frontiers in
PSYCHOLOGY

MINI REVIEW ARTICLE
published: 10 June 2011
doi: 10.3389/fpsyg.2011.00117

Extraordinary claims require extraordinary evidence: the case of non-local perception, a classical and Bayesian review of evidences

Patrizio E. Tressoldi*

Dipartimento di Psicologia Generale, Università di Padova, Padova, Italy

Edited by:
Holmes Finch, Ball State University, USA

Reviewed by:
Jelte M. Wicherts, University of Amsterdam, Netherlands
Jeff Rouder, University of Missouri, USA

****Correspondence:***
Patrizio E. Tressoldi, Dipartimento di Psicologia Generale, Università di Padova, Via Venezia, 8, 35131 Padova, Italy.
e-mail: patrizio.tressoldi@unipd.it

Starting from the famous phrase "*extraordinary claims require extraordinary evidence*," we will present the evidence supporting the concept that human visual perception may have non-local properties, in other words, that it may operate beyond the space and time constraints of sensory organs, in order to discuss which criteria can be used to define evidence as extraordinary. This evidence has been obtained from seven databases which are related to six different protocols used to test the reality and the functioning of non-local perception, analyzed using both a frequentist and a new Bayesian meta-analysis statistical procedure. According to a frequentist meta-analysis, the null hypothesis can be rejected for all six protocols even if the effect sizes range from 0.007 to 0.28. According to Bayesian meta-analysis, the Bayes factors provides strong evidence to support the alternative hypothesis (H1) over the null hypothesis (H0), but only for three out of the six protocols. We will discuss whether quantitative psychology can contribute to defining the criteria for the acceptance of new scientific ideas in order to avoid the inconclusive controversies between supporters and opponents.

Keywords: meta-analysis, frequentist, Bayes, non-local perception

INTRODUCTION

"*Extraordinary claims require extraordinary evidence*" was a phrase made popular by Carl Sagan who reworded Laplace's principle, which says that "the weight of evidence for an extraordinary claim must be proportioned to its strangeness" (Gillispie et al., 1999). This statement is at the heart of the scientific method, and a model for critical thinking, rational thought and skepticism everywhere. However, no quantitative standards have been agreed upon in order to define whether or not extraordinary evidence has been obtained. Consequently, the measures of "extraordinary evidence" are completely reliant on subjective evaluation and the acceptance of "extraordinary claims." In science, the definition of extraordinary evidence is more a social agreement than an objective evaluation, even if most scientists would state the contrary (see, for example, the recent debate about climate change: Anderegg et al., 2010; Bodenstein, 2010).

However, a relevant example of an agreement about the strength of evidence has been defined in the field of clinical medicine and psychology in order to grade evidence to recommend the application of treatments for physical and mental clinical conditions. Recommendations that are based on evidence can be of different levels of quality. The sources of evidence, range from small laboratory studies or case reports to large, well-designed clinical studies that have minimized bias to a large extent. As poor-quality evidence can lead to recommendations that are not in the patient's best interests, it is essential to know whether a recommendation is strong (i.e., we can be confident about the recommendation) or weak (we cannot be confident). The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group (Guyat et al., 2008), for example, states that strong recommendations, meaning that most patients who are provided with the information would choose the recommended management and that clinicians can structure their interactions with patients accordingly, must derive from consistent evidence from a comprehensive meta-analysis of all of the evidence available or from at least two well-performed, randomized and controlled trials. If an agreement can be obtained in such an important field as that of human physical and mental health, we think that it should be possible to reach an agreement in the field of "human knowledge," where there are fewer risks of harming people.

AIM OF THE STUDY

In this paper, we will present a quantitative review of the evidence which is available today, supporting the hypothesis that the human mind may have non-local properties, that is, that some of its functions, i.e., perceptual abilities may extend beyond its local functions, that is beyond the space and time constraints of sensory organs. This quantitative review will be presented using both a classical frequentist and a new Bayesian meta-analytic approach. Before we justify our choice to use these two statistical approaches, a brief explanation of what we mean by non-local perception (NLP) is necessary.

NON-LOCAL PERCEPTION

We prefer the term NLP to the old-fashioned term extrasensory perception (ESP), because NLP allows us to use the non-local properties of physical "objects" such as photons, atoms, etc. and the laws of quantum mechanics as analogies. The main non-local properties which are studied within the realm of quantum physics and which are supported by "extraordinary evidence" (see Genovese, 2005,2010; Zeilinger, 2010), are "entanglement" and "measurement

interference." The first property, entanglement, allows two or more physical objects to behave as one even if they are separated in space and time. This "strange" property allows a form of immediate communication of the objects' characteristics over distances between or among the entangled objects, as has been observed in teleportation experiments (i.e., Bouwmeester et al., 1997).

The possibility that quantum-like properties may be observed not only in physics but even in biology and psychology has not only been studied theoretically (Khrennikov, 2010; Walach and von Stillfried, 2011) but even experimentally (see Gutierrez et al., 2010 for biology and Aerts, 2009 for psychology).

With regard to the methodology for studying NLP, the basic methods are the free-response and forced-response protocols. In the free-response protocol, participants are invited to perceive information, usually images or short video clips, using only their minds. This is because this information is only available at a distance or is chosen after their description and, consequently, no conventional (local) ways to perceive it are possible. During the task, the participants' normal state of consciousness may be altered with some techniques, e.g., they may be immersed in a ganzfeld environment, or put under hypnosis, meditation, etc. In contrast to the forced-choice protocol, the participants are allowed to describe verbally or through drawing what they perceive, allowing them all the time which is necessary to complete the task. With the forced-choice protocol, participants are simply required to quickly choose the distant (in terms of space or time) information from among a set, usually ranging from two to five. Obviously, in order to prevent the participants from guessing the target information using conventional means or using explicit or implicit strategies, all of the necessary safeguards for the experimental settings should be adopted, i.e., sensory shielding from the target information, proper randomization of the stimuli, etc. The results are the ratio of hits with respect the mean chance expectations.

A variant of the free-response protocol with participants in a normal state of consciousness is a procedure commonly referred to as the remote vision (RV) method. In a typical protocol, the participant is asked to describe the physical surroundings of a distant experimenter or to describe a target that they will see in a short while; this is also called precognitive targeting. A trial requires a participant, a monitor who will remain with the participant throughout the trial, a second experimenter and an analyst. Once the monitor and the participant have been sequestered in a laboratory, the second experimenter (E2) chooses one physical location at random from a predefined set called a target pool. At this moment, the monitor and the participant are blind to the choice of target. E2 then travels to that location and remains there for about 15 min during which time he/she attempts to experience the site as much as possible. Meanwhile, back in the laboratory, the monitor is free to ask the participant non-leading questions in order to illicit as much information as possible about the site where E2 is currently located. The participant is encouraged to write down and to draw his/her impressions. When the session is over, the data are copied and secured. Then the monitor and the participant travel to the selected site as a form of feedback. Naturally, this does not imply or constitute an analytical procedure.

There are many ways in which to analyze the output of such trials. The most common technique in use in remote viewing studies is the rank-order method. In general, an analyst who is blind to the target choice is presented with the original response and a set of targets which include the intended target for the trial. The analyst's task is to pick which of the targets best matches the response, and then the second best, the third best, and so on. After a number of such trials, the null hypothesis of no NLP can be tested using simple statistical methods.

In addition to the free- and forced-choice protocols, there is a new method which is used to study NLP using psycho-physiological responses (i.e., skin conductance, heart rate, EEG, fMRI). The basic procedure consists of the random presentation of two categories of information (i.e., emotional vs. neutral pictures) and the recording of the psycho-physiological responses *prior to* the presentation. If the statistical comparison between the anticipatory responses before the first and the second categories of information is significant, this is deemed to provide support for implicit NLP.

STATISTICAL ANALYSIS

The aim of the present paper is not to demonstrate that NLP is a quantum-like property of the human mind, but only to offer an update to the experimental evidence supporting NLP, letting the readers decide for themselves whether or not this evidence can be considered to be extraordinary.

Why use a classical frequentist and a Bayesian meta-analytical approach? The classical approach is that which was introduced by Glass and colleagues in the early 1980s (Glass et al., 1981; Hedges and Olkin, 1985). In extreme synthesis, it consists of a weighted inverse variance average of standardized measures (effect sizes) observed in all of the available studies relating to a specific topic, i.e., medical or educational interventions, psychological effects, etc. The strength of the evidence is demonstrated by the amount of studies retrieved and the measure of the average effect sizes with their confidence intervals and the associated probability of the null hypothesis being rejected. According to the fixed-effect model, we can assume that there is one true effect size (hence the term fixed effect) which underlies all of the studies in the analysis, and that any differences between this value and the observed effects are due to sampling errors. In contrast, under the random-effects model, we accept that the true effect could vary from study to study as a consequence of the influence of so-called moderator variables (i.e., participants or stimuli characteristics). The effect sizes in the studies that were actually performed are assumed to represent a random sample of these effect sizes, leading to the term "random effects."

In contrast to the classical approach, Bayesian meta-analysis (Rouder and Morey, 2011) provides a probability ratio as a summary of the results, called the Bayes factor (BF), a well-calibrated measure of the evidence of the ratio of probabilities of the data given two contrasting hypotheses, i.e., the reality of a phenomenon (H1) and its non-existence (H0). These latter quantities are called the posterior odds, and are the product of the BF and the prior odds. For example, a BF of three indicates that the observed level of evidence favors the alternative over the null hypothesis by a ratio of 3:1. Further details are given in the Methods section.

It is left to the reader to evaluate whether or not classical and Bayesian statistics obtained from studies testing the existence of NLP may be considered to be "exceptional." This is the aim of this study.

MATERIALS AND METHODS

THE DATABASE

The database comprises five meta-analyses which have already been published in different papers, all related to different aspects of NLP and from which it was possible to obtain the raw data from each study. Two more meta-analyses were used from which it was only possible to analyze the summary data.

The five meta-analyses with raw data include one which is related to NLP when participants are in the special altered state of consciousness (ASC) defined as the ganzfeld effect (Storm et al., 2010) which covers all of the available studies up to 2009. A second one is related to all of the studies available up to 2010 which are related to "anticipatory psycho-physiological responses" (Mossbridge et al., submitted). The third one is a meta-analysis related to NLP in participants with non-ASC using a forced-choice protocol (Storm et al., submitted), covering all of the relevant studies from 1987 to 2010. Of the two remaining meta-analyses, one is related to NLP in participants who are not in a ganzfeld state but who are in other ASC and studies relating to NLP in participants in a normal state of consciousness but using a free-choice protocol, covering all of the available studies from 1992 to 2009 (Storm et al., 2010).

The two meta-analyses which provide only summary data are related to the special protocol called RV with participants in non-ASC using free-response procedures. The first one was published by Milton (1997) and covers all of the studies related to this line of investigation from 1964 up to 1992. The second one is a summary of all of the studies conducted by Brenda Dunn and Robert Jahn within the Princeton Engineering Anomalies Research (PEAR) program from 1976 to 1999.

RESULTS

FREQUENTIST META-ANALYSIS

The raw data, effect sizes and standard errors were obtained from the databases of each of the five meta-analyses[1] and were analyzed by testing the fixed and random effect models with Comprehensive Meta-analysis Software® (Borenstein et al., 2005). This analysis provided the average weighted effect sizes with 0.95 confidence intervals and the corresponding *Z* values in order to test the null hypothesis.

BAYESIAN META-ANALYSIS

As discussed by Rouder et al. (2009), BFs respect the resolution of data: when the sample size is small, small effects may be considered as evidence for the null hypothesis as the null hypothesis is the more parsimonious description given the resolution provided by the data. As the sample size increases, however, the resolution provided for the data is finer, and small effects are more concordant with the alternative hypothesis.

Rouder and Morey's (2011) approach is to consider two hypotheses for a sequence of experiments. The first one, the null hypothesis, is that the true effect size is zero for all experiments. The second is that there is a single true effect size greater than zero which underlies all of the experiments.

Rouder and Morey approach considers a sequence of *t*-values, *t*1, *t*2…tM, from *M* comparisons.

$$\mathrm{BF} = \frac{\int \prod_{m=1}^{M} Pr(t_m \mid \delta) f(\delta) d\delta}{\prod_{i=1}^{M} Pr(t_m \mid \delta = 0)}$$

where δ is the effects size and *f* is the prior weights on parameter δ, the true effect size. A more conservative two-tailed prior on δ, was placed as default, following a *t*-distribution with a single degree of freedom (Jeffreys,1961; Rouder et al., 2009). The key property of this meta-analytic approach is that the true effect size is assumed to be constant across each experiment. In this sense it has the same assumption of the frequentist fixed model.

Following this approach, we calculated the *t*-value corresponding to the effect size and the number of participants in each study in the five meta-analyses which provided all of the raw data, and from the summary effect sizes of the two meta-analyses, and obtained a BF of H1/H0 that is NLP yes/NLP no.

The descriptive statistics and the statistical results plus the estimate of the file drawer effect of the five meta-analyses which include all of the raw data are presented in **Table 1**. The statistical results obtained using the summary data of the two meta-analyses relating to the RV procedure are presented in **Table 2**.

More detailed information about, e.g., effect size relationship with papers' methodological quality, number of participants, participants' characteristics, etc., is available in the original papers.

DISCUSSION

What evidence is there to support the existence of NLP? If we use the results obtained with the frequentist statistical approach, i.e., P(Data/H0), apart from the results obtained using participants in normal states of consciousness and the free-response protocol, all of the statistics in the remaining meta-analyses lead to the rejection of the null hypothesis, even if the measures of effect size are clearly greater using the free-response protocol.

In contrast, if we refer to the results obtained with the Bayesian statistical approach, i.e., P(H0/Data), only for the three meta-analyses which relate to the ganzfeld condition, the RV procedure and anticipatory responses, there is an high probability that H1, the hypothesis supporting the existence of NLP, may be true.

However, meta-analysis is sometimes criticized for mixing together good and bad studies from a methodological point of view. This criticism, is known as the "*garbage in and garbage out*" issue (Hunt, 1997, p. 42). One may hence wonder if beyond the quantitative evidences there are also qualitative ones. If the observed effects are due to "*garbage*" we should expect a negative correlation between effect size and study quality. In all three meta-analysis with the highest BF, the correlation between quality of study (obtained by at least two independent coders using predefined criteria) and effect size, ranged from $r = 0.05$ in Milton (1997) to $r = 0.36$ in Storm et al. (2010), suggesting a modest positive relationship. Another more specific criticism is that large-scale studies, that is those with more statistical power, fail to replicate the findings of many small-scale experiments, a clear paradox given than the opposite should be expected when the estimated effect size is low (see **Table 1**). If this is true, we should obtain a negative correlation between effect size and the number of participants. It was possible to calculate this measure for two out three meta-analysis with the highest BF, Ganzfeld:

[1]Available upon request from the author.

Table 1 | Descriptive and statistical results of the five meta-analysis with all raw data.

Meta-analysis	N. studies	N. participants	Fixed ES (0.95 CI)	*Z*	Random ES (0.95 CI)	*Z*	Bayes factor (H1/H0, 2-tailed)	File drawer effect
Ganzfeld[1]	108	3650	0.12 (0.11–0.14)	19.36	0.13 (0.09–0.17)	6.39	18861051*	357§
ASC[1]	16	427	0.12 (0.09–0.15)	8.63	0.11 (0.03–0.19)	2.86	0.04764247	13§
Anticipatory responses[2]	37	1064	0.26 (0.19–0.37)	8.7	0.28 (0.20–0.32)	6.07	2.891308e + 13	954#
Normal SC[1] (free response)	14	1026	–0.015 (–0.03–0.005)	–1.48	–0.03 (–0.06–0.002)	–1.84	0.02924606	–
Normal SC[3] (forced choice)	72	69726	0.007 (0.006–0.007)	16.2	0.011 (0.006–0.015)	4.88	0.003162905*	187§

*[1]Storm et al. (2010); [2]Mossbridge et al. (submitted); [3]Storm et al. (submitted); *one study excluded because N participants = 1; § = Darlington and Hayes's (2000) formula; # = Orwin's (1983) fail-safe N.*

Table 2 | Results of the two meta-analysis related to remote vision.

Meta-analysis	N. studies	N. participants	Fixed ES (0.95 CI)	*Z**	Bayes factor (H1/H0, 2-tailed)	File drawer effect
Dunne and Jahn (2003)	not defined	366	0.34 (0.19–0.49)	6.3	25424503838	849**
Milton (1997)	78	1158	0.16 (0.10–0.22)	5.7		866**

**Stouffer $Z = \sum z / \sqrt{\text{Number of studies}}$; **Rosenthal's fail-safe N.*

$r = -0.097$; Anticipatory Responses: $r = -0.054$. Even if we cannot forecast the correlation for the other meta-analysis, if present, this correlation is not generalized.

Are these converging results with these three protocols "extraordinary" evidence? Perhaps. Surely these results are well beyond the standards for a "strong recommendation" suggested by the GRADE system. However, the results presented in this study concern the "recommendation" to accept the existence of NLP and not to apply medical or psychological interventions to ameliorate human health. Do we need more stringent standards to enable us to accept phenomena that apparently seem to violate our common beliefs regarding physical laws? However, if results analyzed with both frequentist and Bayesian statistical approaches from more than 200 studies conducted by different researchers with more than 6000 participants in total and three different experimental protocols are not considered "extraordinary," or at least "sufficient" to suggest that the human mind may have quantum-like properties, what standards can possibly apply? Or we should accept that, in order to accept new hypotheses about the functioning of the human mind, it is necessary for us to abandon quantitative standards and in this case quantitative methods are useless?

As extensively discussed by Toomela (2010, p. 9) *how to behave if the theory about underlying processes has not been created yet? Here quantitative methods become valuable: it is possible to create useful generalizations without knowing the processes that underlie the events. This was exactly what Thurstone, for instance, aimed at: It is the faith of all science that an unlimited number of phenomena can be comprehended in terms of a limited number of concepts or ideal constructs. Without this faith no science could ever have any motivation. To deny this faith to affirm the primary chaos of nature and the consequent futility of scientific effort* (Thurstone, 1935, p. 44).

I would conclude citing some excerpts from Osborne's (2010) editorial about quantitative psychology:…*Through quantitative study of the human condition, we hope to gain insight into basic, fascinating questions that humans have pondered for millennia*…the promise of quantitative study of psychology is also one of its greatest challenges: demonstrating in a convincing way that quantification of behavioral, cognitive, biological, and psychological processes is valid, and that the analyses we subject the numbers to are honest efforts at elucidation rather than obfuscation."

Is it, therefore, hopeless to attempt to arrive at a consensus regarding what may be considered as "extraordinary evidence" or at least "sufficient" evidence to support new scientific claims as in the realm of human health, without resorting to inconclusive rebuttals between the supporters and opponents of new ideas?

ACKNOWLEDGMENTS

We thanks the Proof Reading Service for revising English, Jeff Rouder for providing the code to implement the Bayesian meta-analysis and the reviewers and the associate editor for their helpful comments.

REFERENCES

Aerts, D. (2009). Quantum structure in cognition. *J. Math. Psychol.* 53, 314–348.

Anderegg, W. R. L., Prall, J. W., Harold, J., and Schneider, S. H. (2010). Expert credibility in climate change. *Proc. Natl. Acad. Sci.* 107, 12107–12109.

Bodenstein, L. (2010). Regarding Anderegg et al. and climate change credibility. *Proc. Natl. Acad. Sci.* 107, E188.

Borenstein, M., Hedges, L., Higgins, J., and Rothstein, H. (2005). Englewood, NJ: Biostat.

Bouwmeester, D., Pan, J. W., Mattle, K., Eibl, M., Weinfurter, H., and Zeilinger, A. (1997). Experimental quantum teleportation. *Nature* 390, 575–579.

Darlington, R. B., and Hayes, A. F. (2000). Combining independent p values:

extensions of the Stouffer and binomial methods. *Psychol. Methods* 5, 496–515.

Dunne, B. J., and Jahn, R. G. (2003). Information and uncertainty in remote perception research. *J. Sci. Explor.* 17, 2, 207–241.

Genovese, M. (2005). Research on hidden variable theories, a review of recent progresses. *Phys. Rep.* 413, 319–396.

Genovese, M. (2010). Interpretations of quantum mechanics and measurement problem. *Adv. Sci. Lett.* 3, 249–258.

Gillispie, C. C., Gratton-Guinness, I., and Fox, R. (1999). *Pierre Simon Laplace, A Life in Exact Science*. Princetion, NJ: Princeton University Press.

Glass, G. V, McGaw, B., and Smith, M. L. (1981). *Meta-Analysis in Social Research*. Beverly Hills, CA: SAGE Publications.

Gutiérrez, R., Caetano, R., Woiczikowski, P. B., Kubar, T., Elstner, M., and Cuniberti, G. (2010). Structural fluctuations and quantum transport through DNA molecular wires, a combined molecular dynamics and model Hamiltonian approach. *New J. Phys.* 12, 023022.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., and Schünemann, H. J. (2008). GRADE, an emerging consensus on rating quality of evidence and strength of recommendations. *Br. Med. J.* 336, 924. doi: 10.1136/bmj.39489

Hedges, L. V., and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press.

Hunt, M. (1997). *How Science Takes Stock*. New York: Russell Sage Found.

Jeffreys, H. (1961). *Theory of Probability*, 3rd Edn. New York: Oxford University Press.

Khrennikov, A. Y. (2010). *Ubiquitous Quantum Structure from Psychology to Finance*. Berlin, Heidelberg: Springer-Verlag.

Milton, J. (1997). Meta-analysis of free-response ESP studies without altered states of consciousness. *J. Parapsychol.* 61, 279–319.

Mossbridge, J., Tressoldi, P. E., and Utts, J. (submitted). Physiological anticipation of unpredictable stimuli, a meta-analysis.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *J. Educ. Stat.* 8, 157–159.

Osborne, J. W. (2010). Challenges for quantitative psychology and measurement in the 21st century. *Front. Psychol.* 1:1. doi: 10.3389/fpsyg.2010.00001

Rouder, J. N., and Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychon Bull Rev*. doi: 10.3758/s13423-011-0088-7

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* 16, 225–237.

Storm, L., Tressoldi, P. E., and Di Risio, L. (2010). Meta-analyses of free-response studies 1992-2008, assessing the noise reduction model in parapsychology. *Psychol. Bull.* 136, 491–494. doi: 10.1037/a0019840

Storm, L., Tressoldi, P. E., and Di Risio, L. (submitted). Meta-analysis of ESP studies, 1987-2010, assessing the success of the forced-choice design in parapsychology.

Thurstone, L. L. (1935). *The Vectors of Mind: Multiple-Factor Analysis for the Isolation of Primary Traits*. Chicago: University of Chicago Press.

Toomela, A. (2010). Quantitative methods in psychology: inevitable and useless. *Front. Psychol.* 1:29. doi: 10.3389/fpsyg.2010.00029

Walach, H., and von Stillfried, N. (2011). Generalised quantum theory. Basic idea and general intuition: a background story and overview. *Axiomathes*. doi: 10.1007/s10516-010-9145-5

Zeilinger, A. (2010). *The Dance of the Photons. From Einstein to Quantum Teleportation*. NY: Farrar, Straus and Giroux.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 February 2011; accepted: 19 May 2011; published online: 10 June 2011.

Citation: Tressoldi PE (2011) Extraordinary claims require extraordinary evidence: the case of non-local perception, a classical and Bayesian review of evidences. Front. Psychology ***2****:117. doi: 10.3389/fpsyg.2011.00117*

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.