A vindication of psychological science

By Ludvig Renbo Olsen, June 7th, 2016

In an attempt to directly replicate 100 experimental and correlational studies from 3 important psychology journals, the Open Science Collaboration only succeeded in replicating 38% of the effects tested (Open Science Collaboration, 2015). This was the culmination of the so-called "replication crisis" that has brought into question the methodological and theoretical foundation of psychology and related fields. In this essay we will discuss the implications of these findings and it will be argued that Cognitive Science is a new paradigm of psychology and that interdisciplinarity will be one of the ways forward for psychology.

To understand the scope of the crisis we will start by creating a common understanding of what science is. There are many perspectives to this matter, so we will only cover some of the most influential ones in contemporary science, including the problem of induction, Popper's falsificationism, Kuhn's scientific paradigms, and the scientific method. This understanding will be used to discuss the validity of statistical sciences and will allow us to demarcate science in novel way, including psychology as science.

The journal Nature (Baker, 2016) surveyed 1,500 scientists from natural sciences and found that 52% thought there was a significant reproducibility crisis in science, while 38% thought there was a slight crisis. Along with other findings from the survey, this indication that the crisis expands far beyond psychology will be used to question the structures of our scientific institutions, such as systems of funding. Two novel types of scientific journals will be suggested as tools for promoting objectivity in science.

Science

A hunt for new discoveries and intellectual challenges, science is driven by human curiosity. We want to deepen our understanding of the natural world and use this knowledge to better our lives. It is a belief-system that constantly challenges itself in the aim of getting closer to the truth. A truth that may never be reached as human perception does not seem geared to experience the pure objective world, making it a tough job to develop methods able to reveal and confirm this true nature of nature.

One of the hardest questions facing science is that of when something is to be considered a science. There must be some demarcations of when a discipline is a science, a pseudo-science, or a non-science, so that we can disqualify theories of the latter two and meaningfully base our actions and future research on previous findings. If we do at some point arrive at truth we will only recognize it if we are capable of trusting the methods used to arrive at it.

By constantly challenging the ground on which we stand, the principles of good science slowly develop and form the mindsets and practices of scientists, while discarding those previous findings and assumptions that do not live up to the principles. Science keeps moving and new eras with scientific breakthroughs become possible, leading to large structural changes in our societies and in human lives - often for the better.

Science uses scientific method (that we will cover later) to objectify knowledge of the world. Though subjective experience and motive seem impossible to exclude entirely, it is supposed to be downplayed to gain validity for one's findings. Problematically though, many funding systems emphasize publication, meaning that scientists risk loosing their jobs, their statuses and possibly future chances of making meaningful contributions to science, if they do not "find" publishable results. This obviously affects the choice of methods and scientific principles, as we will discuss later.

It is vital for the role of science in society and thus science itself that leaders are able to trust the science on which they base their decisions. If science looses its public legitimacy it will require a much greater education of leaders and the people to distinguish between good and bad science to avoid handpicking of theories on the one criterion that it fits a political agenda. On matters such as climate change, the few scientists disagreeing with the majority are highlighted for convenience. Finding opposing arguments to the view of the majority is not itself a problem, but the dismissal of the majority is, as it signals that most scientists on the subject are wrong and that the political leaders are better capable of evaluating the science in the field than those educated to do so. The legitimacy of science is thus transferred to the gut feelings of the often ill-educated public and politicians.

In later parts of this essay we will discuss how science can establish better practices to become more transparent to both scientists and the public.

The problem of induction

We humans constantly try to predict the world. We need to know the outcome of eating a certain food or jumping in front of a car. This knowledge is vital for our survival. But logically we cannot predict the future. We cannot know that the plant that gave us food poisoning will give us food poisoning again. We cannot know that the car will continue its movement in our direction and hit us. I cannot know that all humans need to predict the world in order to survive. These are only guesses based on our experience, which seem to work a lot of the time.

This is the problem of induction, acknowledged by David Hume (Vickers, 2016). That, since we cannot experience all fish in the sea, we cannot infer from even the sum of all human experience that all fish are capable of swimming. (Unless, of course, we use this capability as part of our definition of a fish. But then "do all fish swim?" seizes to be an empirical/inductive matter.)

An example of a valid argument form would be this:

Premise 1: If P, then Q

Premise 2: P

Conclusion: Q

E.g. "if our theory P is true, then we will observe Q. Our theory P is true, therefore we will observe Q."

When using induction this is what we do instead:

P1: If P, then Q

P2: Q

C: P

E.g. "if our theory P is true, then we will observe Q. We observe Q, therefore our theory P is true."

This is not logically valid since, even if the premises are true, it is perfectly possible that the conclusion is false. So we infer a general rule when we cannot.

According to the views of Hume, we humans believe that P is true because we have observed Q multiple times and without exception. It is this observed regularity (that Hume calls a “constant conjunction”) that we in our day-to-day lives use to try to predict the future (Vickers, 2016). Thus it would seem that induction is a necessary tool, as without it we would meet every new situation with an empty mind. But there is tension between our need for inductive inferences and its logical deficiencies. Can we ever justify relying on inductive inferences? Even if we attempt, by arguing that induction is what has helped us in the past, this adds induction as a premise in the defense of itself, making it a circular argument.

Falsificationism

If we cannot infer from experience what is true about the world beyond the observed, how can we then do anything in science? Karl Popper, wanting to find the demarcations between science and non-science, believed that induction should not be a part of science. Instead of verifying theories by gathering empirical evidence, scientists should think up scenarios where the theories could be rejected using deduction. Theories with high falsifiability, i.e. that one could think up many falsifying scenarios, and which had been subject to many of such tests without being falsified could then be used in science, until rejected or replaced by a better theory, but never confirmed as a fact (Vickers, 2016). A form of Darwinian evolution, the theories most strongly tested and unfalsified survives (Creative Conflict Wisdom, 2010)

It is questionable, however, whether Popper can consistently prefer one unfalsified theory to another without relying on induction. Where induction tries to infer a generality from the observed data, Popper instead ascribe value to the attempts of falsification in order to rank theories as closer to, or further from, truth (verisimilitude). The more a theory has been attempted falsified, the closer it is to truth (“The Problem of Verisimilitude”, n.d.) But whether a theory (assuming it is false) withstands attempts of falsification or not, might be coincidental, as a theory having withstood an attempted falsification has actually just been verified inductively. Preferring one theory to another based on inductive verifications is, according to Popper, completely irrational, and so Popper seems forced to consider unfalsified competing theories as equals.

A research hypothesis looking to falsify a theory would be “if P is true, then Q cannot be false.” If one then finds that Q is false, P must be rejected. Popper did though distinguish between the logic of falsifiability and its applied methodology. In reality errors do occur and as such it usually takes more than one conflicting instance to reject a theory. Popper also noted, that before abandoning the present theory, a better theory must be established, as science must involve theoretical progress (Thornton, 2015).

Scientific paradigms

Thomas Kuhn also wrote about the substitution of theories. He found that some historians of science wanted to label rejected theories from the past as scientific, as the approaches of attaining them had been just as scientific as those of today. Kuhn concluded that "out-of-date theories are not in principle unscientific because they have been discarded" (Kuhn, 1996, p. 3) In order to study the old theories in relation to the era in which they were born, Kuhn established the notion of scientific paradigms by describing the 3 stages that sciences historically have gone through. In the first stage, pre-science, observations are made, data are collected, and ideas of the world are created without an overall framework for thinking. Then in the second stage a paradigm is established. A way of thinking about the subject is shared by scientists that attempt to solve the same problems. Kuhn described this as follows: "Men whose research is based on shared paradigms are committed to the same rules and standards for scientific practice."(Kuhn, 1996, p. 11) Because of this cognitive bias, this stage can slow down the rate of big findings as these often lie outside the realm of normal thinking. At one point this paradigm will be broken down by substantial new findings that totally dismiss the, now, old way of thinking. This is the revolutionary stage(Kuhn, 1996) that leads to a new cycle often including pre-scientific methods.

Kuhn thought it was wrong to evaluate the quality of a theory using the standards of evaluation from a different paradigm. He saw incommensurability between the two, as the standards of evaluation are themselves subject to change (Bird, 2013). Differing methods and use of language could easily make what was once the best science around seem crazy and non-scientific. Instead we should judge from the practices and assumptions in the paradigm from which the theory came.

The scientific method

In many science textbooks there is a description of a single common method, "The Scientific Method", used to conduct science. It is a procedure of four or five steps leading the scientist from observation to theory. First the scientist will observe or imagine something interesting, perhaps in a problem-solving situation. From this a testable hypothesis will be formulated. An experiment capable of testing the hypothesis will be designed and run, before conducting an analysis of the results using statistical tools. The scientist will then discuss the implications of the findings to previous theories and if suitable formulate a new theory (Friedenberg and Silverman, 2011; Andersen and Hepburn, 2015).

While this idea of a universal method used by all fields of science is pleasing when explaining science and when needing to effectively demarcate science from non-science, this regimentation has its disadvantages such as excluding important scientific operations. Exploratory experimentation and descriptive science are both very important ways of building knowledge. Exploratory experimentation is often based on theory but without the restrictions of a hypothesis. One may consider it a pre-scientific element and that findings should either be tried falsified in a new hypothesis-driven experiment or labelled descriptive science, simply stating what has been observed. The latter might later inform theories and models of the world as scientists, or computers, recognizes patterns in the data (Andersen and Hepburn, 2015).

Astronomers and historians would have a hard time without exploratory descriptive work. These fields attempt to study very large and old systems by exploring the available snapshots. To try to make sense of the observed, they tend to use abduction, meaning that they look for the best and simplest explanations available. These serve as testable theories, though technological limitations or lack of new findings could delay the testing process. An example could be the travel of the human species from Africa to the other continents. By dating findings of humans' physical remains it has been possible to estimate some of our travel details. It only requires a single finding to falsify this though. If we find human remains in a different continent - or possibly even on a different planet - dating further back in history than those found in Africa, we would have to rewrite history.

Psychology

Returning to our focus on the crisis in psychology it seems relevant to establish what psychology is. In this section we will look at the aims and methods of psychology and compare them to those of cognitive science. Cognitive science is an interdisciplinary science spanning psychology, artificial intelligence, linguistics, neuroscience, anthropology, and philosophy (Thagard, 2014). It will be discussed whether cognitive science is simply a paradigm of psychology.

Friedenberg and Silverman (2011) defines psychology as “the scientific study of mind and behavior”(p. 57) with a focus on “internal mental events, such as perception, reasoning, language, and visual imagery”(p. 57) and external events, the latter being human behaviour. Knowledge is acquired using the scientific method discussed previously, while also utilizing more qualitative data from introspection. Introspection, simply put asking participants to describe their subjective experience, has been criticized for not being generalizable, and while one could argue that it is simply a matter of gathering enough qualitative data and treating them as quantitative data, e.g. by finding patterns using classification algorithms, this is not the type of investigation that the Open Science Collaboration tried to replicate and it will not be further discussed in this essay.

The history of psychology is a brilliant example of Kuhn's description of paradigms as it has been filled with movements with each their own theories and practices. A roughly described example, behaviourism was the idea that we should only study measurable external events as we could not objectively study the contents of the mind at the time. This paradigm led to many wrong conclusions but left us with both practices and aims of objectifying science. It was abandoned because many important aspects of human and animal behaviour could not be understood without including the mind in the investigations and theories. Some have argued that behaviourism have gained new momentum in neuroscience with its focus on behaviour, this time the measurable mechanisms of the brain (Thompson, 1994). The view that "mind" is simply what the brain does.

Cognitive psychology was the counterrevolution to behaviourism. It was focused on mental processes and utilized advances in technology to examine the brain *in vivo*. The mind-as-a-computer metaphor, in which the brain is seen as a modular information processor, became the framework, aiming to understand how the brain represents, processes, and stores information. Theories were set up as process models using boxes and arrows to visualize information flow and processing. Each box in these models could be tested and falsified, making it easy to understand, teach, and investigate current theories. Models could further be simulated by computer algorithms as a supplementary experimental tool (Friedenberg and Silverman, 2011).

When considering cognitive science an umbrella discipline containing psychology, it suggests that methods and aims of psychology are embedded. Especially the computational theories have formed the mindsets of cognitive scientists. As such it is a symbiotic relationship and not based on a revolution. With the exception of philosophy, the other disciplines of cognitive science have in a way all sprung from psychology. Neuroscience tries to find the neural correlations to theories of the mind. Linguistics study language, which must be considered a big part of human psychology. Anthropology could be considered a kind of social psychology. Only artificial intelligence seems to deviate from the aims of psychology by aiming to create intelligent machines instead of investigating human intelligence. But we must ask, if exploration of possible systems of intelligence could not to a high degree impact theories of psychology? All these disciplines could thus be considered a study of mind and/or behaviour. The role of philosophy seems just as justified in psychology as in cognitive science, and so, even if not a result of a scientific revolution, cognitive science could be considered a paradigm of psychology. That said, the paper describing the replication crisis that this essay wishes to investigate has not drawn such conclusion and so we will treat psychology as a subfield of cognitive science unless otherwise stated. The interdisciplinarity of cognitive science will nonetheless be a vital part of developing the questions asked, and methods used, by psychology. The awareness of the replication crisis is in itself a healthy sign, as a science must continuously question itself. By looking to other sciences, such as those of cognitive science, problems and possible solutions become more apparent. One might even suggest the performance of a larger meta-analysis of all disciplines to possibly inspire creation or adoption of stronger practices. At the same time it could lead to stronger demarcation criteria for separating science, pseudo-science or non-science.

The crisis

Many sciences such as psychology use statistics to find "significant" patterns in the world. In much contemporary literature a probability threshold at 0.05 is accepted, meaning that if the null-hypothesis is only 5%, or less, likely to be true, it suggests a causal or correlative relationship between the measured variables. Calculation of the probability value is affected by the amount of data collected and how it is distributed. As such you will potentially have a better value to report, the more you have sampled nature.

With our coverage of Popper it should, though, be apparent that it is not possible to falsify a statistical science. An attempt to falsify must include a hypothesis yielding a binary result. Statistical sciences predict that a pattern will be true at least 95%[1] of the time, which is indeed a binary prediction as findings will either be within or outside this threshold. But the falsification of a statistical hypothesis will itself rely on induction, thus making the falsification invalid. If the null-hypothesis is found to be more probable than the set threshold, a falsification will rely on the inductive premise that the sample is representative and not just accidental. As such, statistical sciences are not deductive, which, according to Popper's demarcation criterion, means that they are not sciences at all.

1 Other common thresholds exist but for simplicity we will stay with 0.05

If the statistics-based parts of psychology are to be considered pseudo-science it makes no sense to discuss the lack of replicability, since this is a demand of sciences. Considering the importance of many statistical disciplines to society, we shall therefore, for the sake of the argument, instead attempt to find an acceptable procedure demarcating these as science.

Psychology becomes science

If we accept induction as an unavoidable part of science, must we then accept that science is illogical? This will make it harder to defend science as more legitimate than pseudosciences and non-sciences, as science will be just another belief-system, a matter of faith.

Actually no. There is a way to combine Popper's demand of logical validity with statistical sciences.

First, to make induction a valid form of inference, we will assume axiomatically that nature generally has a high degree of uniformity. This will allow inductive inferences to have the following form:

P1: If P, then Q
P2: P
C: Q
E.g., "if our sample shows x (P), x is likely to be true beyond our sample (Q). Our sample shows x (P), therefore it is likely that x is also true beyond our sample (Q)"

In contrast to the previous examples of argument forms, P is now the particular observation and Q is the general theory.

This is formally valid but still an informal fallacy, since we generalize from particulars. However, assuming the uniformity of nature, i.e. that our sample is representative for instances beyond the sample, we are no longer fallacious in generalizing from particulars.

As the rationale behind Popper's falsificationism was that induction is invalid, we have eliminated the need for falsification, making statistical science possible.

Boiling water is a fine example from the literature (Garvey and Stangroom, 2013). To our knowledge it has always held true that water boils at 100°C. By assuming uniformity of nature, we get to use this information to guide our actions.

To invoke uniformity of nature as an underlying assumption in an argument, certain conditions must be met. There must be sufficient empirical data that was randomly sampled, and interpretation must be (attempted) unbiased. The experiment must, to a certain degree, be replicable. If these conditions are not met, uniformity of nature cannot be assumed.

Even though our inductive inference is justified, our conclusions could still be wrong. Hence we will assume fallibilism, i.e. that we could be wrong about everything and that we are ready to reject all our beliefs.

As much of psychology is not based on falsification, we need a new demarcation criterion. A tentative, but immediate, criterion is that psychology must use formally and informally valid argument forms. Assuming uniformity of nature allows for this when using induction. We cannot prove this assumption though. It is a fundamental assumption in the contemporary paradigm of psychology; necessary for making useful scientific progress, until we find a useful alternative that unnecessitates it.

With this new demarcation criterion, we turn to the implications of the replication crisis in psychology. For a statistical statement to be formally and informally valid, it must invoke uniformity of nature, as this is what makes induction valid. To do so, it must be replicable. If it is not replicable, it is fallacious, and so, it is not science.

By paradigmatically assuming uniformity of nature and fallibilism, we have made induction logically valid and thereby both solved Hume's problem and fulfilled Popper's demands of logical consistency. This has made Popper's demarcation criterion insufficient, why we have replaced it with a novel criterion that allows psychology and other statistical disciplines to be sciences.

Psychology can rely on induction, but statistical statements must live up to certain requirements, such as replicability. Hence the replication crisis is merely a technical problem, and to solve this, we finish off with a couple of solutions.

Solutions

Too much variety between participants could very well make it too hard to make generalizations about many of the questions psychology examines. If considering the mind as a dynamical system there are simply too many variables to control for. The interdisciplinarity of cognitive science could help attack this by discovering new methods for measuring these variables and by suggesting ways of narrowing the breadth of the inquiries.

In the survey conducted by Nature (Baker, 2016), which was mentioned in the introduction, it was apparent that the replication crisis extends beyond psychology and even beyond statistical sciences. More than 70% of the responding researchers had tried and failed to replicate another scientist's experiments, though less than 20% had ever been contacted by another researcher unable to replicate their work (Baker, 2016). While this could be taken as evidence that nature is not uniform, abandoning this assumption seems a last resort due to the implications it would have on science.

More than 60% responded that pressure to publish and selective reporting always or often contributed to problems with replication (Baker, 2016).

Selective reporting can take many forms but is basically the selection of results that one believes publishable. This is often motivated by the pressure to publish, indicating that the funding systems are to a large degree a part of the problem. Publishing articles does though appear to be the only systematised way for scientists to share details of their research. A promotion of openness about such details of tax-paid projects is positive, and so it seems very reasonable to demand that funded scientists publish their work. The question then is, if it has to be significant new discoveries that are published? If less than half of these are reproducible, does it not only clutter the field, making it untransparent to even experienced researchers? It seems incompatible to expect researchers to promote objectivity if reporting objective results can have such a negative influence on the researcher's career. Therefore we will now discuss how two new types of journals could lessen the problem.

1. The journal of independently replicated research. This journal should only accept results that have been replicated by 3 or more independent research groups. This will help researchers find more trustable foundations for their future research and formulations of theories. To be accepted in this journal, the original investigators will have to maximize chances of replicability in all steps of the scientific method. Researchers failing to replicate already accepted results should be accepted fairly easy and appended the original article.

2. The journal of non-significant results. If researchers have a place to check if what they are about to investigate has been investigated previously, it minimizes the risk of wasting time and money. Research should still be peer-reviewed to secure high quality science, but by removing the need for significant results, it increases objectivity in science and helps researchers live up to the demands accompanying funding. It must be pressured that publication in this journal should be considered as prestigious as in other journals.

In Nature's survey, nearly 90% of respondents listed "more robust experimental design", "better statistics", and "better mentorship" as approaches to improve replication success in science (Baker, 2016). This indicates a need to focus on better education of scientists. With the two new types of journals it is the hope that educational institutions and mentorships, along with already established researchers, will be incentivised to focus on performing quality science instead of certain p-values and effect sizes. This will hopefully lead to better replicability and hence help strengthen the legitimacy of science.

References

Andersen, H. & Hepburn, B. (2015, Winter). Scientific Method. The Stanford Encyclopedia of Philosophy. Retrieved from http://plato.stanford.edu/archives/win2015/entries/scientific-method/

Baker, M. (2016, May 25). 1,500 scientists lift the lid on reproducibility. Nature. Retrieved from http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Bird, A. (2013, Fall). Thomas Kuhn. The Stanford Encyclopedia of Philosophy. Retrieved from http://plato.stanford.edu/entries/thomas-kuhn/

Creative Conflict Wisdom. (2010, August 14). Thinkers on Conflict: Popper versus Kuhn [Blog Post] Retrieved from https://creativeconflictwisdom.wordpress.com/2010/08/14/thinkers-on-conflict-popper-versus-kuhn/

Friedenberg, J., & Silverman, G. (2011). Cognitive science: An introduction to the study of mind. Sage.

Garvey, J., & Stangroom, J. (2013). The Story of Philosophy. Quercus.

Kuhn, T. (1996). The Structure of Scientific Revolutions, 3E.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.

Thagard, P. (2014, Fall). Cognitive Science. The Stanford Encyclopedia of Philosophy. Retrieved from http://plato.stanford.edu/archives/fall2014/entries/cognitive-science/

The Problem of Verisimilitude. (n.d.). [Information on a page]. Retrieved from http://philosophy.wisc.edu/forster/220/notes_10.html

Thompson, R. F. (1994). Behaviorism and neuroscience.

Thornton, S. (2015, Winter). Karl Popper. The Stanford Encyclopedia of Philosophy. Retrieved from http://plato.stanford.edu/entries/popper/

Vickers, J. (2016, Spring). The Problem of Induction. The Stanford Encyclopedia of Philosophy. Retrieved from http://plato.stanford.edu/entries/induction-problem/