## Verisimilitude, Belief, and Progress in Psychological Science

Does science offer a way to learn what is true about our world? According to the perspective in philosophy of science known as *scientific realism*, the answer is ‘yes’. Scientific realism is the idea that successful scientific theories that have made novel predictions give us a good reason to believe these theories make statements about the world that are at least partially true. Known as the *no miracle argument*, only realism can explain the success of science, which consists of repeatedly making successful predictions (Duhem, 1906), without requiring us to believe in miracles.

Not everyone thinks that it matters whether scientific theories make true statements about the world, as scientific realists do. Laudan (1981) argues against scientific realism based on a pessimistic meta-induction: If theories that were deemed successful in the past turn out to be false, then we can reasonably expect all our current successful theories to be false as well. Van Fraassen (1980) believes it is sufficient for a theory to be ‘empirically adequate’, and make true predictions about things we can observe, irrespective of whether these predictions are derived from a theory that describes how the unobservable world is in reality. This viewpoint is known as *constructive empiricism*. As Van Fraassen summarizes the constructive empiricist perspective (1980, p.12): “Science aims to give us theories which are empirically adequate; and acceptance of a theory involves as belief only that it is empirically adequate”.

The idea that we should ‘believe’ scientific hypotheses is not something scientific realists can get behind. Either they think theories make true statements about things in the world, but we will have to remain completely agnostic about when they do (Feyerabend, 1993), or they think that corroborating novel and risky predictions makes it reasonable to believe that a theory has some ‘truth-likeness’, or *verisimilitude*. The concept of verisimilitude is based on the intuition that a theory is closer to a true statement when the theory allows us to make more true predictions, and less false predictions. When data is in line with predictions, a theory gains verisimilitude, when data are not in line with predictions, a theory loses verisimilitude (Meehl, 1978). Popper clearly intended verisimilitude to be different from belief (Niiniluoto, 1998). Importantly, verisimilitude refers to how close a theory is to the truth, which makes it an ontological, not epistemological question. That is, verisimilitude is a function of the degree to which a theory is similar to the truth, but it is not a function of the degree of belief in, or the evidence for, a theory (Meehl, 1978, 1990). It is also not necessary for a scientific realist that we ever know what is true – we just need to be of the opinion that we can move closer to the truth (known as comparative scientific realism, Kuipers, 2016).

Attempts to formalize verisimilitude have been a challenge, and from the perspective of an empirical scientist, the abstract nature of this ongoing discussion does not really make me optimistic it will be extremely useful in everyday practice. On a more intuitive level, verisimilitude can be regarded as the extent to which a theory makes the most correct (and least incorrect) statements about specific features in the world. One way to think about this is using the ‘possible worlds’ approach (Niiniluoto, 1999), where for each basic state of the world one can predict, there is a possible world that contains each unique combination of states.

For example, consider the experiments by Stroop (1935), where color related words (e.g., RED, BLUE) are printed either in congruent colors (i.e., the word RED in red ink) or incongruent colors (i.e., the word RED in blue ink). We might have a very simple theory predicting that people automatically process irrelevant information in a task. When we do two versions of a Stroop experiment, one where people are asked to read the words, and one where people are asked to name the colors, this simple theory would predict slower responses on incongruent trials, compared to congruent trials. A slightly more advanced theory predicts that congruency effects are dependent upon the salience of the word dimension and color dimension (Melara & Algom, 2003). Because in the standard Stroop experiment the *word* dimension is much more salient in both tasks than the *color* dimension, this theory predicts slower responses on incongruent trials, but only in the color naming condition. We have four possible worlds, two of which represent predictions from either of the two theories, and two that are not in line with either theory.

|  |  |  |
| --- | --- | --- |
|  | Responses Color Naming | Responses Word Naming |
| World 1 | Slower | Slower |
| World 2 | Slower | Not Slower |
| World 3 | Not Slower | Slower |
| World 4 | Not Slower | Not Slower |

In an unpublished working paper, Meehl (1990b) discusses a ‘box score’ of the number of successfully predicted features, which he acknowledges is too simplistic. No widely accepted formalized measure of verisimilitude is available to express the similarity between the successfully predicted features by a theory, although several proposals have been put forward (Niiniluoto, 1998; Oddie, 2013, for an example based on Tversky's (1977) contrast model, see Cevolani, Crupi, & Festa, 2011). However, even if formal measures of verisimilitude are not available, it remains a useful concept to describe theories that are assumed to be closer to the truth because they make novel predictions (Psillos, 1999).

As empirical scientists, our main job is to decide which features are present in our world. Therefore, we need to know if predictions made by theories are corroborated or falsified in experiments. To be able to falsify a theory, it needs to forbid certain states of the world (Lakatos, 1978). This is not easy, especially for probabilistic statements, which is the bread and butter of psychological science. Where a single black swan is clearly observable, probabilistic statements only reach their true predicted value in infinity, and every finite sample will have some variation around the predicted value. However, according to Popper, probabilistic statements can be *made* falsifiable by interpreting probability as the relative frequency of a result in a specified hypothetical series of observations, and decide that reproducible regularities are not attributed to randomness (Popper, 2002). Even though any finite sample will show some variation, we can decide upon a limit of the variation. Researchers can use the limit of variation that is allowed as a *methodological rule*, and decide whether a set of observations falls in a ‘forbidden’ state of the world, or in a ‘permitted’ state of the world, according to some theoretical prediction.

This *methodological falsification* (Lakatos, 1978) is clearly inspired by a Neyman-Pearson perspective on statistical inferences. Popper (2002, p. 168) acknowledges feedback from the statistician Abraham Wald, who developed statistical decision theory based on the work by Neyman and Pearson (Wald, 1992). Lakatos (1978, p. 25) writes how we can make predictions falsifiable by “specifying certain rejection rules which may render statistically interpreted evidence 'inconsistent' with the probabilistic theory” and notes: “this methodological falsificationism is the philosophical basis of some of the most interesting developments in modern statistics. The Neyman-Pearson approach rests completely on methodological falsificationism”. To use methodological falsification, Popper describes how empirical researchers need to decide upon an interval within which the predicted value will fall. We can then calculate for any number of observations the probability that our value will indeed fall within this range, and design a study such that this probability is very high, or that it’s complementary probability, which Popper denotes by ε, is small. We can recognize this procedure as a Neyman-Pearson hypothesis test, where ε is the Type 2 error rate. In other words, high statistical power, or when the null is true, a very low alpha level, can corroborate a hypothesis.

Popper distinguishes between subjective probabilities (where the degree of probability is expressed as feelings of certainty, or, belief), and objective probabilities (where probabilities are relative frequencies with which an event occurs in a specified range of observations. Popper strongly believed that the corroboration of tests was based on Frequentist, not Bayesian, probabilities (Popper, p. 434): “As to degree of corroboration, it is nothing but a measure of the degree to which a hypothesis h has been tested, and of the degree to which it has stood up to tests. It must not be interpreted, therefore, as a degree of the rationality of our belief in the truth of h [the hypothesis]”. For a scientific realist, who believes the main goal of scientists is to identify features of the world that corroborate or falsify theories, what matters is whether theories are truthlike, not whether you *believe* they are truthlike. As Taper and Lele (2011) express this viewpoint: “It is not that we believe that Bayes' rule or Bayesian mathematics is flawed, but that from the axiomatic foundational definition of probability Bayesianism is doomed to answer questions irrelevant to science. We do not care what you believe, we barely care what we believe, what we are interested in is what you can show.” Indeed, if the goal is to identify the presence or absence of features in the world to develop more truth-like theories, we mainly need procedures that allow us to make choices about the presence or absence of these features with high accuracy. Subjective belief plays no role in these procedures.

To identify the presence or absence of features with high accuracy, we need a statistical procedure that allows us to make decisions while controlling the probability we make an error. This idea is translated into practice in hypothesis testing procedures put forward by Neyman and Pearson (1933): “We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis. But we may look at the purpose of tests from another view-point. Without hoping to know. whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.” Any procedure with good error control can be used (although Popper stresses that these findings should also be replicable). Some authors prefer likelihood ratios where error rates have maximum bounds (Royall, 1997; Taper & Ponciano, 2016), but in general, frequentists hypothesis tests are used where both the Type 1 error rate and the Type 2 error rate are controlled.

Meehl (1978) believes “the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology”. Meehl is of this opinion, not because hypothesis tests are not useful, but because they are not used to test risky predictions. Meehl remarks that “When I was a rat psychologist, I unabashedly employed significance testing in latent-learning experiments; looking back I see no reason to fault myself for having done so in the light of my present methodological views” (Meehl, 1990a). When one theory predicts rats learn nothing, and another theory predicts rats learn *something*, even Meehl believed testing the difference between an experimental and control group was a useful test of a theoretical prediction. However, Meehl believes that many hypothesis tests are used in a way such that they actually do not increase the verisimilitude of theories are all. If you predict gender differences, you will find them more often than not in a large enough sample. Because people can not be randomly assigned to gender conditions, the null hypothesis is most likely false, not predicted by any theory, and therefore rejecting the null hypothesis does not increase the verisimilitude of any theory. But as a scientific realist, Meehl believes accepting or rejecting predictions is a sound procedure, as long as you test risky predictions in procedures with low error rates. Using such procedures, we have observed an asymmetry in the Stroop experiments, where the interference effect is much greater in the color naming task than in the word naming task, which leads us to believe the theory that takes into account the salience of the word and color dimensions has higher truth-likeness.

From a scientific realist perspective, Bayes Factors or Bayesian posteriors do not provide an answer to the main question of interest, which is the verisimilitude of scientific theories. Belief can be used to decide which questions to examine, but it can not be used to determine the truth-likeness of a theory. Obviously, if you reject realism, and follow anti-realist philosophical viewpoints such as Fraassen’s constructive empiricism, then you also reject verisimilitude, or the idea that theories can be closer to an unobservable and unknowable truth. I understand most psychologists do not choose their statistical approaches to follow logically from their philosophy on science, and instead follow norms or hypes. But I think it is useful to at least reflect upon basic questions. What is the goal of science? Can we approach the truth, or can we only *believe* in hypotheses? There should be some correspondence between your choice of statistical inferences, and your philosophy of science. Whenever I tell a fellow scientist that I am not particularly interested in evidence, and that I think error control is the most important goal in science, people often look at me like I’m crazy, and talk to me like I’m stupid. I might be both – but I think my statements follow logically from a scientific realist perspective on science, and are perfectly in line with thoughts by Neyman, Popper, Lakatos, and Meehl.

A final benefit of being a scientific realist is that I can believe it is close to 100% certain that this blog post is wrong, but testing my ideas against the literature, it seems to have pretty high verisimilitude. Nevertheless, this is a topic I am not an expert on, so use the comments to identify features of my blog that are incorrect, so that we can improve its truth-likeness.

References

Cevolani, G., Crupi, V., & Festa, R. (2011). Verisimilitude and belief change for conjunctive theories. *Erkenntnis*, *75*(2), 183.

Feyerabend, P. (1993). *Against method* (3rd ed). London ; New York: Verso.

Kuipers, T. A. F. (2016). Models, postulates, and generalized nomic truth approximation. *Synthese*, *193*(10), 3057–3077. https://doi.org/10.1007/s11229-015-0916-9

Lakatos, I. (1978). *The methodology of scientific research programmes: Volume 1: Philosophical papers* (Vol. 1). Cambridge University Press.

Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science*, *48*(1), 19–49.

Meehl, P. E. (1978). Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.

Meehl, P. E. (1990b). *Corroboration and verisimilitude: Against Lakatos’ “sheer leap of faith.”* Working Paper, MCPS-90-01). Minneapolis: University of Minnesota, Center for Philosophy of Science. Retrieved from http://meehl.umn.edu/sites/g/files/pua1696/f/146corroborationverisimilitude.pdf

Melara, R. D., & Algom, D. (2003). Driven by information: A tectonic theory of Stroop effects. *Psychological Review*, *110*(3), 422–471. https://doi.org/10.1037/0033-295X.110.3.422

Neyman, J., & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *231*(694–706), 289–337. https://doi.org/10.1098/rsta.1933.0009

Niiniluoto, I. (1998). Verisimilitude: The Third Period. *The British Journal for the Philosophy of Science*, *49*, 1–29.

Niiniluoto, I. (1999). *Critical Scientific Realism*. Oxford University Press.

Oddie, G. (2013). The content, consequence and likeness approaches to verisimilitude: compatibility, trivialization, and underdetermination. *Synthese*, *190*(9), 1647–1687. https://doi.org/10.1007/s11229-011-9930-8

Popper, K. R. (2002). *The logic of scientific discovery*. London; New York: Routledge.

Psillos, S. (1999). *Scientific realism: how science tracks truth*. London; New York: Routledge.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. London ; New York: Chapman and Hall/CRC.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643.

Taper, M. L., & Lele, S. R. (2011). Philosophy of Statistics. In P. S. Bandyophadhyay & M. R. Forster (Eds.), *Evidence, evidence functions, and error probabilities* (pp. 513–531). Elsevier, USA.

Taper, M. L., & Ponciano, J. M. (2016). Evidential statistics as a statistical modern synthesis to support 21st century science. *Population Ecology*, *58*(1), 9–29.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford : New York: Clarendon Press ; Oxford University Press.

Wald, A. (1992). Statistical Decision Functions. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics* (pp. 342–357). Springer New York. https://doi.org/10.1007/978-1-4612-0919-5\_22