The conceptual, cunning and conclusive experiment in psychology

early draft (February 19, 2020)

Stijn Debrouwere

Abstract

The ideal experiment in physics must be conceptual, cunning and conclusive. Adoption of these same standards in psychology has led to experiments that are uninformative and frivolous.

Contents

Introduction

A psychologist wonders whether people eat more when portions look smaller, and with some tinkering is able to engineer a soup bowl that surreptitiously replenishes itself. It does in fact cause people to eat more soup.

A psychologist wonders if natural selection might have produced heuristics for assessing risks and probabilities that work well in realistic settings but are awful at equivalent problems stripped of their context. An online survey queries test subjects on their probabilistic aptitude in various scenarios, and that survey vindicates the hypothesis that we are bad at solving abstract problems, devoid of context.

A psychologist wonders what comes first: thoughts or feelings, so she has people take either a leisurely stroll or a brisk run before meeting with a handsome experimenter, then compares the average level of arousal in both groups. If these differ, feelings are clearly post-hoc interpretations of a pre-existing bodily state.

Seen from one perspective, these three vignettes are evidence of a mature science: one that knows how to unravel the mysteries of human behavior through cunning interventions, one that is not content to catalogue human behavior but instead seeks theories that might explain it, one that uses statistical inference to ensure experiments are not merely suggestive, but lend firm support to the hypothesis at hand.

Seen from another perspective, these vignettes represent an experimental science whose interventions lack ecological validity, a science that theorizes freely but misses some of the basic facts, a science that uses statistics to whitewash uncertainty.

At the heart of psychology's ambivalent status as a science stand these three methodological pillars: experiments must be cunning, conceptual and conclusive. We argue that these standards have led to psychological research that is frivolous and uninformative. Yet the irony is that these same features are also those that fuel the successes of theoretical physics. We will attempt to untangle why these ideals work so well in physics, and what goes wrong when these same methodological standards are unwittingly applied to psychological research.

Experiments in physics

Scientific experiments come in all shapes and sizes, but the experiments that speak most to our imagination tend to be conceptual, cunning and conclusive. Before we consider what that means for psychology, let us take a look at how this ideal plays out in theoretical physics, where this scientific style originated.

Experiments serve to confirm a concept, hypothesis or theory and measurements hold no value outside of that confirmatory context.

Newtonian mechanics predicts that light from distant stars passing close by the sun will not travel in a straight line but should be deflected by some minute amount. Albert Einstein concurred, but calculated that light ought to bend by twice that angle. During a solar eclipse in 1919, scientific expeditions to Príncipe and Brazil confirmed Einstein's prediction and corroborated general relativity theory (Coles 2019). These measurements held no intrinsic value. Nobody especially cared about how light would bend in this unique scenario... except insofar as these measurements allowed us to pit classical mechanics against general relativity theory.

Cunning in experimental setup is required because many physical phenomena are elusive and cannot be readily observed nor measured in everyday environments. Furthermore, competing theories nevertheless make identical predictions under a wide

range of conditions and only differ on the fringes, which means not just any experiment will be able to differentiate winner from loser.

It took physicists over 40 years to conceptualize, engineer and build the experimental apparatus required to verify the existence of Peter Higgs' conjectured elemental particle, the Higgs boson. Said apparatus: a particle accelerator that tunnels under France and Switzerland, measures 26.659 meter in circumference, consumes as much electricity as a small city and smashes particles into each other at very near the speed of light.

A single experiment is conclusive, or very nearly so. Hypotheses in physics are often so precisely formulated and the experiments so well-crafted that even a single demonstration carries enormous significance.

Harvey Fletcher and Robert Millikan's careful measurement of tiny ionized oil drops conclusively showed that electric charge is always a multiple of some basic unit, thought to be the charge of a single proton. Many physicists repeated the oil drop experiment after Millikan, but out of marvel, not skepticism.

From physics to psychology

The early psychophysicists showed that human perception is lawlike and was best studied in a controlled setting. The behaviorists brought an appreciation for objective measurement and stressed the need to translate everyday concepts into operational definitions. Educational scientists showed that statistical methods borrowed from agricultural science could avoid arbitrariness in the evaluation of experimental results. (Gigerenzer et al. 1989, ch. 6) Finally, under the aegis of social psychology, these early influences were blended together into a methodology that strongly favors physics-like experimentation as the route to knowledge about human behavior.

As psychology has become more scientific, its experiments have also come to resemble those of physicists.

Thus, the ideal experiment in psychology happens in a lab and takes great care to insulate the phenomenon of interest from outside influences; human behavior is not so much explored or investigated, but rather a hypothesis about human behavior is verified or rejected; and the use of statistical testing is thought to make even an individual experiment sufficiently trustworthy to tell us something about human behavior and to guide public policy.

The archetype of the cunning, confirmatory and conclusive experiment is admittedly something of a methodological PR story, told mostly by philosophers and not taken too seriously by historians of science, but it's a useful myth. In physics, the ideals it embodies help to establish a healthy research culture.

Transplanted to psychology, however, these ideals damage the scientific integrity of our studies rather than reinforcing it, and we will now explain why too much theory is

harmful to psychology, why cunning in experimental setup is unnecessary for psychology and why bold predictions in psychology do not lead to conclusive evidence.

Why too much theory is harmful to psychology

In physics, theory has an uncanny ability to add structure to complex and seemingly random phenomena. But physics is the odd one out. "It is difficult to avoid the impression that a miracle confronts us here", is how nobel laureate Eugene Wigner described the remarkable extent to which a handful of mathematical equations can describe essential aspects of how matter behaves in the natural world (Wigner 1960). In psychology, theory has not been equally effective.

We know many things about how homo sapiens think, behave and interact that we did not know one hundred years ago: we know eyewitnesses are not always reliable and we know how easy it is to generate false memories, we know that you can't judge a person by their handwriting or through the shape of their skull, we know that punishment and criticism almost never leads a student to learn more or more quickly, we know that women and men have equal intellectual capabilities, we know how to ask people about their political preferences and sexual activity even if they'd rather not tell.

Insight into human behavior has grown enormously in the 20th and 21st centuries. Yet for all of this knowledge, no field that studies human behavior has much in the way of good theory: no explanations, frameworks or truths at a basic level that we might build on.

More than a century after Alfred Binet's pioneering intelligence tests, we're still debating whether intelligence is general or multifaceted. Categorization of human personalities according to openness, conscientiousness, extraversion, agreeableness and neuroticism – the big five – is based on a mere statistical regularity with no link to brain or biology. Attempts to formulate laws in economics, such as the Phillips curve that predicts inflation will go up as unemployment goes down and vice versa, tend not to survive when exposed to empirical evidence. Political science rarely generalizes beyond the country in which a theory finds its origin.

Alan Newell, discussing the state of cognitive psychology in 1973, praised the "veritable horn of plenty for our experimental life", the vast amount of psychological phenomena that have been discovered in the 20th century, but at the same time feared that "far from providing the rungs of a ladder by which psychology gradually climbs to clarity, this form of conceptual structure leads rather to an ever increasing pile of issues, which we weary of or become diverted from, but never really settle." (Newell 1973, 7)

Some have claimed that psychology is not amenable to theoretical development because humans are unpredictable, their behavior too variable to summarize into laws and principles. That's probably true, as far as it goes, but a much more fundamental consideration is that psychology is not like physics because both layman and academics

do not *want* psychology to be like physics. We want psychology – even when we study memory, cognition, perception – to be able to teach us something about actual human behavior and individual cases.

Consider these questions that a psychologist might care to investigate:

- Is it better to track students according to their abilities, or are there cognitive or social benefits to putting students of different ability together?
- Is it better to let a baby cry it out, or to provide immediate attention and comfort?
- Can I change my personality, or was I born with it?
- How accurate are our memories?
- Do antidepressants actually work?
- How conformist are people, when are we more likely to stand up to others?
- How do people react when confronted with evidence counter to what they believe?
- What works best, encouragement or criticism?
- Does an addiction ever go away?
- Have people become less racist over time, or are we just better at hiding it?
- What comes first, thoughts or emotions?

You might recognize among these questions some of psychology's most famous experiments. Elizabeth Loftus showed that asking eye witnesses about a *crash* rather than an *accident* would subsequently make them overestimate the speed of a collision. John Darley and Bibb Latané filled a room with smoke and found that, if confederates of the researchers feigned ignorance of the emergency, so would the participant. Cognitive dissonance is the brainchild of Leon Festiger. Anthony Greenwald and his colleagues at Harvard showed that many hold an implicit bias against other races even though they might not display any outward racism. Stanley Schachter and Jerome Singer held that it is how we interpret an arousing event that gives form to our emotional state, rather than the other way around.

On closer evaluation, it should be clear that none of these questions are particularly theoretical, but simply ask whether this or that phenomenon works like this or like that. To be absolutely clear, in all of these cases we may expect that a little theory could guide the empirical investigations and streamline their conclusions. And we'll need a bit of theory to make sense of what we see, for there is no such thing as raw data or unmediated observation. We will need to carry out our investigations according to the method of our choice, and enlist the help of psychometrics and its underlying theories and assumptions. Without *any* theory, science cannot exist. But first and foremost, the questions of psychology are empirical questions that require empirical answers.

There is no reason, no historical precedent and no epistemological argument, to suspect that we will find any great commonalities and structural relationships both between these cases and within each case. Instead, given the level of abstraction at which these questions are posed, it is all but guaranteed that the answers to these questions will depend on culture, environment, circumstance and that the findings and answers – plural

– to each question will follow their own logic.

Insofar as the role of theory is precisely to unearth such structural relationships, in experimental psychology, theory can never be the protagonist but must be relegated to the role of helpful sidekick.

In short, (1) theory in psychology has historically speaking never been a particularly fruitful endeavor, (2) given the variability of human behavior, regularities are hard to find and (3) the kinds of questions we ask are at the wrong level of abstraction for theory to do much good.

Despite this, psychologists remain convinced that one cannot conduct a proper experiment if one doesn't have an accompanying theory. While it is unclear whether this belief is harmful to psychological science, it is also unclear that it contributes much if anything to psychology's ability to generate knowledge.

Why cunning in experimental setup is unnecessary in psychology

In physics, cunning experimental setup is often necessary to bring elusive phenomena to the fore. Benjamin Franklin's attempts to figure out the secret workings of electricity at one point involved electrocution of turkeys, while like-minded scientists toyed around with Leyden jars, induction coils and all number of wonderful contraptions. No particle accelerator, no Higgs boson. In genetics, the short life cycle of fruit flies allows for rapid testing of genetic mechanisms and new techniques for genetic manipulation applicable to *all* plant and animal life.

Educational psychologists, similarly, might test whether adding a bit more struggle to instructional materials and slowing down students aids their understanding not through something a teacher might actually try to implement, e.g. changes to a curriculum that encourage slow, deliberate processing, but instead by comparing regular course materials with materials identical in content but printed in a grayed out and hard to read font. Whether or not a teacher would ever consider intentionally obfuscating their course materials in such a fanciful manner is besides the point. Ecological veracity of the intervention is an irrelevant concern for the theoretically minded researcher, just as it would be for the physicist and her particle accelerator or the biologist and his fruit flies: it's a good intervention if it's a good test of our theory.

When testing a theory, there is no requirement for our test subjects and manipulation to be anything that is intrinsically interesting – certainly, fruit flies are not interesting. Instead we should use manipulations that foreground the defining features of our theory, with an outcome that cannot easily be guessed and that hopefully provides a unique contrast with competing theories.

Unusual and cunning interventions, then, are par for the course in a theoretical science. They are in fact often the *only* way in which we can properly grill a theory. The natural

sciences' academic literature is full of poetic phrases to this end: we must interrogate nature, we must trick nature into giving up its secrets, we must design experiments that let us peek behind the curtains, and so forth.

However, studying one thing in order to learn about another thing is a dangerous bet. In the effort to see a behavior in its purest form, we may induce behavior that would never occur in a natural setting, and therefore can hardly be said to teach us anything about human nature.

If information from *this specific intervention* does not readily flow back to the underlying theory at test, and from there to the entire class of problems that the theory might inform, then you end up with something not very different from testing whether painting purple dots on rhinoceros in Mozambique affects soil quality in Nunavut. In the absence of greater motive, fanciful interventions are no longer science, they are nonsense.

The requirement that information from fanciful interventions will flow to the theory, and back again to all phenomena the theory covers, is hard to uphold in the social sciences. There is simply too much variability in the phenomena at hand, and it is impossible to know how our particular experimental setup and our experimental sample, chosen for convenience and assumed to be equivalent to the population rather than verified to be so, will affect the outcome. (Henrich, Heine, and Norenzayan 2010)

We cannot be assured that, if we investigate what we believe to be a "pure" version of a particular phenomenon, we will observe similar behaviors albeit in attenuated form in more realistic settings.

The relative lack of power of cunning experiments to establish universal truths about human behavior hits particularly hard for psychology, where laboratory experiments are common.

While cunning experimental setups are not a universal feature of psychological research, and while psychologists do sometimes police each other's work to ensure some degree of ecological validity, there remains a particular flavor of psychological research that exhibits a great fondness for the cunning experiment. A quick look at the American Psychological Association's *Particularly Exciting Experiments in Psychology* newsletter shows that fanciful interventions that lack ecological validity can find great appreciation not just in mainstream media but among academics as well.

Such psychological research ends up being useless because, if the motivating theory does not matter, then the intervention does not matter either, and doesn't provide empirical material that may stand on its own.

When a historian provides an account of institutional racism and segregation in early 20th century North Carolina, that's interesting. When a psychologist takes a broad look at the patterns of thought that characterize those with major depression, that's interesting. When another psychologist asks people to read the word "red" rendered green and monitors their reaction time, that's much narrower in scope but still interesting,

and it might teach us a thing or two about the brain. But when a third psychologist does a lab experiment in which they test how well subjects perform on a sudoku puzzle after listening to half a minute of classical music versus half a minute of white noise, that's only interesting if you happen to have some sort of theory that would *make* it interesting.

In the words of Roy D'Andrade:

> As seen by an outsider, the problem of social psychology is not that nothing is being found out–there is considerable new activity each year, and novel findings and new theoretical formulations. The major problems that seem to bedevil the field are those like the degree to which the findings of the laboratory have any relation to the things that people do in ordinary life, the degree to which findings represent historically conditioned cultural attitudes rather than deep psychological process, and the way in which the exploration of a particular phenomenon seems to lead to a dead end, with no general findings outside of the understanding of how to produce a certain limited class of effects. (D'Andrade 1986, 37)

As Kenneth Ring observed long ago, psychological experiments whose prime concern is to test a theory run the risk of becoming weird or zany (Ring 1967). What started out as genuine science more and more appears as a game. But it is only a game because the underlying strategy – trick nature into letting us peek behind the curtains through ingenious tests – does not seem to work as well as we had hoped in the social sciences, certainly not as well as it does in the natural sciences.

Instead, it is probably more fruitful if psychologists model their experiments on those of applied science. When engineers wish to test the crack resistance of different formulations of concrete and different reinforcements, they will test *exactly that* and preferably in conditions that resemble how the concrete will be employed in real building projects. They will rely on technology and theory as needed, but not more than is necessary.

Why bold predictions do not lead to conclusive evidence in psychology

In physics, as we noted earlier, hypotheses are often translated into predictions that are so precise that, if corroborated by the experimental evidence, any other explanation would be literally incredible.

These are the terms of the hypothetico-deductive quid pro quo: make predictions that are very unlikely to hold, and in exchange you will not have to compile an endless book of proof, but instead may take a small handful of positive instances as conclusive evidence of your favored theory.

Other physicists may wish to repeat the experiment to rule out measurement artifacts or other mishaps, and they frequently do, but these are usually limited to close replications

of the original experiment. The weak point of an experiment is usually taken to be the auxiliary hypotheses that hold it up: is the experimental apparatus up to snuff, etc. The fact that universal claims are being made on the basis of a limited test in one particular setting, on the other hand, is not generally a cause of worry.

This is, ultimately, the most fundamental consideration in favor of a theoretical science, with its cunning interventions and its bold predictions: the use of theory allows a shift from plebeian to aristocratic induction (Laudan 1981, ch. 6). Instead of generalizing from many, many individual cases to a general statement, we may instead generalize from a handful of well-crafted experiments to a theory. This is also known as the hypothetico-deductive method, in contrast to the more pedestrian *induction by enumeration*.

The inductive scientist may note: "Never in all the years that I have lived have I seen ice sink in water, so I guess that it never does. *Here's a million row spreadsheet for proof.*"

The deductive scientist may instead say: "Lighter substances float, denser ones sink. Ice is lighter than water, hence ice will float on water. *Here's a quick demonstration.*"

When it works, hypothetico-deductive inference is a brilliant epistemic trick that produces large amounts of information from small amounts of observational feedstock. There is no need to go to Charles Darwinesque lengths to observe nature for years on end before one may venture a guess as to how it all fits together. Instead, a stroke of insight and one well-placed observation could be enough to validate a theory.

Basic science in physics possesses unique epistemic and social features that renders the hypothetico-deductive mode of science not merely possible but fruitful:

- theories make precise predictions which are by definition bold predictions, as the probability of any outcome in a continuous distribution of probabilities is technically zero – an important point to which we will devote our special attention later in this paper
- experimental setup can create a sufficiently closed system to test these predictions and to expect them to hold unconditionally
- social standards that frown upon corrections, auxiliary variables and other ad hoc adjustments to theories that would make them fit the facts
- a theoretical spider web that connects every theory to every other theory such that ad hoc adjustments in one place are likely to wreak havoc and lead to faulty predictions elsewhere, limiting the scope of abuse.

In such an environment, hypothesis testing works exactly as philosophers of science such as William Whewell and Karl Popper said it ought to: subject theories to rigorous tests and quickly dispatch of theories that do not fit the facts, with the minimum amount of fuss.

Psychologists often take it for granted that they may use the hypothetico-deductive mode, but in line with its pedigree, this style of experimental reasoning makes high

demands of our theories and observations, demands that are impossible to meet in experimental psychology.

In psychology, predictions are often so loosely connected to hypotheses, hypotheses so loosely connected to theories, and post-hoc modifications to evidence, theory and hypothesis are so common that it is impossible for experimental evidence to truly favor one theory over another.

Predictions are loosely connected to hypotheses.

When Heinrich Hertz sought to test James Clerk Maxwell's hypothesis that light was a form of electromagnetic radiation, he did so through a series of experiments that proved the existence of electromagnetic waves, that they move at the speed of light and that they "exhibit lightlike behaviors of reflection, refraction, diffraction, and polarization" (Rautio 2014). Hertz did not set out merely to show the *existence* of electromagnetic waves, but also that these waves had all the *properties* Maxwell thought they should have, and that any measurable *quantities* did not deviate from Maxwell's equations.

In the example of Hertz and other famous experiments from physics, a prediction is merely a hypothesis in which one has filled in the blanks, and prediction and hypothesis are tightly coupled.

In psychology, on the other hand, predictions are usually (1) solitary, postulating one particular effect rather than multiple conditions that each must hold and (2) vague, postulating quality but not quantity, or put differently: postulating that a relationship between two variables will be found, but not the magnitude of that relationship.

In theoretical physics, measurement is aimed at *verification*. In experimental psychology, measurement is aimed at informing us about the magnitude of an effect. This, too, is very different. The goal of the Eddington expedition was not to investigate, empirically, how light from distant stars is deflected by the gravity of the sun and add this to our repository of physical facts, because Einstein's theory already *told* us what kind of deflection to expect in this scenario: a little over 0.000483 degrees.

We do not wish to overstate the issue. Mediation analysis narrows down the predictive leeway of experiments in psychology by requiring one to specify the mechanism that links two variables together, not just that a relationship exists. And because experiments in psychology tend to be fairly small, in practice a minimal effect size is implicitly a part of the hypothesis. On the other hand, qualitative research in physics is common, especially in the context of discovery. Nevertheless, predictions in physics and predictions in psychology tend to be on opposite sides of the evidentiary spectrum.

Verifying a bold theoretical prediction is different from estimating an empirical quantity. Nevertheless, scientists in various fields seem to believe they do the former while actually engaged in the latter.

To further illustrate this contrast between physics and psychology, one may compare

how these two disciplines deal with statistical testing. In psychology, we aim to reject a precisely quantified null hypothesis in favor of its unquantified complement ("anything not null"). In physics, the aim is generally to *confirm* the null hypothesis (our theory) and reject the alternative hypothesis (any other theory). It needs little explanation that the latter approach leads to bold conjectures whereas the former provides only very meager evidence in favor of the theory under investigation. (Meehl 1967)

Hypotheses are loosely connected to theories.

Niels Bohr once said that the opposite of a great truth is also true. In psychology we come across such Bohrian paradoxes all the time. For example, we posit that someone with high social standing will tend to do what they can to maintain that standing (our theory) and that they might therefore be inclined to boycott social competitors or withhold access to important resources (our hypothesis)... but if it turns out that instead, those of high social standing tend to be magnanimous, that too is easily explained: generosity is a great way to showcase one's superior status. Our theory aligns equally well with both hypotheses, even though they are diametrically opposed to each other.

Often it's not quite clear what if anything a theory implies for the empirical world. In the worst case, this is a sign of a lack of scientific rigor – if a theory does not predict some and rule out other states of the world, it cannot be falsified, and hence it is not scientific. This was the essence of Karl Popper's critique of early twentieth-century psychology. A methodologist more favorably disposed towards psychology might instead say that our theories' vagueness is merely a reflection of how context-dependent human behavior can be.

Regardless of whether we judge the lack of a strong connection between hypothesis and theory harshly (as Popper did) or whether we consider it to be an unavoidable aspect of the behavioral sciences, indeterminacy and under-identification of theories undermines the abductive argument underlying the hypothetico-deductive style of reasoning. An experiment whose outcome lends support to a hypothesis will not transitively lend support to the theory that inspired the hypothesis if competing theories align equally well with that hypothesis. Abductive reasoning consists of inference to the most plausible explanation, and such inference is impossible if the evidence equally favors multiple explanations.

(To compound the problem, in much of psychology an experiment can hardly be said to lend much support to a hypothesis because of how broad the alternative hypothesis is construed, as we saw in the previous section.)

Theories are loosely connected to other theories.

After studying 19th century science, William Whewell concluded that while successful *predictions* are essential to science, what is really impressive is science's *consilience* (how a theory may inform cases beyond its original scope) and *coherence* (how different theories dovetail with each other). Knowledge does not stand alone, but fits together.

While it is common for experimental psychologists to claim to contribute to a particular research program or to start from a particular framework such as Daniel Kahneman's *heuristics and biases*, in practice results from one experiment rarely if ever make us question the validity of other, unrelated experiments, and theories themselves rarely clash with each other. In physics, it is true that one may explain the working of some kinds of laser beam while blissfully unaware of quantum physics, but on the other hand any theory in fluid dynamics that does not conserve energy will clash with everything we know from thermodynamics. In psychology, such restrictions rarely apply. One may hold the most outlandish views on intelligence without gainsaying prevailing views on working memory, childhood learning, language acquisition and much more.

As a result of these weak to non-existent links between theories, experimental psychologists cannot avail themselves of a kind of sanity check that is very important in physics and that in physics is performed well in advance of any sort of experimental verification: *should I even bother to test this idea, or would it be nonsensical given everything else we know?* In psychology, there are very few hypotheses that one would have to reject prima facie for being incompatible with firmly grounded knowledge about human nature, and in fact journals tend to encourage research that at first blush seems to have a counterintuitive premise, because it fits well with the scientist's rebellious self-image of not succumbing to taboo or conventional wisdom.

Because predictions are loosely connected to hypotheses, hypotheses are loosely connected to theories and theories are loosely connected to other theories, psychology gets very little mileage out of theoretical developments and cannot use experiments to separate good theories from bad ones.

Without a convincing connection between theory and hypothesis, hypothesis and prediction, prediction and observation, we are back where we started: the conditions for successful application of the hypothetico-deductive mode are not fulfilled and we must default to plebeian induction, in which we have to laboriously enumerate positive instances of our claim until, statistically, it counts as sufficient evidence. Plebeian induction takes more time, but it makes fewer demands: its applicability is limited only by our persistence.

Psychologists have often sought to circumvent the requirement for laborious induction by shifting the burden of proof and assuming *ceteris paribus*: unless we have specific reasons to assume that the behavior under investigation would manifest differently (or not at all) in people of different ages, nationalities, intelligence, etc. then we will rely on the old psychological axiom that behavior is universal. This is, of course, wishful thinking.

Summary

Under the right circumstances, we may accelerate the scientific process through carefully chosen interventions, by increasing the level of abstraction from observation to theory, and by deductively testing those theories rather than building up evidence inductively.

When we expect that behavior in the lab will match behavior at large, we may avail ourselves of crafty interventions in lieu of naturalistic observations. When we expect to find strong regularities, theories organize our knowledge. And when those theories make bold predictions, an ingenious intervention may permit us to make sweeping conclusions from a small amount of evidence.

We have earlier referred to this unique mode of science as *aristocratic induction.*

Natural scientists have worked hard over the centuries to ensure that the preconditions for this unique mode of science are met, and when they are, it enables spectacular advances in knowledge.

When, however, the criteria for aristocratic induction are not met, yet we pretend that they are, the result is not one but two steps backwards: research that is lousy as per the standards of aristocratic induction, but equally lousy as per the standards of straightforward empirical science and plebeian induction: what little amount of data is procured is usually irrelevant to everyday behaviors.

Psychologists have been eager to use the hypothetico-deductive method as the ultimate epistemic time-saving hack, but have not generally heeded its requirements. Theories are not clearly tied to observational outcomes. Hypotheses cannot discriminate between theories. Conclusions from the lab do not translate to scenarios of intrinsic interest. These are not failings or shortcomings but rather inevitable *properties* of the study of human behavior. And because of these properties, the hypothetico-deductive method simply does not apply, and stubborn attempts to apply it will pervert science.

Experimental psychology's focus on the confirmation of theories rather than the collection of facts, means that for most kinds of behavior we have all sorts of ideas about what might drive that behavior, but often no idea of how prevalent that behavior is, what encourages and what supresses it. Because a single experiment is taken as conclusive, a lot of purported behaviors and behavioral moderators and mediators might simply not exist. And because the preferred mode of checking other scientists' work is through close replication, and because the original experiments tend to be a bit weird and do not care about ecological validity, these kinds of erroneous beliefs tend to live on much longer than one would think, before being corrected.

What it all means for experimental psychology

A vocal minority of experimental psychologists thinks that there is something wrong with our field (John, Loewenstein, and Prelec 2012), and talk of a replication crisis. They would like to change how psychology operates as a science, make it better. However, proposals to "fix" psychology typically start from the assumption that experimental psychology as an approach to the investigation of human behavior is fine the way it is, it's just that we need to be a bit more strict ($p < 0.005$), a bit more honest (preregistration) and a bit more diligent (replication, meta-analysis).

All such proposals to make psychology more scientific still consider the cunning, confirmatory and conclusive experiment to be the engine for scientific progress in psychology.

However, many in vogue proposals to fix psychology do not get at the heart of the issue and offer only marginal improvements over the status quo:

- greater strictness does not catch unwarranted generalizations,
- replications do not check the validity of operationalizations,
- curbing questionable research practices does nothing to curb our credulity,
- preregistered studies barely increase the severity of a hypothesis test when there is only a loose connection between theory, hypothesis and prediction, such that the results of even a preregistered experiment cannot count as decisive evidence in favor of one theory or another anyway.

The crisis in psychology is more fundamental: psychology lacks a convincing story of how fact relates to theory, what counts as evidence and how strong that evidence is.

Instead, or in addition to more replication, pregistrations, better statistical methods, stricter standards and other proposals to fix psychology, psychology must take an empirical turn and move away from an unhealthy obsession with theory. Psychology must realize that statistical evidence in isolation is not convincing. Psychology must learn again the importance of careful observation and limited generalizations. In this field, a little theory goes a long way.

An empirical psychological science would look for sources of variation rather than trying to eliminate them, would share data and explore it freely, and would encourage null results because variation would be considered the norm, universal applicability of a theory the exception.

An empirical psychological science would not care about whether a theoretical construct such as ego depletion "exists", instead it would try to figure out in what situations the phenomenon does and where it doesn't occur, and whether ego depletion can have a big or a small effect on willpower across a range of settings. Such a science would consider dichotomous evaluations of a theory utterly uninteresting: human behavior is so malleable that *surely* there must be some situation or other in which ego depletion occurs, what really matters is whether it occurs often and whether it explains a substantial

proportion of the variability in (some kinds of) human behavior (Greenwald et al. 1986).

An empirical psychological science would get rid of psychologists' unshakable belief that scientific progress equals theoretical progress. For a psychologist, it is never enough to simply ask a question about human behavior and seek an answer. Instead, we must first search for a theory that might predict this or that kind of behavior, and straightforward empirical findings without theory are considered to be uninteresting.

An empirical psychological science would be more content to explore the great diversity in human behavior and to catalogue how different situations can give rise to different behaviors. It will not try to subsume everything under a theory that will later turn out to be too simplistic to do any good, tested using an intervention too strange to interpret.

Bibliography

Coles, Peter. 2019. "Einstein, Eddington and the 1919 eclipse." *Nature* 568 (7752): 306–7. https://doi.org/10.1038/d41586-019-01172-z.

D'Andrade, Roy. 1986. "Three Scientific World Views and the Covering Law Model." In *Metatheory in Social Science*, 400. Chicago: University of Chicago Press.

Gigerenzer, Gerd, Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty, and Lorenz Kruger. 1989. *The Empire of Chance: How Probability Changed Science and Everyday Life*.

Greenwald, Anthony G., Anthony R. Pratkanis, Michael R. Leippe, and Michael H. Baumgardner. 1986. "Under What Conditions Does Theory Obstruct Research Progress?" *Psychological Review* 93 (2): 216–29.

Henrich, Joseph, Steven J Heine, and Ara Norenzayan. 2010. "The weirdest people in the world?" *The Behavioral and Brain Sciences* 33: 61–83; discussion 83–135. https://doi.org/10.1017/S0140525X0999152X.

John, L. K., G. Loewenstein, and D. Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling." *Psychological Science* 23 (5): 524–32. https://doi.org/10.1177/0956797611430953.

Laudan, Larry. 1981. "Hume (and Hacking) on Induction." In *Science and Hypothesis*, 72–85. Springer Netherlands. https://doi.org/10.1007/978-94-015-7288-0_6.

Meehl, Paul E. 1967. "Theory-testing in psychology and physics: a methodological paradox." *Philosophy of Science* 34: 265–66.

Newell, Allen. 1973. "You Can't Play 20 Questions With Nature and Win: Projective Comments on the Papers of This Symposium." In *Visual Information Processing*. New York: Academic Press. https://doi.org/10.1016/b978-0-12-170150-5.50012-3.

Rautio, James. 2014. "The Long Road to Maxwell's Equations." https://spectrum.ieee.org/tech-history/dawn-of-electronics/the-long-road-to-maxwells-equations.

Ring, Kenneth. 1967. "Experimental Social Psychology: Some Sober Questions about Some Frivolous Values." *Journal of Experimental Social Psychology* 3: 113–23.

Wigner, Eugene P. 1960. "The unreasonable effectiveness of mathematics in the natural sciences." *Communications on Pure and Applied Mathematics* 13: 1–14. https://doi.org/10.1179/030801811X13082311482537.