lead to spurious positive results. In Experiment 4, where the overall result was positive and significant, and where 75% of the tests were incomplete, could optional stopping have accounted for the positive results? No. Hit rates were in fact lower in complete tests than in incomplete tests, showing that optional stopping could not explain these positive results (Table 3).

We looked at the data from Experiment 4 in more detail to find out whether the overall positive effect could be explained by high scores from a few participants and scores close to chance levels by most others. Such a pattern might suggest that some people were cheating. The data in Table 4 show that this was not the case. If participants were simply guessing, by chance roughly equal numbers of tests should have had positive and negative scores. In tests with 5, 3 and 1 trials, all results were either positive or negative because scores of 2.5, 1.5 and 0.5 were impossible. "Positive" results in complete tests with 6 trials were those in which receivers gave 4, 5, or 6 correct responses, "at chance" those with 3, and "negative" those with 2, 1, or 0. Likewise, for tests with 5 trials, "positive" results were those with 3, 4, or 5 hits, and "negative" those with 2, 1, or 0 hits.

In fact, 17 tests had negative scores and 36 positive scores. Thus, out of a total of 53 tests with hit rates that were either positive or negative, 36 were above chance, a significant positive effect, p = .006. This analysis confirms that the overall positive effect in Experiment 4 could neither be explained in terms of optional stopping nor by most people scoring at near-chance levels with a few scoring strongly above chance. Significantly more people had positive than negative scores, showing that the positive scores were not confined to a small minority but were widely distributed.

Table 4.

The Number of Complete and Incomplete Tests with Positive, At Chance or Negative Results

Trials per test	Negative	At chance	Positive
6	4	5	2
5	6		15
4	1	4	3
3	4		6
2	0	8	4
1	2		6
Total	17	17	36

Discussion

We attempted to devise easy-to-use telephone telepathy tests in which all participants were connected together in a conference call format. All three versions this type of experiment gave results that were close to the chance level, showing no detectable signs of telepathy. However, when we reverted to an earlier type of experiment, in which there were longer intervals between trials and the receivers and callers went about their normal lives when not being tested, we obtained an above-chance result, comparable to the hit rate in a previous automated telepathy test (Sheldrake et al., 2015).

The failure of our first three tests may well have been because all three participants were connected together continuously. Although this experimental design reduced the proportion of incomplete tests, it may have led to telepathic confusion. The participants may have been unable to detach their minds from an awareness of each other, even when they were not callers. They also may have been disengaged by the automated call environment the tests employed. The systems we used, especially in Experiments 1-3, had a somewhat frustrating pace, typical of automated calls. Thus, while these programmable systems technically allow almost any test design to be constructed, they are unlikely to put participants at ease. In Experiment 4, the participants were free to go about their lives, with the system designed to be as unobtrusive as possible. Participants had no need to think about the procedure except during a trial. The small but significant positive effect, in line with previous results, suggests that this procedure was more effective, perhaps because there was less interference from potential callers who were not involved in the trial, and because the calls were relatively quick and therefore less likely to disengage participants. However, as in previous experiments of this type, many of the tests were not completed. This was often because the randomly selected callers did not answer their phones.

The fact that the first three experiments showed no significant effect is reassuring in one way: it suggests that the participants were not cheating. The data from Experiment 4 (Tables 3 and 4) lead to a similar conclusion. As in previous studies (Sheldrake et al., 2015), a small effect widely distributed among participants is not the pattern that would be expected if some people were cheating. If telephone telepathy did not exist and positive scores were the effects of cheating by a minority of participants, then a few people would have had high hit rates and others would be around the chance level, which is not what we found. Nevertheless, in unsupervised automated tests, it is impossible to completely eliminate the suspicion that some participants might cheat, and these tests cannot therefore provide conclusive evidence for telephone telepathy. Instead, they provide a system whereby participants may be able to practice and

increase their sensitivity, and also provide a way of identifying potentially talented participants.

In Experiment 4 and in other comparable tests of telephone, email, and SMS telepathy, the results were positive and significant, but were not far above chance, implying that most people have a limited ability to detect who is calling telepathically under these artificial test conditions. Nevertheless, some people do better than others, as discussed earlier. This is not surprising, given that human abilities and sensitivities are unevenly distributed: some people have a better sense of smell than others; some are unusually musical; some can hear high-pitched sounds that others can barely detect. Thus, the best uses for future automated tests would be to enable people to practice and, if possible, improve their telepathic abilities. Such tests could be incorporated into an intuition training app.

In this discussion, we have assumed that above chance results in telephone telepathy tests are indeed a result of telepathy, but there is another theoretical possibility, namely precognition. Could participants have picked up who was about to speak to them by anticipating this future experience, rather than telepathically detecting the callers' intention to call? In previous research using automated tests both with telephone calls and SMS messages, the effects seemed to be telepathic rather than precognitive. In telephone tests with three callers, with a mean chance expectation of a 33.3% hit rate, with more than 400 trials, the hit rate was 42%, ($p < .0001$). By contrast, under precognitive conditions, where the receiver was asked to guess who would be calling *before* the caller was selected at random, in more than 700 trials, the hit rate was 33%, almost exactly at chance level. In similar tests using SMS messages, under telepathic conditions in more than 800 trials the hit rate was 38% ($p < .01$), whereas under precognitive conditions, with 340 trials, the hit rate was 32%, again very close to chance level (Sheldrake, 2014). We therefore think it probable that the effect we observed in Experiment 4 was telepathic rather than precognitive.

In order to develop an automated training procedure, it may not be necessary to use the forced choice format employed in all the tests conducted so far. One general problem with forced-choice tests for psi abilities is that they may inhibit the very phenomena they are designed to investigate because they create artificial conditions that make participants self-conscious. In real-life conditions, most psi phenomena occur spontaneously and do not involve discursive thought or the conscious consideration of multiple options. In telephone telepathy, people "just know" or "feel" who is calling, rather than thinking about the possible alternatives. Moreover, the callers have a motive to call the receiver.

In our tests, by contrast, people had no motive except to follow automatically randomized instructions, and the receiver was confronted with a forced choice that raised the possibility of responding incorrectly, inducing doubt and uncertainty. It would be easier to encourage participation and also easier for the participants if there were a procedure that could be used as part of everyday life in connection with telephone calls or online messages they are receiving anyway. Here is one possible design using online messages:

Participants who are acting as receivers encourage some of the people who message them relatively frequently to do so through a training app. Instead of sending the message direct to the receiver, they send it to the training app which then sends a message to the receiver saying that "One of your friends is getting in touch with you right now. Please indicate who you think it is." They would then speak or type in, or simply select, the name of the person they thought was trying to contact them. The message would then be delivered, giving immediate feedback. An online database would keep track of their hit rates in these tests. Such a training app could be integrated with any major social media platform via their Application Programming Interfaces (APIs), effectively creating a substitute interface for widely used messaging systems. In some cases, participants may have been expecting a call from a particular person at a particular time, or know that person's habits in such a way that they can predict who is calling without the need for telepathy. Hence some of the positive responses may not depend on telepathy, but the receivers themselves would probably be aware of this fact. After the call has ended, they could be asked to rate how expected or unexpected the call was on a 1 – 5 scale.

A graphical display within the training app could show their weekly hit rates, and whether these were going up or down or remaining more or less the same. Such an approach could also overcome the most significant challenges of performing this research. The biggest barrier to participation was the requirement for three people to coordinate being online at the same or similar times. In practice, many people are online or available through their phones at the same time and responses are sometimes but not always almost instantaneous. The time between message and response would be recorded, and the effect of delayed interactions could be explored. This approach would remove the need for coordination by participants, utilizing instead the ubiquity and ease of online communications via platforms that participants already use.

If such an automated training system existed, people who claim to be able to teach others how to improve their intuitive skills could monitor their students' progress by finding out whether their hit rates were in fact increasing. This would enable intu-

ition training methods to be improved by providing an ongoing objective measurement of telepathic abilities. Then those who have been able to improve their telepathic skills, or who appear to be naturally talented, could be invited to take part in more rigorous tests, possibly with the incentive of payment. These tests would be done under supervised conditions with the participants being filmed in a way that could rule out cheating, similar to the original telephone telepathy tests that preceded the development of automated methods, in which the receivers responded by saying who they felt was calling, rather than responding to a forced-choice list (Sheldrake & Smart, 2003b; Schmidt et al., 2009).

There is much potential for automated tests for psi using phones, but the forced-choice methods used so far, including those described in this paper, do not work very well. Systems that go with the flow of people's everyday interactions seem more likely to be successful in both detecting and training telepathic abilities.

References

Lobach, E., & Bierman, D. (2004). Who's calling at this hour? Local sidereal time and telephone telepathy. In *Proceedings of the Parapsychological Association Annual Convention*, Vienna, 91-97.

Pinker, S. (2021). *Rationality: What it is, why it seems scarce, why it matters*. Viking.

Reber, A.S., & Alcock, J.E. (2020). Searching for the impossible: Parapsychology's elusive quest. *American Psychologist, 75*(3), 391-399. doi.org/10.1037/amp0000486

Schmidt, S., Erath D., Ivanova V., & Walach H. (2009) Do you know who is calling? Experiments on anomalous cognition in phone call receivers. *The Open Psychology Journal*, *2*, 12-18**.** doi: 10.2174/1874350100902010012

Sheldrake, R. (2003). *The sense of being stared at, and other aspects of the extended mind.* Coronet.

Sheldrake, R. (2014) Telepathy in connection with telephone calls, text messages and emails. *Journal of the Society of Life Information Sciences, 32*(1), 7-10. www.sheldrake.org/files/pdfs/papers/Telepathy-in-Connection-with-Telephone-Calls-Text-Messages-and-Emails.pdf

Sheldrake, R., & Avraamides, L. (2009) An automated test for telepathy in connection with emails. *Journal of Scientific Exploration*, *23*(1), 29-36. www.sheldrake.org/files/pdfs/papers/An-Automated-Test-for-Telepathy-in-Connection-with-Emails.pdf\

Sheldrake, R., Avraamides, L., & Novák, M. (2009). Sensing the sending of SMS messages: an automated test. *Explore: The Journal of Science and Healing*, *5*(5), 272-276. doi.org/10.1016/j.explore.2009.06.004

Sheldrake, R., & Beharee, A. (2009). A rapid online telepathy test. *Psychological Reports 104*(3), 957-990. doi.org/10.2466/PR0.104.3.957-970

Sheldrake, R., Godwin, H., & Rockell, S. (2004). A filmed experiment on telephone telepathy with the Nolan sisters. *Journal of the Society for Psychical Research*, *68*(3) 168-172. www.sheldrake.org/files/pdfs/papers/A-Filmed-Experiment-on-Telephone-Telepathy-with-the-Nolan-Sisters.pdf

Sheldrake, R., & Lambert M. (2007). An automated online telepathy test. *Journal of Scientific Exploration, 21*(3), 511-522. www.sheldrake.org/files/pdfs/papers/An-Automated-Online-Telepathy-Test.pdf

Sheldrake, R., & Smart, P. (2003a). Experimental tests for telephone telepathy. *Journal of the Society for Psychical Research, 67*(3), 184-199. www.sheldrake.org/files/pdfs/papers/Experimental-Tests-for-Telephone-Telepathy.pdf

Sheldrake, R., & Smart, P (2003b). Videotaped experiments on telephone telepathy. *Journal of Parapsychology, 67*(2), 147-166. www.sheldrake.org/files/pdfs/papers/Videotaped-Experiments-on-Telephone-Telepathy.pdf

Sheldrake, R. & Smart, P (2005). Testing for telepathy in connection with emails. *Perceptual and Motor Skills*, *101*(3), 771-786. doi.org/10.2466/pms.101.3.771-786

Sheldrake, R., Smart, P. & Avraamides, L. (2015). Automated tests for telephone telepathy using mobile phones. *Explore*, *11*(4), 310-319. doi: org/10.1016/j.explore.2015.04.001

Comparaison de Quatre Nouveaux Tests Automatisés de Télépathie par Téléphone

Rupert Sheldrake Tom Stedall

Résumé: *Objectif*: développer des tests automatisés de télépathie par téléphone, qui soient faciles à utiliser. *Méthode*: dans un type de test, trois participants qui se connaissaient étaient reliés entre eux de manière continue sous forme de conférence téléphonique. Dans chaque essai, le destinataire était choisi au hasard. Les deux autres participants ont été mis en sourdine et l'un d'entre eux a été choisi au hasard pour être l'appelant et il lui a été demandé de penser au destinataire avant d'être connecté à cette personne. On a demandé au destinataire d'identifier qui était en ligne, puis l'appelant et le destinataire ont été mis en relation et ont pu parler. Dans le second type de test, les essais étaient espacés sur des périodes plus longues et les appelants et les destinataires poursuivaient leur vie normale entre les essais. *Résultats*. Dans aucun des tests de "conférence téléphonique", le taux de réussite n'a été significativement différent du niveau de chance de 50 %. Dans le deuxième type de test, avec un total de 266 essais, le taux de réussite était de 57 % ($p = 0,01$). Conclusion. L'échec de nos tests de "conférence téléphonique" à montrer des effets télépathiques significatifs pourrait être dû au fait que les trois participants étaient continuellement engagés dans le test, ce qui a pu perturber toute influence télépathique. Les tests dans lesquels les personnes qui n'appelaient pas n'étaient pas impliquées dans l'expérience ont donné de meilleurs résultats. Nous suggérons de développer une application d'entraînement à l'intuition qui fonctionnerait en même temps que les appels et les messages habituels. Une telle application pourrait être plus conviviale et permettre aux participants d'ex-

ercer leurs compétences intuitives, tout en permettant d'identifier les participants particulièrement doués afin de mener avec eux des tests plus rigoureux.

Translation into French by Antoine Bioy, Ph. D.

Ein Vergleich von vier neuen automatisierten Tests zur Telefon-Telepathie

Rupert Sheldrake Tom Stedall

Zusammenfassung: Zielsetzung. Die Entwicklung benutzerfreundlicher automatischer Tests zur Telefon-Telepathie. Methode. Bei einer Art von Test wurden drei Teilnehmer, die sich kannten, kontinuierlich in Form einer Telefonkonferenz miteinander verbunden. Bei jedem Versuch wurde der Empfänger zufällig ausgewählt. Die beiden anderen Teilnehmer wurden stummgeschaltet, und einer wurde zufällig als Anrufer ausgewählt und gebeten, an den Empfänger zu denken, bevor er mit dieser Person verbunden wurde. Der Empfänger wurde gebeten, die Person in der Leitung zu identifizieren, und dann wurden Anrufer und Empfänger miteinander verbunden und konnten sprechen. Bei der zweiten Art von Test wurden die Versuche über längere Zeiträume verteilt, und Anrufer und Empfänger gingen zwischen den Versuchen ihrem normalen Leben nach. Ergebnisse. Bei keinem der "Konferenzgespräch"-Tests unterschied sich die Trefferquote signifikant von der Zufallsrate von 50 %. Bei der zweiten Testart mit insgesamt 266 Versuchen lag die Trefferquote bei 57 % ($p = .01$). Schlussfolgerung. Die Tatsache, dass unsere "Telefonkonferenz"-Tests keine signifikanten telepathischen Effekte zeigten, könnte daran liegen, dass alle drei Teilnehmer kontinuierlich mit dem Test beschäftigt waren, was eventuelle telepathische Einflüsse beeinträchtigen könnte. Tests, bei denen die Nicht-Anrufer nicht mit dem Experiment beschäftigt waren, ergaben bessere Ergebnisse. Wir schlagen vor, eine Anwendung für das Intuitionstraining zu entwickeln, die parallel zu den normalen Anrufen und Nachrichten der Teilnehmer funktionieren würde. Eine solche App könnte benutzerfreundlicher sein und es den Teilnehmern ermöglichen, ihre intuitiven Fähigkeiten zu trainieren und talentierte Teilnehmer für strengere Tests zu identifizieren.

Translation into German by Eberhard Bauer, Ph. D.

Uma Comparação de Quatro Novos Testes de Telepatia Automatizados por Telefone

Rupert Sheldrake Tom Stedall

Resumo: Objetivo. Desenvolver testes de telepatia automatizados por telefone que sejam user-friendly. Método. Em um tipo de teste, três participantes que se conheciam foram mantidos continuamente ligados em um formato de chamada em conferência. Em cada tentativa, o receptor era selecionado aleatoriamente. Os outros dois participantes foram silenciados e um foi selecionado aleatoriamente como o aquele que ligava e solicitado a pensar no receptor antes de ser conectado àquela pessoa. O receptor foi solicitado a identificar quem estava na linha e, em seguida, o chamador e o receptor foram conectados e pud-

eram conversar. No segundo tipo de teste, as tentativas foram espaçadas ao longo de períodos de tempo mais longos e os chamadores e receptores seguiram suas vidas normais entre as tentativas. Resultados. Em nenhum dos testes de "chamada em conferência" a taxa de acerto foi significativamente diferente do nível de chance de 50%. No segundo tipo de teste, com um total de 266 tentativas, a taxa de acerto foi de 57% (p = .01). Conclusão. A falta de efeitos telepáticos significativos em nossos testes de "chamada em conferência" pode ter se dado devido ao fato de os três participantes estarem continuamente envolvidos com o teste, o que pode ter confundido quaisquer influências telepáticas. Testes nos quais aqueles que não telefonaram não estavam envolvidos com o experimento apresentaram melhores resultados. Sugerimos o desenvolvimento de um aplicativo para treinamento de intuição que funcionaria junto às chamadas e mensagens regulares das pessoas. Tal aplicativo poderia ser mais fácil de ser utilizado e permitir que os participantes praticassem suas habilidades intuitivas, além de possibilitar a identificação de participantes talentosos para testes mais rigorosos.

Translation into Portuguese by Antônio Lima, Ph. D.

Comparación de Cuatro Nuevas Pruebas Automatizadas de Telepatía Telefónica

Rupert Sheldrake Tom Stedall

Resumen: *Objetivo*. Desarrollar pruebas automatizadas de telepatía telefónica fáciles de usar. *Método*. En un tipo de prueba, tres participantes que se conocían entre sí se conectaron continuamente en formato de conferencia telefónica. En cada prueba, el receptor se seleccionaba al azar. Se silenciaba a los otros dos participantes y se elegía a uno al azar como interlocutor y se le pedía que pensara en el receptor antes de conectarse con esa persona. Se pedía al receptor que identificara quién estaba en la línea y, a continuación, se conectaba a la persona que llamaba y al receptor, que podían hablar. En el segundo tipo de prueba, los ensayos se espaciaban durante periodos de tiempo más largos y los llamantes y los receptores hacían su vida normal entre las pruebas *Resultados*. En ninguna de las pruebas de "conferencia telefónica" el porcentaje de aciertos fue significativamente diferente del nivel de azar del 50%. En el segundo tipo de prueba, con un total de 266 pruebas, el porcentaje de aciertos fue del 57% (*p* = .01). *Conclusiones*. El hecho de que nuestras pruebas de "conferencia telefónica" no mostraran ningún efecto telepático significativo podría deberse a que los tres participantes estaban continuamente ocupados con la prueba, lo que podría haber confundido cualquier influencia telepática. Las pruebas en las que las personas que no participaban en la llamada no estaban involucradas en la prueba dieron mejores resultados. Sugerimos desarrollar una aplicación de entrenamiento de la intuición que funcione junto con las llamadas y mensajes habituales de la gente. Esta aplicación sería más fácil de usar y permitiría a los participantes practicar sus habilidades intuitivas, así como identificar a los participantes con talento para someterlos a pruebas más rigurosas.

Translation into Spanish by Etzel Cardeña, Ph. D.

Letter to the Editor

Moon Phases and Online Tests of Precognition

Julia Mossbridge, Ph.D.[1]

University of San Diego Dept. of Physics and Biophysics

In an article in this volume, "*State, trait, and target parameters associated with accuracy in two online tests of precognitive remote viewing*," my co-authors and I present several exploratory and some confirmatory analyses of data from two online remote viewing experiments. After the article was accepted for publication, my colleague Dean Radin reminded me that in 2020 I had written a note on a blog post about moon effects on remote viewing (https://adventuresofgreg.com/blog/2020/04/19/does-the-moon-effect-our-ability-to-make-decisions/). In It, I said I would like to analyze remote viewing data according to moon phase, but needed a reliable algorithm to do so. After Dean's reminder, I found a PHP script to turn UNIX timestamps into moon phases (https://minkukel.com/en/various/calculating-moon-phase/), translated that script into Matlab, and analyzed data from the first experiment in this study. I chose this experiment because it had enough participant-instigated trials spread over enough moon phases (two years) to make the analysis meaningful.

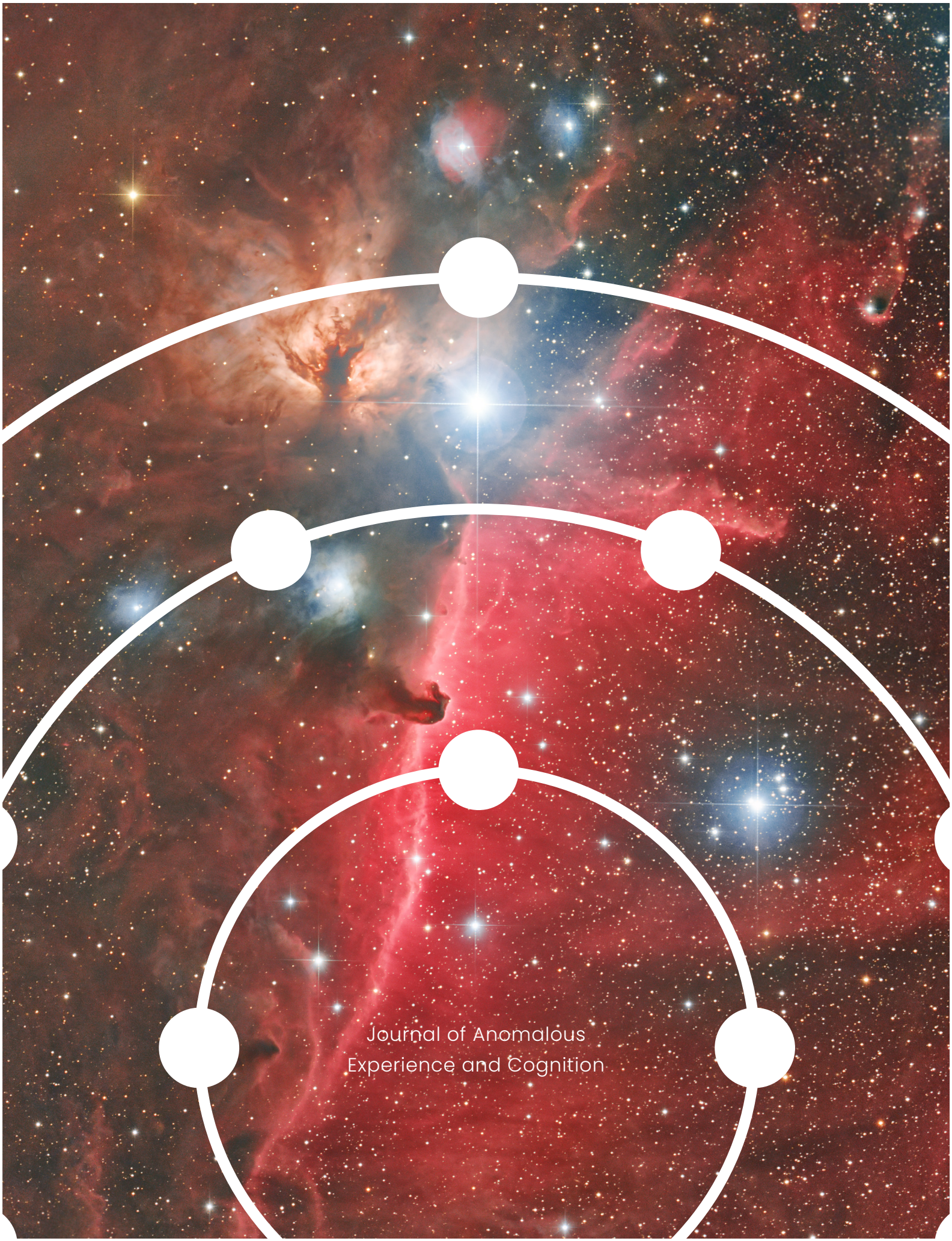
Using a Matlab script (available on request), I examined trials for targets presented in the second (non-biased) position in each of eight moon phases grouped into four composite phases with equal likelihood: new/waxing crescent, first quarter/waxing gibbous, full/waning gibbous, and last quarter/waning crescent. The analysis revealed that participants were more likely than expected by chance to instigate a trial during the last quarter/waning crescent phases, and less likely than chance to do so during the full/waning gibbous phases. This pattern held and was significant for both data batches (batch 1, n = 1,452: last qtr/waning crescent: proportion of trials .33, $p < 1x10^{-6}$ [binomial test, chance = .25]; full/waning gibbous: .21, $p < .0005$; batch 2, n = 2,659; last qtr/waning crescent: .35, $p < 1x10^{-6}$; full/waning gibbous: .19, $p < 1x10^{-6}$).

1 Address correspondence to **jmossbridge@sandiego.edu**

https://doi.org/10.31156/jaex.26006

A similar trend occurred for trials that matched the randomly selected targets, which also held for both data batches. During the moon phases when participants were more likely to choose to instigate a trial, they were also more likely to perform better, in contrast to the moon phases when participants were less likely to instigate trials (batch 1, last qtr/waning crescent: 220 hits, 266 misses, ratio = .83; full/waning gibbous: 112 hits, 194 misses, ratio = .58; $\chi^2_{(2, N=792)} = 5.79$, $p < .017$; batch 2, last qtr/waning crescent: 443 hits, 490 misses, ratio = .90; full/waning gibbous: 204 hits, 301 misses, ratio = .68; $\chi^2_{(2, N = 1438)} = 6.65$, $p < .01$).

Although it is not surprising that participants chose not to participate in the task during times when they did not perform well and chose instead to participate when they would perform better, it is surprising that these times were correlated so consistently with phases of the moon. The finding suggests a potential gravitational or geomagnetic influence on psi. I encourage other scientists studying precognition and psi phenomena in general to consider this type of simple but potentially revealing analysis.


Journal of Anomalous
Experience and Cognition