

MAS.S73 Moving Beyond the Replication Crisis

How to Spot Misleading Social Science and Design Better Experiments

MAS.S73: Moving Beyond the Replication Crisis
Lecture 4: Moving Forward (a pseudo-religious case for DAGs & PPLs)

David Ramsay

I. METHODS

MISAPPLIED STATISTICS

P-Hacking, Publication Bias, File Drawer Effects, HARKing, QRPs, Researcher DOF

II. SYSTEMS

SYSTEMIC INCENTIVES

Fraud, Hype, Motivated Misinterpretation, Overgeneralization, Narrative Support, Status Quo Bias

III. ONTOLOGIES

FUNDAMENTAL ASSUMPTIONS

Psychometrics, Taxometrics, Analytic/Gestalt, Idiographic/Nomothetic, Statistical/Causal Reasoning

Today is about part 3.

Make your Scientific Model Explicit.

Part I: Why?

Part II: How?

Part III: Where does it breaks down?

This is the message that I want to drive home.
why to do that, how to do it, and where it breaks down.

Recap

- The failures of null hypothesis significance testing
- Philosophy of Science: Induction, Causation, and Popperian Falsifiability

Categorical thinking is the cardinal sin of modern statistical practice.

- Meta-statistical, practical tools to evaluate and extract value from existing literature

Methods to tackle each potential problem	
I. Uncommon	Cataloging publications that need to be retracted Disclosure of all data and materials Fraud detection flags
II. Common, but not primary culprit	Pre-registration & journal acceptance Statistical checks Power analysis
III. Very common	Meta-Statistics (Pre-registration/Registered reports) Data Protection Packages - Undercover feasibility in data collection and analysis www.undercover.org.uk/undercover/2012/



Recap.

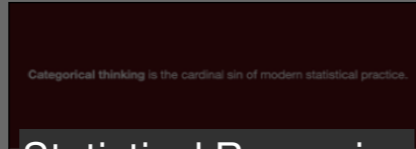
categorical thinking about p-values and conflating it with whether something is true has led to a really sorry state of research literature and our own skills as researchers to evaluate what we read. Issues with how we use statistics.

We looked at how to extract value out of the statistics we have remaining. statistical tests. Blogs. Databases.

Underpinning of philosophy of science. The notion of falsifiability, the problem of induction, randomization, some fascinating ideas about our ability to actually model and generalize complexity.

Recap

- The failures of null hypothesis significance testing

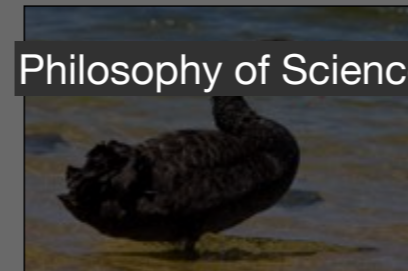


Statistical Reasoning

- Meta-statistical, practical tools to evaluate and extract value from existing literature

Methods to tackle each potential problem	
I. Uncommon	Cataloging publications that need to be retraced Disclosure of all data and materials Practitioner tags
II. Common, but not primary culprit	Pre-registration & journal acceptance Statistical checks Power analysis
III. Very common	Meta-statistics (Pre-registration, Registered reports) Data sharing (e.g., Open Science Framework) Pre-registered analysis www.rosalind.org.uk (Rosalind Institute)


- Philosophy of Science: Induction, Causation, and Popperian Falsifiability



We will be talking about the relationship between statistical techniques we covered in the first class, and how they relate to the philosophy of science Matt covered last class.

DEDUCE	INDUCE
axioms -> specific instances	specific instances -> axioms
A1 = Hume was a philosopher. B1 = All philosophers are great. C1 = All great people support crypto.	a1 = Hume was great and a philosopher. a2 = Plato was great and a philosopher. a3 = Nietzsche was great and a philosopher.
Therefore Hume is great. Therefore Hume supports crypto.	Therefore all philosophers are great. Therefore all great people are philosophers.

PROBLEM OF INDUCTION
(==)
PROBLEM OF CAUSALITY



First, a recap of what Matt covered. Recall the problem of Induction.

Hume also argued there is no such thing as causality, only 'constant conjunction'. This is the same argument as the black swan argument. You come out in the world, flail your arm, you knock over a cup, over and over your whole life, and you assume hitting a cup knocks it over. Causality is an irrational illusion you've created for yourself, same problem as induction.

2 big problems !!!



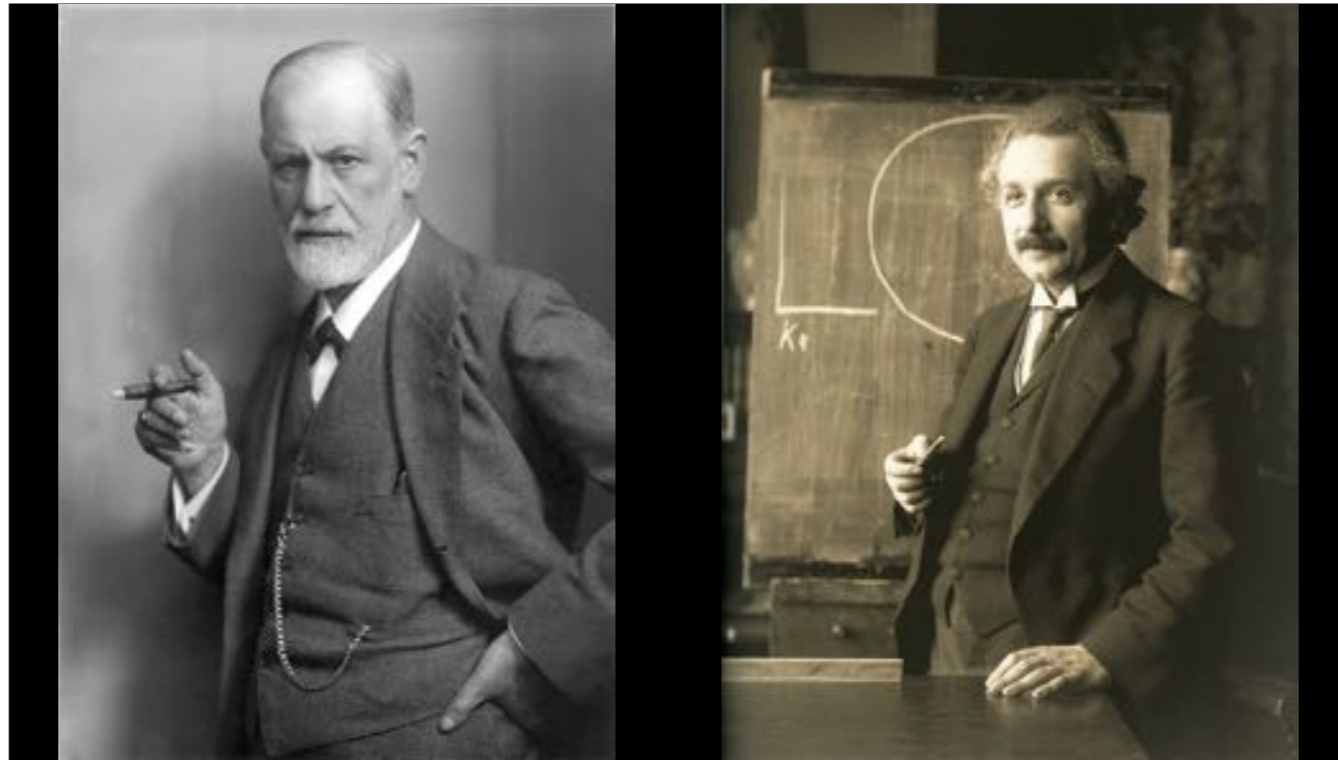
Correlation does not imply Causation.

Might not buy Hume's argument that causation is an illusion. I hope not.

But we might accept the problem of induction applies here, and we can't ever really be rationally sure that contingent, constant conjunction implies causation.

We probably buy this to some extent; that there's some fundamental difficulty going from correlation to causation.

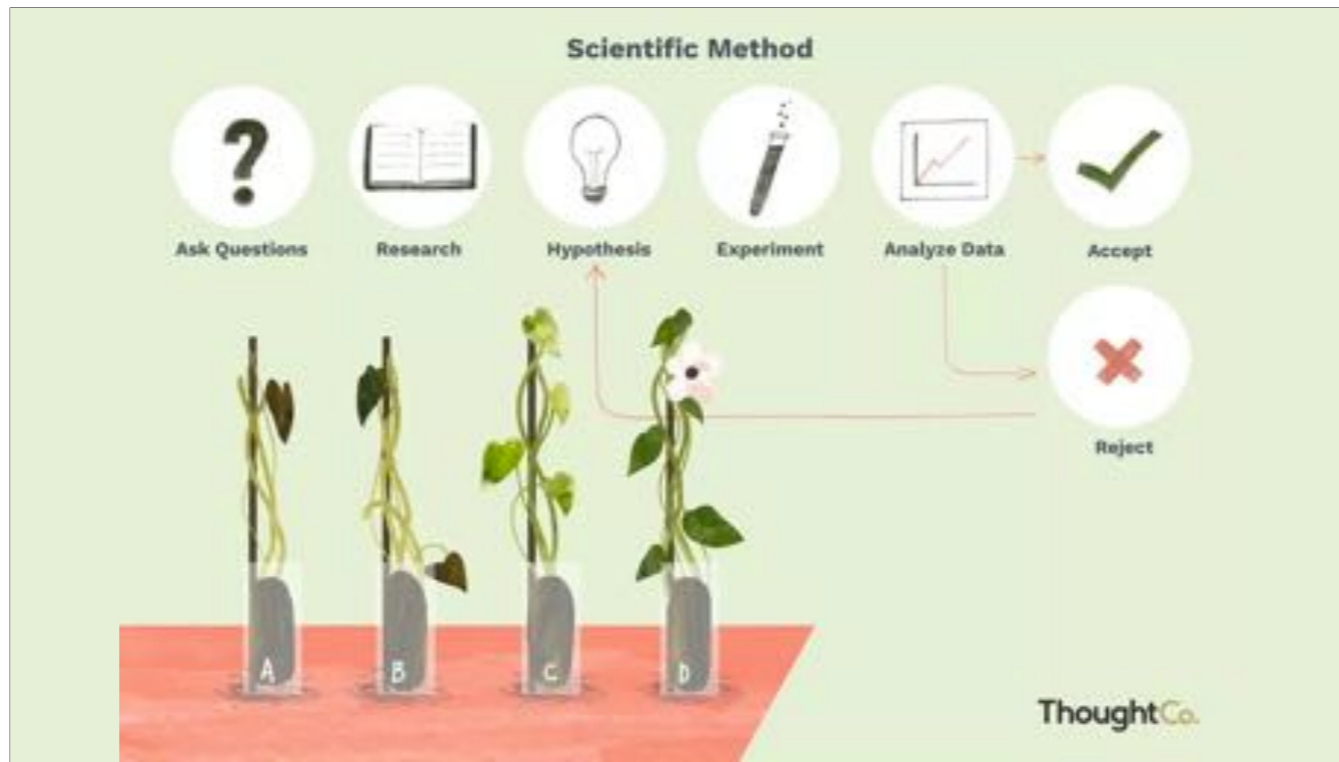
Bracket this, we'll come back. This is a problem we're gonna have to deal with.



Problem of Induction applies to science. Popper comes along and sees these two guys and says one 'feels like science' and one doesn't.

Demarcation problem— both of these people are empiricists. Both of them are reasoning inductively. Freud claiming people's psychopathology is due to latent psychosexual trauma; Einstein claiming experiences of time and space are due to relativity.

Empirical falsifiability separates them. Can't trust induction. failures of falsification corroborate.

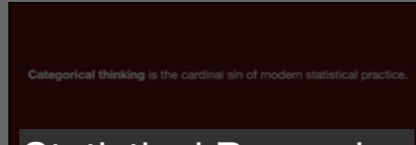


So you must create a theory with falsifiable predictions and test those predictions. Science *is* when you have falsifiable claims, otherwise it's not science.

This is the scientific method we're taught in school. Forget induction; create a fantasy hypothesis and deduce from it a prediction. Can't prove anything. Corroborate a bunch or disprove.

Recap

- The failures of null hypothesis significance testing

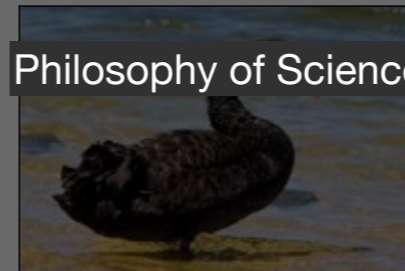


Statistical Reasoning

- Meta-statistical, practical tools to evaluate and extract value from existing literature

Methods to tackle each potential problem	
I. Uncommon	Categorizing publications that need to be retraced Disclosure of all data and materials Practitioner tags
II. Common, but not primary culprit	Pre-registration & journal acceptance Statistical checks Power analysis
III. Very common	Meta-statistics (Pre-registration, Registered reports) Data sharing (e.g., Open Science Framework) www.rosalind.org.uk (Rosalind Institute for Open Science)

- Philosophy of Science: Induction, Causation, and Popperian Falsifiability



Philosophy of Science

So that leaves us here. We have:

- (1) statistical tools to do null hypothesis testing.
- (2) philosophy of science that says we need to falsify scientific hypotheses.

How do we combine them?

The greatest obstacle that I encounter among students and colleagues is the tacit belief that the proper objective of statistical inference is to test null hypotheses. This is the proper objective, the thinking goes, because Karl Popper argued that science advances by falsifying hypotheses.

But the above is a kind of folk Popperism, an informal philosophy of science common among scientists but not among philosophers of science. Science is not described by the falsification standard, as Popper recognized and argued. In fact, deductive falsification is impossible in nearly every scientific context.

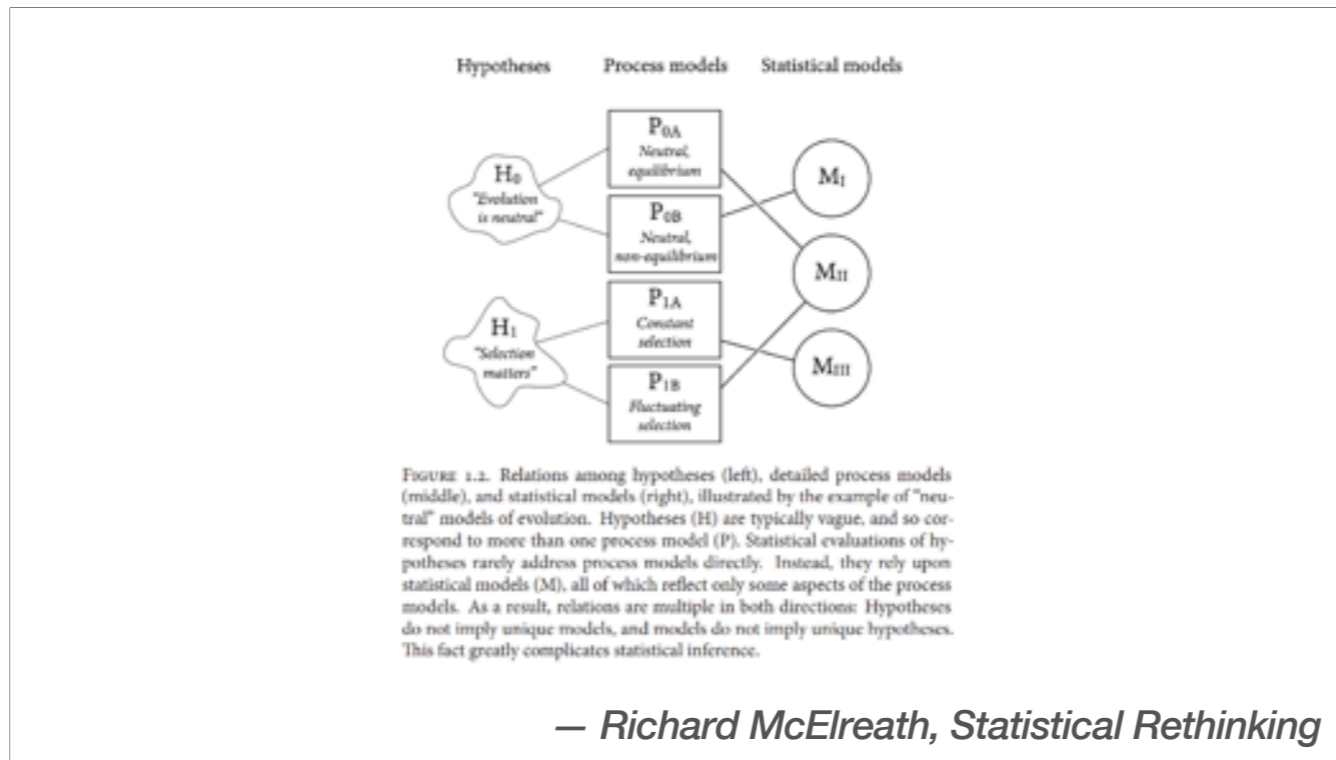
— Richard McElreath, Statistical Rethinking

First address how they are *not related* and how we should *not* combine them.

2 claims in this quote here.

1 is about the relationship between scientific hypotheses and null hypotheses; theory or model of the world, and the statistical tests you're doing.

2 is that falsification only takes us so far and is an imperfect description of how science actually proceeds.



Here is an important, real and wonderful example from statistical rethinking. Real debate from evolutionary science that is now settled. The question is whether evolution happens because of random variation/mutations in our genes (evolution is neutral) or whether it happens because of variation + natural selection a la 'survival of the fittest' (selection matters). We believe this second one now because the dispute is resolved.

Notice there is no null hypothesis here. What we have to do is create what Richard McElreath calls a 'Process Model', what Judea Pearl would call a 'Data Generating Model', what I'm calling a 'Scientific Model', to explicitly turn our fuzzy hypothesis into something we can test.

You can see each hypothesis maps to a few process models.

P0A assumes equilibrium — it's population size is constant— from this model we would expect to see a power-law distribution of alleles (some are rare, some are not). The existence of this power law distribution of alleles is a hypothesis we can test with a statistical tool.

P1A assumes trait pressures are constant (certain traits are always useful), while P1B posits that certain traits will be more advantageous in certain seasons or at certain times. We can see that P1B also would lead to predict a power law distribution of alleles.

Multiple process models might generate similar distributions of alleles. Testing against these predictions using statistical tests might not give us certainty about which model and hypothesis are correct. We might need to query different types of predictions. We might have priors over our process models based on prior experience or heuristics like which ones are more 'law-like'.

Make your Scientific Model Explicit.

The major takeaway from the previous slide is that you should draw an explicit process model (or data generating model (DGM) or scientific model). You should be explicit about it for yourself, and you should communicate it to others also working in your discipline.

Hypotheses Process models Statistical models

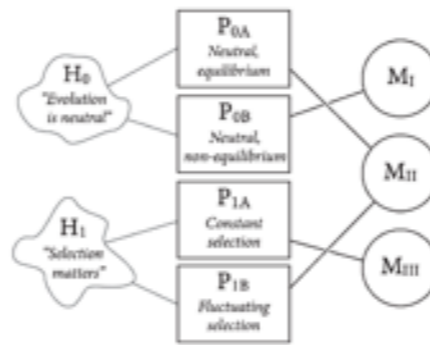


FIGURE 1.2. Relations among hypotheses (left), detailed process models (middle), and statistical models (right), illustrated by the example of "neutral" models of evolution. Hypotheses (H) are typically vague, and so correspond to more than one process model (P). Statistical evaluations of hypotheses rarely address process models directly. Instead, they rely upon statistical models (M), all of which reflect only some aspects of the process models. As a result, relations are multiple in both directions: Hypotheses do not imply unique models, and models do not imply unique hypotheses. This fact greatly complicates statistical inference.

— Richard McElreath, *Statistical Rethinking*

You have to do the middle part!

Popper's Problems #1: Abduction vs. Falsification

Now you may notice that this model of science is not like falsification; falsification is a subset of it. It's really what we'd call 'abduction' or inducing to the best explanation.



Falsification is great. And it works for something like the Large Hadron Collider, where we're looking for the Higgs Boson.

We have a precise claim. An unlikely claim. A Claim that doesn't jive with other hypotheses we would have. We have not every observed data that would corroborate or disprove this prediction before ever.

Falsification is great here. But notice all the features that *make* it great.

$P(\text{data} \mid \text{our hypothesis})$ is high. $P(\text{data} \mid \text{other hypotheses})$ is low. The data is unobserved, so we're not reasoning about it and making sure our hypothesis is consistent with it and selecting our hypothesis after seeing the data. We can collect the data (albeit at great expense).

Theories of why extroverts are happier than introverts:

H1: We all have to engage socially, extroverts don't pay a cost.

H2: extroverts are more rewarded by pure attention and less responsive to social judgement.

H3: extroverts find 'pleasant' moods high-energy and introverts like to relax quietly, high alertness leads to longer experience.

H4: extroverts are more personally adaptive so positive affect lasts longer for them.

H5: extroverts are less responsive to environment so positive affect lasts longer for them.

- (1) Can a **true theory** fail to have unique, un-intuitive predictions or weak explanatory power? I think so. Depends on the domain of the theory.
- (2) What do you do *after* the large hadron experiment, when *all theories* will be consistent with the data?
- (3) look at this example about Extroverts being happier than introverts; it resembles our natural selection example. There are tons of hypotheses. There is no null hypothesis; these are probabilistic, so no single bit of data will 'falsify' them; many of them will make very similar kinds of predictions.

Theories of why extroverts are happier than introverts:

H1: We all have to engage socially, extroverts don't pay a cost.

H2: extroverts are more rewarded by pure attention and less responsive to social judgement.

H3: extroverts find 'pleasant' moods high-energy and introverts like to relax quietly, high alertness leads to longer experience.

H4: extroverts are more personally adaptive so positive affect lasts longer for them.

H5: extroverts are less responsive to environment so positive affect lasts longer for them.



- (1) For a lot of social science the era of 'large hadron collider' was in the 70s and is over— you could put undergrads in cages and feed them LSD and do all sorts of crazy things. Now we have ethics and can no longer engage in these kinds of crazy, exploitative, large hadron collider style definitive experimentation on people (except on Tuesday nights on ABC if you're a TV producer).
- (2) still predicting everyday behavior accurately is next to impossible— plenty of prediction to be done.



To make this point again, a geocentric solar system can *perfectly predict* celestial motion. Circles on circles gives us a Fourier representation that can draw arbitrary paths.

Perfectly explain the same set of observations with Geocentric and Heliocentric; one is a little more complicated. Descriptively accurate, underlying assumptions incorrect.

Popper's Problems #2: Kuhn's Incommensurability

Falsification has one more problem I'd like to touch on; 'incommensurability'.

Swan

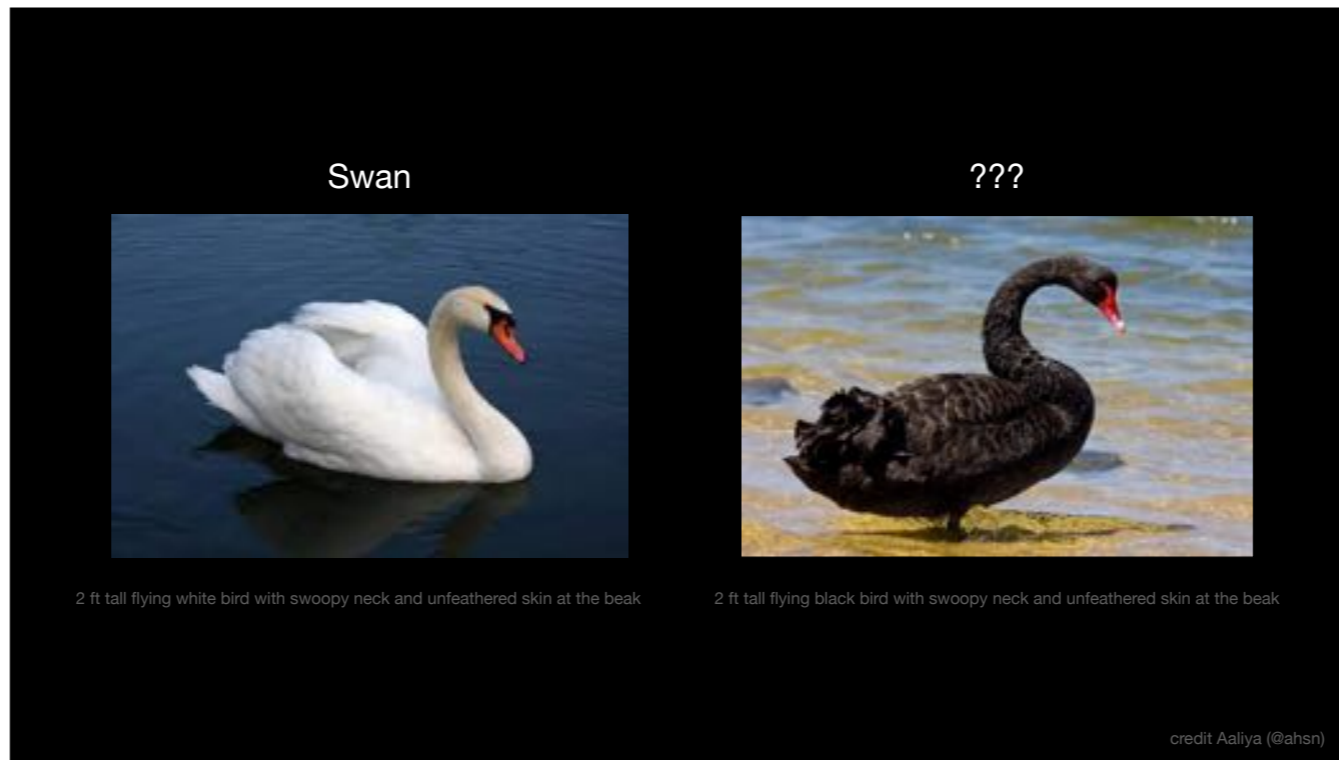


2 ft tall flying white bird with swoopy neck and unfeathered skin at the beak

credit Aaliya (@ahsn)

Based on Aaliya's question last time, which I loved.

Say you're a 15th century European and you define a swan as a 'white flying bird w a swoopy neck and unfeathered beak'.



And you find this bird which has the same neck, beak, traits, etc but it's not white. Well, you didn't falsify your idea of a swan. It's not a swan! Swans are white birds. We just found a new bird that looks a lot like a swan!

HMMM, perhaps we there's something uniting these birds...

Swan



2 ft tall flying white bird with swoopy neck and unfeathered skin at the beak

???



2 ft tall flying black bird with swoopy neck and unfeathered skin at the beak

credit Aaliya (@ahsn)

but seeing this bird might make you think. Hmm. These look like they are the same bird, it's just the color that is different.

HMMM, perhaps we there's something uniting these birds...

Swans = Genus: Cygnus

Species: Cygnus olor



2 ft tall flying white bird with swoopy neck and unfeathered skin at the beak

Species: Cygnus atratus

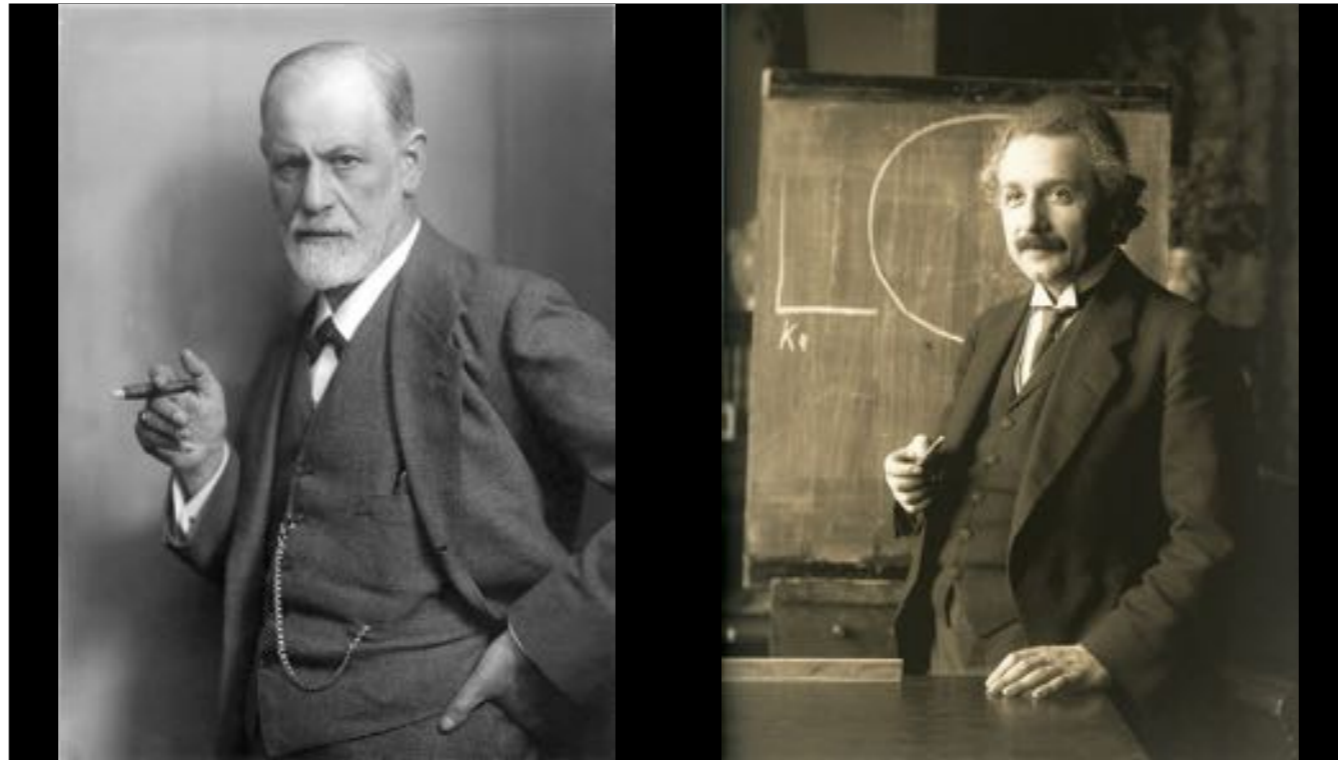


2 ft tall flying black bird with swoopy neck and unfeathered skin at the beak

credit Aaliya (@ahsn)

And so you throw out your old way of classifying birds as white birds and black birds, and you come up with a new taxonomy! You invent 'Genus' and 'Species' and you name the 'Genus' swans!

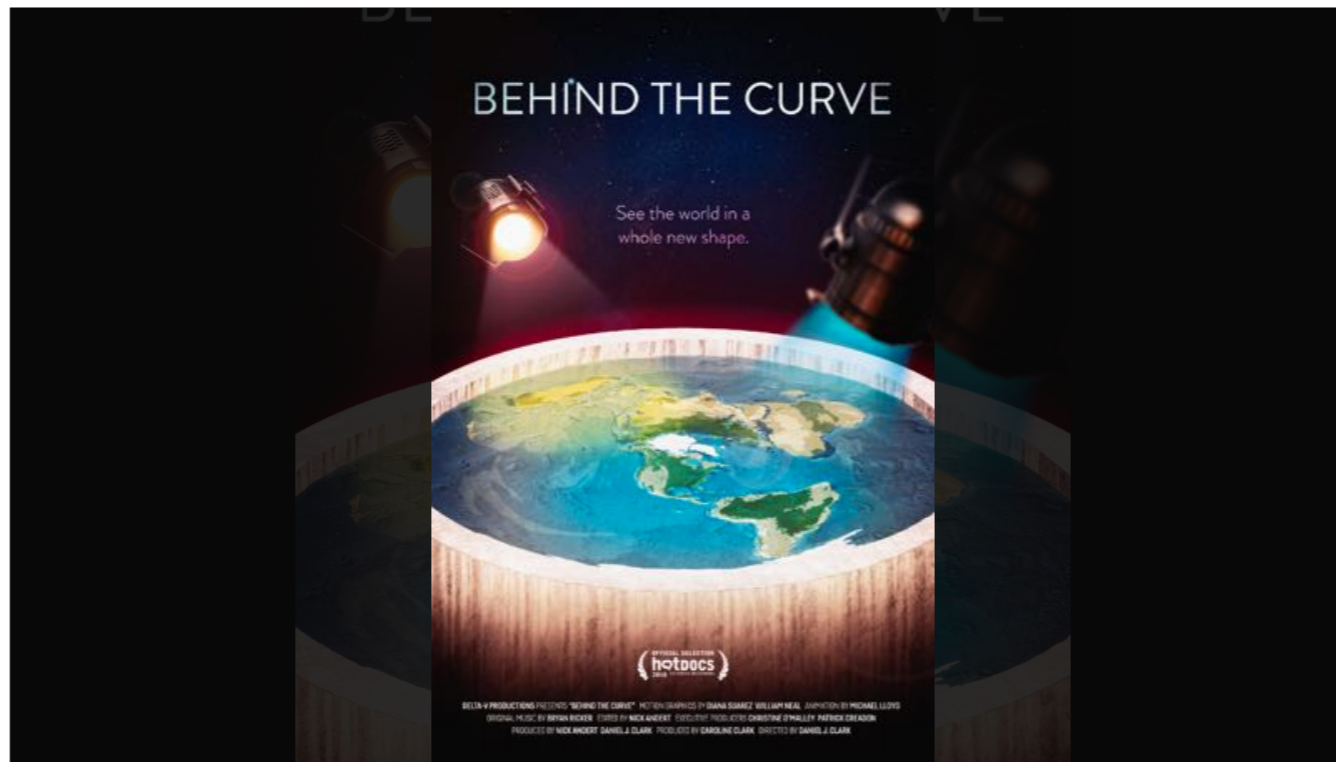
Colloquially we might say we 'falsified swans are white', but we really didn't. We had what Kuhn would call a 'paradigm shift'. There's no way to evaluate competing theories across a paradigm shift; they use different concepts, different language, different assumptions.



Freud did this by popularizing the idea of the subconscious. This is a paradigm shift, despite being not-falsifiable (and 'science' according to Kuhn). Einstein created a paradigm shift too.

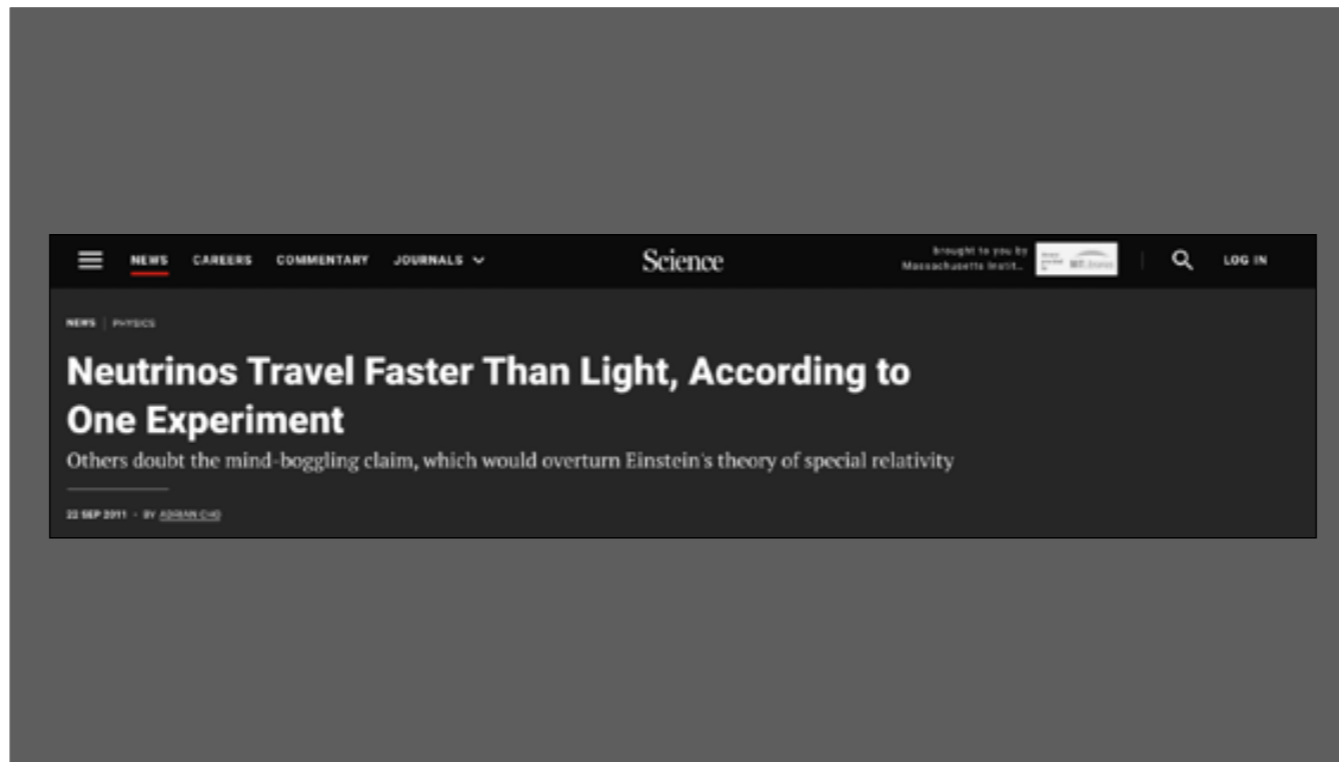
Popper's Problems #3: Paradigms and Falsification

How do we deal with paradigms in science?

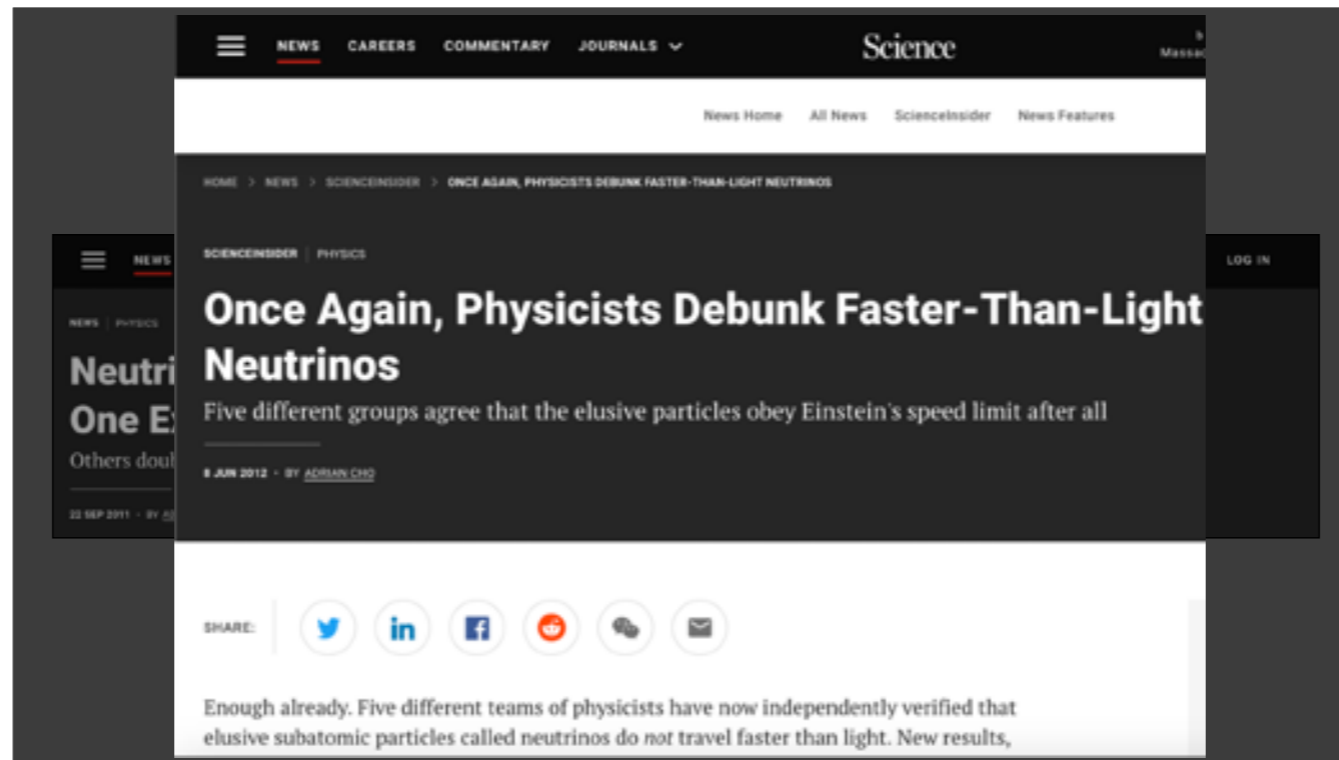


In the flat earth documentary, flat-earthers run a few pretty high quality experiments with gyroscopes and lasers that prove the earth is round. And they scratch their heads and say 'what did we do that is wrong here? We *must* have messed something up because we know the earth is flat.'

We are *exactly* the same. If you went out and proved the earth was flat, how many experiments would it take to convince you? How many experiments would you have to do before you rebuke conventional wisdom?

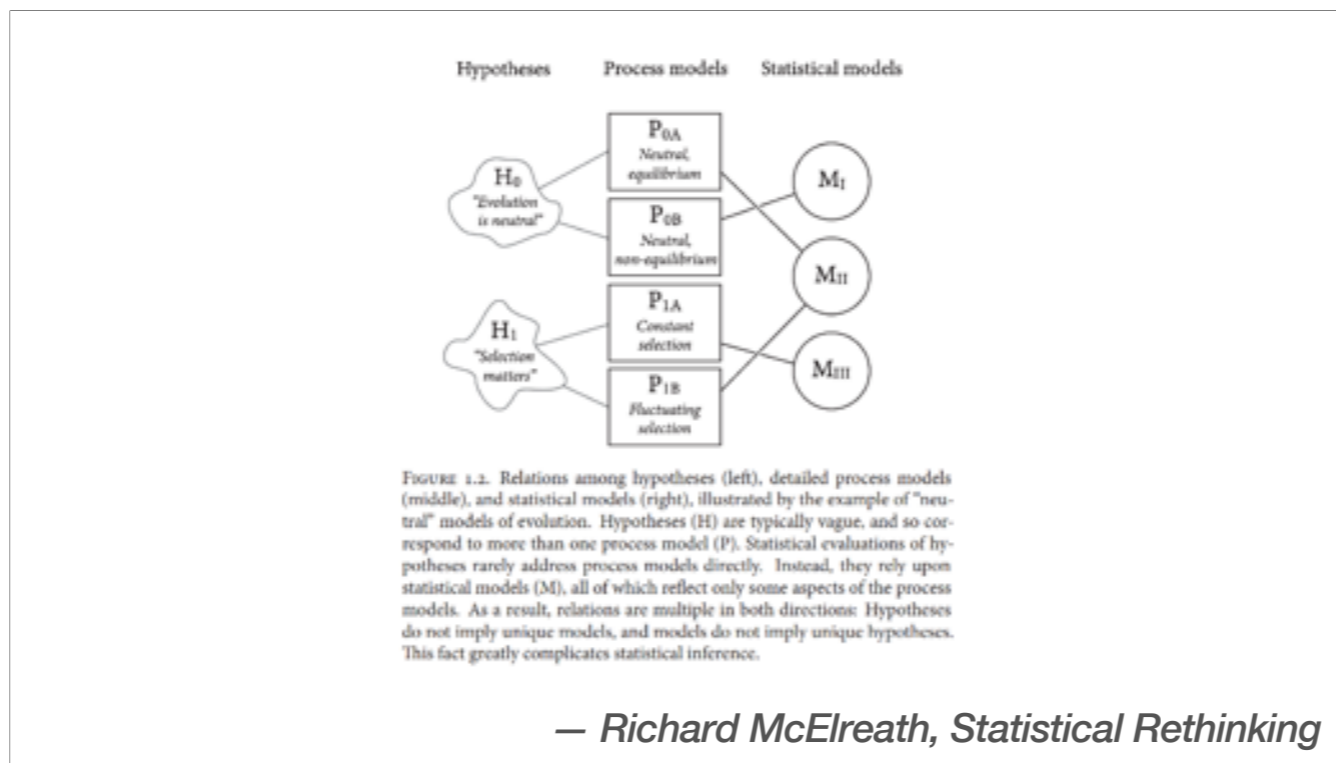


An example from science— faster than the speed of light neutrinos.



Immediately picked apart by physicists around the world. Notice the article says 'Enough already'. This goes against the paradigm! The bar for accepting new findings and trusting scientific methodology scales with how well a claim fits with the overall paradigm.

Bad? No! Bayesian! In my opinion, for an overall system, this is not so bad. Revolutionary findings require revolutionary levels of evidence.



So this is our worldview of how to do science, in my opinion.

- (1) We generalized falsifiability to abduction to the best explanation, after creating explicit scientific models for theories. Our job as scientists is to be explicit about this entire chain.
- (2) We should be very concerned about is the taxonomy underlying our model. Do we have the right concepts and relations? Paradigms and scientific resolutions are based on thinking at this level.
- (3) We should reason like a bayesian; with priors; at all levels of abstraction.

Make your Scientific Model Explicit.

Part I: Why?

Part II: How?

Part III: Where does it breaks down?

Talked about Why. Now let's talk about How.

Hypotheses Process models Statistical models

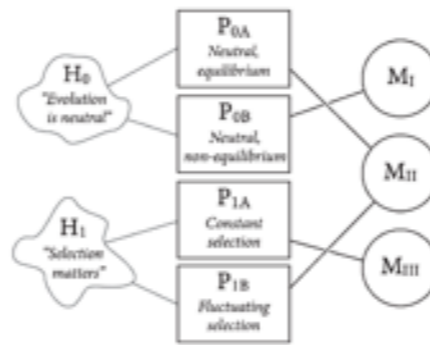
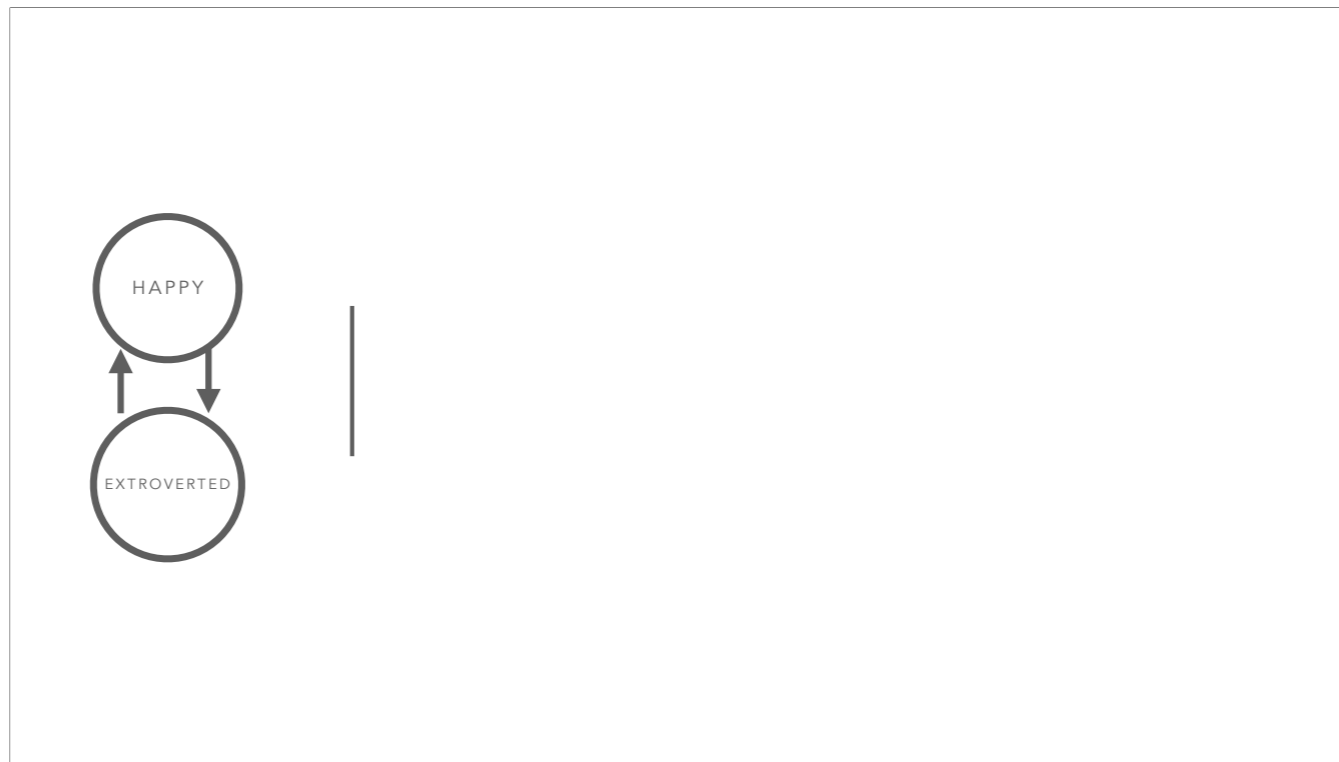


FIGURE 1.2. Relations among hypotheses (left), detailed process models (middle), and statistical models (right), illustrated by the example of "neutral" models of evolution. Hypotheses (H) are typically vague, and so correspond to more than one process model (P). Statistical evaluations of hypotheses rarely address process models directly. Instead, they rely upon statistical models (M), all of which reflect only some aspects of the process models. As a result, relations are multiple in both directions: Hypotheses do not imply unique models, and models do not imply unique hypotheses. This fact greatly complicates statistical inference.

— Richard McElreath, *Statistical Rethinking*

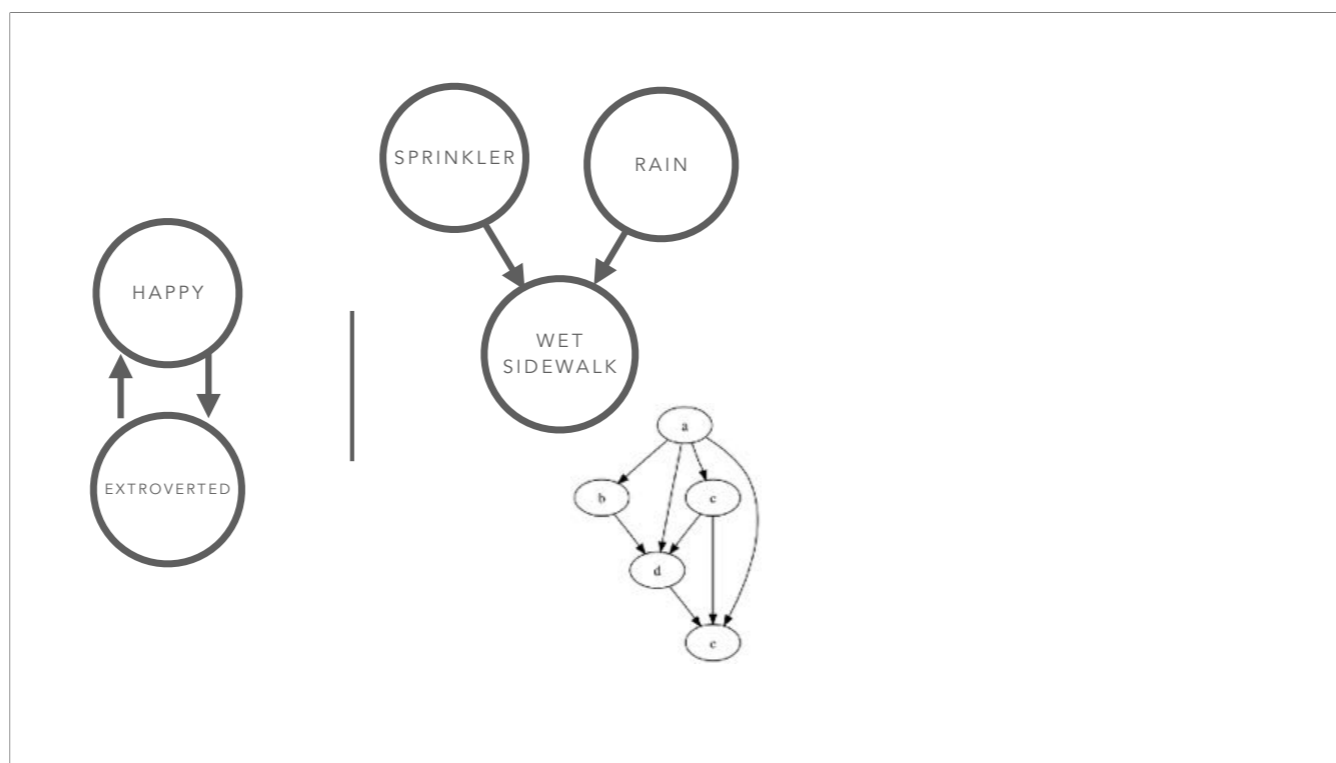
How do we represent this middle column?



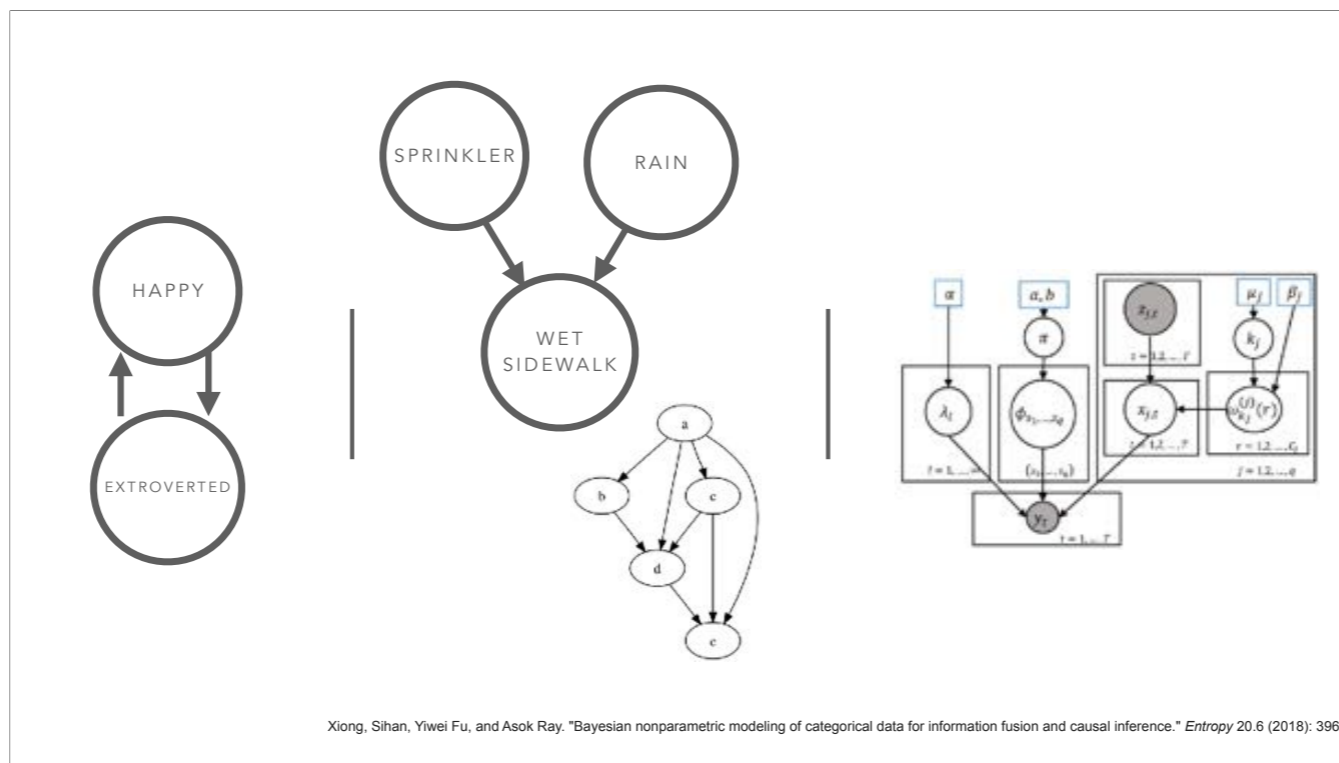
‘Graphical model of causal structure’. A child can do it.

Blobs of concepts, lines to represent causal relationships. Here let’s say we see extroverted behavior interacting with happiness; when I’m happy, I act more like an extrovert, when acting like an extrovert I’m happier. If I want to measure something like happiness, I need to draw another arrow from ‘happy’ (unobservable) to ‘survey answers about happiness’.

Simple but good to remind ourselves what assumptions we’re making and how good or bad they might be. Where do these assumptions break down? What is going into them?



Next layer is a 'DAG' or a directed acyclic graph. If we constrain ourselves not to have cycles in our graph, we can open up an amazing world of computation on our causal model. I'm going to talk about this as we move on, but you should look up Judea Pearl and Josh Tenenbaum to really understand all the possibilities here.



And lastly in this line of complexity is plate notation; parameterized complex probabilistic models.

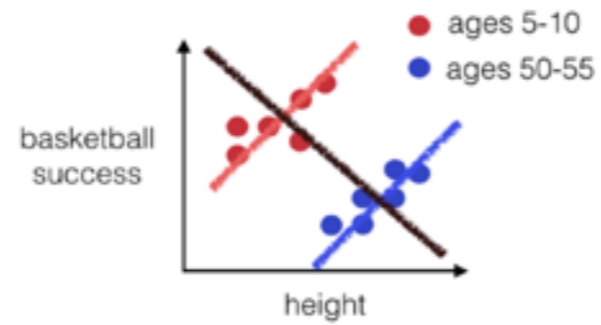
shaded observed RVs, unshaded unobserved (latent) RVs, boxes fixed hyper parameters, plates are copies of what's inside,

Scientific Models are Causal Models.

You might notice that these options are all causal.

YOU CANNOT AVOID CAUSAL MODELS. IT IS IMPOSSIBLE.

Simpson's Paradox



A graphical example of Simpson's paradox. Taller kids and taller adults are both better at basketball, but in general being tall makes you worse!

older people get taller; older people play basketball less and are out of shape and are worse.

trend in subpopulations contradicts overall trend.

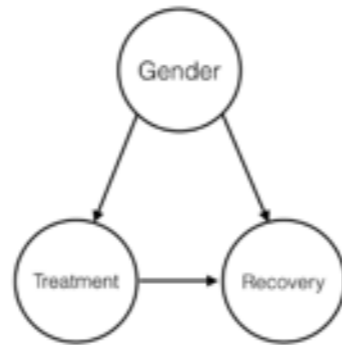
Does being taller make you better at basketball?

	<i>Men</i>	<i>Women</i>	<i>Combined</i>
<i>Treatment</i>	45/50 = 90%	90/150 = 60%	135/200 = 68%
<i>No Treatment</i>	280/350 = 80%	20/50 = 40%	300/400 = 75%

Uh oh. The drug is good for men, good for women, and bad for people.

Now we'll talk about Simpson's Paradox. This is based on a real medical study!

The drug is good for men, good for women, but bad for people. What do you do?



	Men	Women	Combined
Treatment	45/50 = 90%	90/150 = 60%	135/200 = 68%
No Treatment	280/350 = 80%	20/50 = 40%	300/400 = 75%

Uh oh. The drug is good for men, good for women, and bad for people.

A simple causal diagram of what's going on. Gender affects the likelihood of accepting treatment. We also expect the likelihood of recovery to be a function of both receiving treatment AND gender.

We should draw a causal diagram, and from there we can reason that we should control for gender. Gender influences which treatment group they're in. Men are disproportionately in the no treatment and women are in the treatment.

	<i>Men</i>	<i>Women</i>	<i>Combined</i>
<i>Treatment</i>	$45/50 = 90\%$	$90/150 = 60\%$	$(90+60)/2 = 75\%$
<i>No Treatment</i>	$280/350 = 80\%$	$20/50 = 40\%$	$(80+40)/2 = 60\%$

Simpson's paradox is resolved!

In other words, we should just average the percentages of each group (based on how frequently that group appears in our overall population, 50/50 for men and women), without weighting them by size. This removes the effect of uneven treatment selection by the two genders.

We should consider each subgroup equally.

	<i>Low BP</i>	<i>High BP</i>	<i>Combined</i>
<i>Treatment</i>	45/50 = 90%	90/150 = 60%	135/200 = 68%
<i>No Treatment</i>	280/350 = 80%	20/50 = 40%	300/400 = 75%

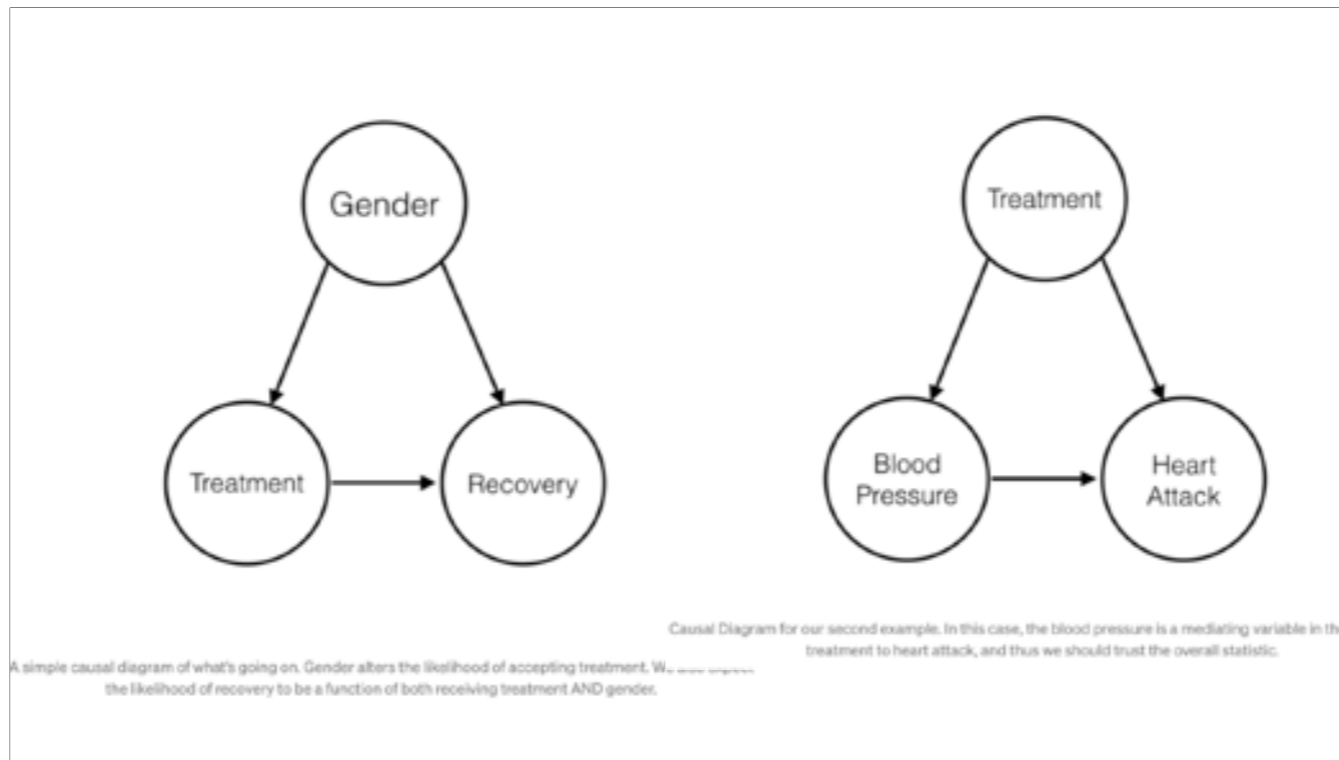
Now we'll do the same example, with the same numbers. But here we give drug for heart attacks, when they come in for a heart attack or a checkup we measure their blood pressure.

notice this!



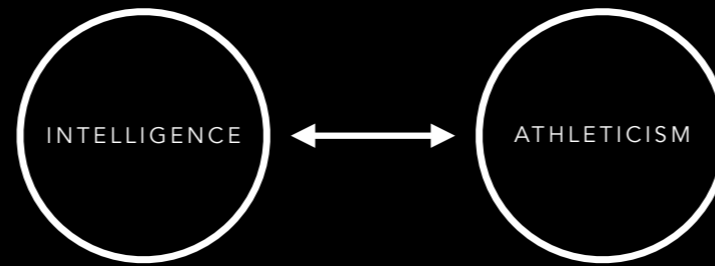
We draw another causal diagram. Here we see that the treatment is working by altering blood pressure. In this case we do not control for blood pressure, we use the combined group.

It's important to point out that this is observational BP data, collected when patients are checking in, and *not* BP data prior to randomization. In the case where BP data was collected prior to randomization, it would not have a causal relationship like this.



In any case, here we have two examples, with the same data exhibiting Simpson's paradox — the trend in each subgroup counteracts the overall trend. We see in one case you should trust the subgroups (because the subgroup affected who got the treatment); and in the other you should trust the combined score (because subgroup affected by treatment).

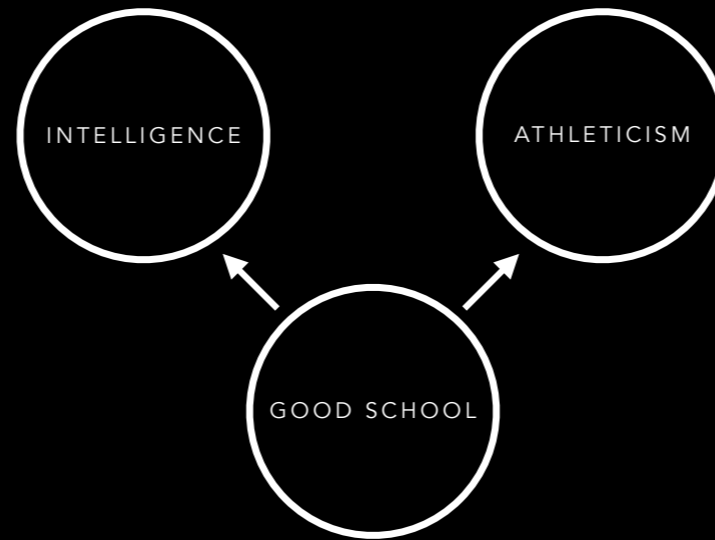
CONDITIONING IMPLIES CAUSAL STRUCTURE



— Judea Pearl

Another example to see if there is a link between intelligence and athleticism in young adults. Let's say some went to a good school.

CONDITIONING IMPLIES CAUSAL STRUCTURE



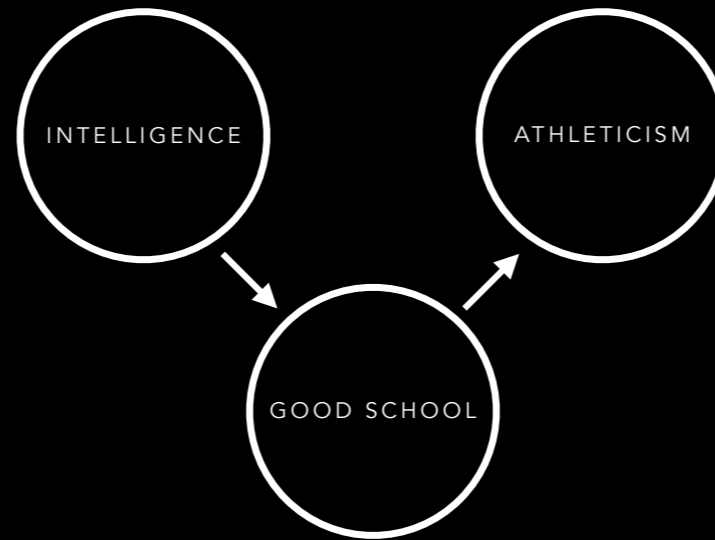
Fork

— Judea Pearl

Perhaps good schools provide opportunity for growth along both of these dimensions.

We should condition here. Like the gender/treatment example, if good school affects both, we want to look at relationship between intelligence and athleticism of subpopulation not at school and subpopulation at school, and treat these as different subgroups.

CONDITIONING IMPLIES CAUSAL STRUCTURE



Chain

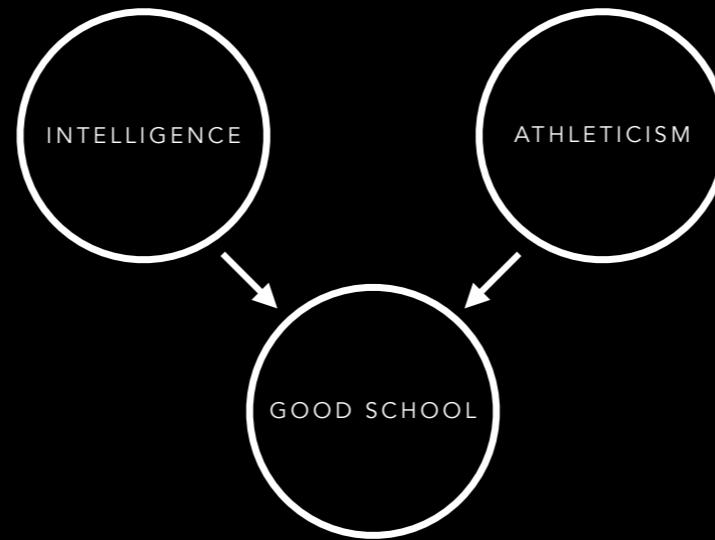
— Judea Pearl

Perhaps though intelligent people get into good schools, and that in turn gives them access to high quality athletic programs, time, and resources. This is a chain.

We shouldn't condition here; intelligence *does lead to athleticism*, and conditioning *breaks the relationship*. There is a correlation between them, and now it's gone when we condition on a good school.

Type II Error. False Negative. Bad. But it gets worse!

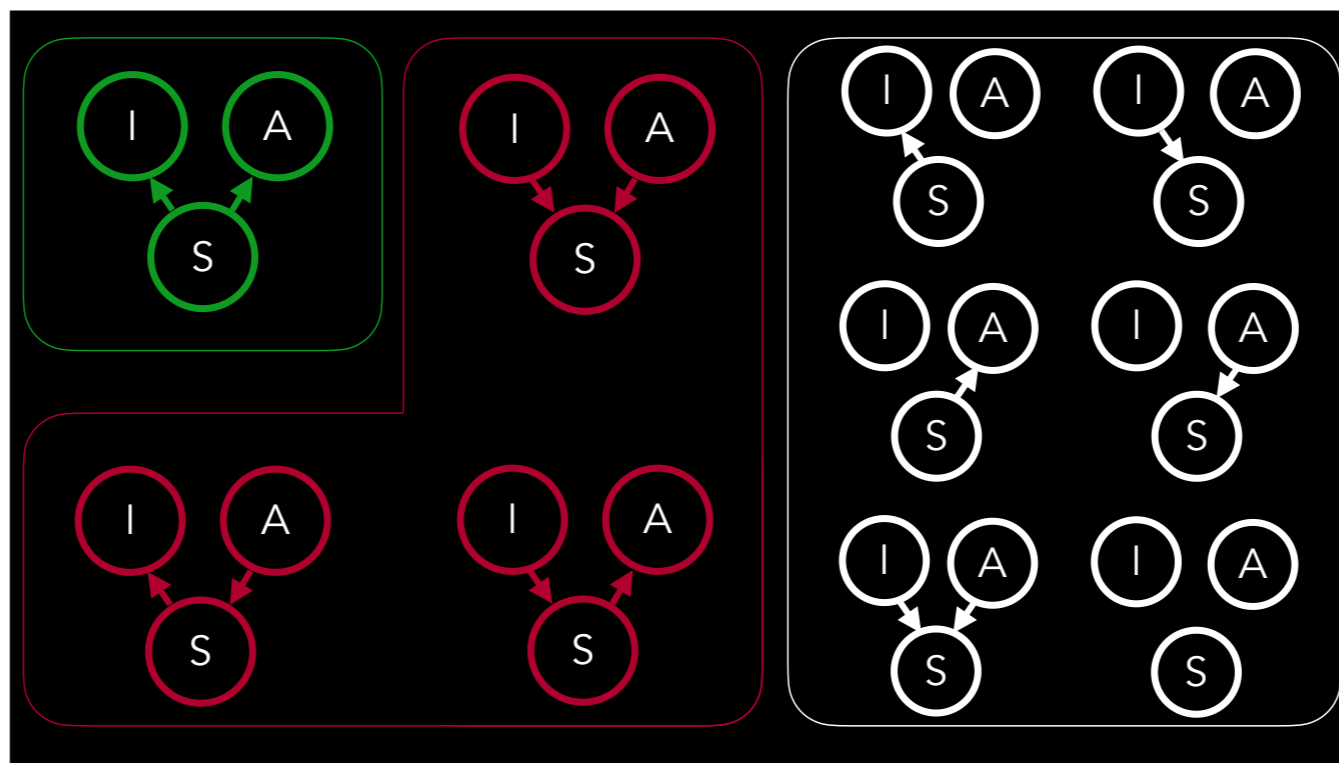
CONDITIONING IMPLIES CAUSAL STRUCTURE



Collider

— *Judea Pearl*

Here's the canonical issue. Let's assume intelligence and athleticism are totally unrelated, BUT you get into college for being either freakishly smart or freakishly athletic. If you condition on the school, you create a negative correlation! People at the school will be really smart when not athletic, and really athletic when not smart. You have introduced this relationship when there wasn't one by conditioning!



If you condition on variables, or decide *not to* condition on variables, you are making assumptions about the underlying causal structure whether you are explicit about it and realize it or not. *Don't do that.* You should be very explicit about what you believe your underlying structure to be.

If we were to condition on schools here, we would get a correct answer in green, a utterly wrong answer in red, and would be doing something 'harmless' but unjustified in white.

How you apply your statistical analysis is *deeply dependent* on this structure, and justified by querying it correctly.



Scientific Models are Causal Models.

Hopefully I've convinced you that we have to make causal assumptions. Even for descriptive science we have to causal sampling assumptions, representativeness assumptions, and SUTVA assumptions even for RCTs.

What is the mechanism for the distribution of the trait you're interested in in the sampling population? What is the sampling process, and does it interact with that trait? How does your intervention interact and what kinds of statistical properties do you expect it to produce in the dependent variable? Draw them out and be explicit about your (causal) assumptions.

but... PROBLEM OF CAUSALITY!

ok so if we have to have a causal model, what do we do about causality?

Make your assumptions *explicit*. This is why we need to know your process model. In some deep sense, causal models are unjustifiable.

But we're forced to compare and reason about causal models.

Correlation does not imply Causation.

Now I'm going to complain about this phrase.

IMPLY here the formal logic word == IS SUFFICIENT CONDITION.

IMPLY the real world == suggests, and of course correlation is suggestive of causation.

Hypotheses Process models Statistical models

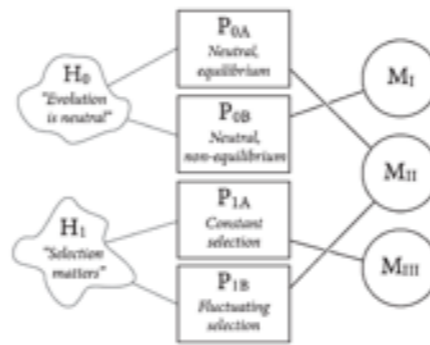


FIGURE 1.2. Relations among hypotheses (left), detailed process models (middle), and statistical models (right), illustrated by the example of "neutral" models of evolution. Hypotheses (H) are typically vague, and so correspond to more than one process model (P). Statistical evaluations of hypotheses rarely address process models directly. Instead, they rely upon statistical models (M), all of which reflect only some aspects of the process models. As a result, relations are multiple in both directions: Hypotheses do not imply unique models, and models do not imply unique hypotheses. This fact greatly complicates statistical inference.

— Richard McElreath, *Statistical Rethinking*

In fact, someone like me might argue that the point of science is to use correlation to predict causal relationships.

correlation w/o causation



But of course we all know it's possible to get this. And in some fundamental sense the choice is unjustified, which is why it's so important to be explicit about the scientific model you're choosing.

intelligence as **pattern-matching** (**correlation**) vs **model-building** (**causation**)

Human-Level Intelligence or Animal-Like Abilities?

By Adrian Dorneliche
Communications of the ACM, October 2018, Vol. 61 No. 10, Pages 56-67
10.1145/3271425
[Comments](#)

VIEW AS:      SHARE:      



"The vision systems of the eagle and the snake outperform everything that we can make in the laboratory, but snakes and eagles cannot build an eyeglass or a telescope or a microscope."
Judea Pearl¹

The recent successes of neural networks in applications like speech recognition, vision, and autonomous navigation has led to great excitement by members of the artificial intelligence (AI) community, as well as by the general public. Over a relatively short time, by the science clock, we managed to automate some tasks that have defied us for decades, using one of the more classical techniques due to AI research.

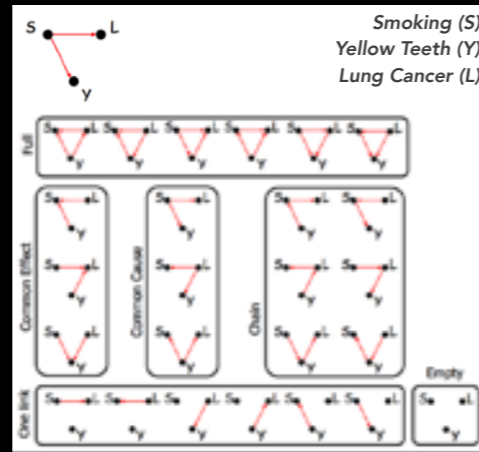
[Back to Top](#)

But humans are are really good at it! Causes seems so obvious!

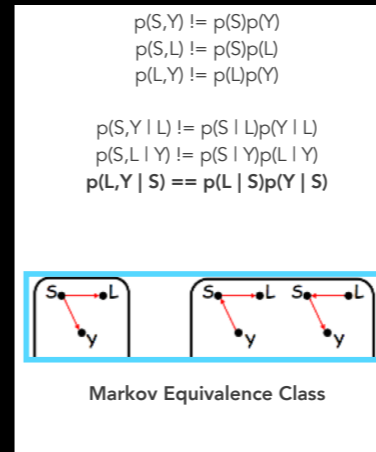
Perhaps if/when we (really) build artificial intelligence it will be able to reason causally. And people are working on this, and there is reason to hope.

CORRELATION *DOES* 'IMPLY' CAUSATION (SOMETIMES)

STRUCTURE LEARNING



from MIT OCW, Tenenbaum Computational Cog Sci Class



from MIT OCW, Tenenbaum Computational Cog Sci Class

Can do structure learning — learn causal structure from correlation.

The first three eliminate the bottom row and the common effect column

The second three eliminate the rest EXCEPT for the markov equivalence class

Markov equivalence class: A set of causal graphs that cannot be distinguished based on (in)dependence relations.

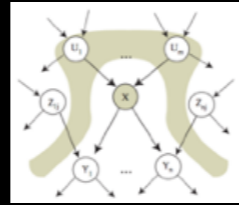
The way to overcome this is through *intervention* or through *prior knowledge* (temporal order or domain knowledge).

This is *constraint based* structure learning. also dynamic greedy search over graph structures, and 'score-based' algorithms that maximize $\text{prob}(\text{Graph Structure} | \text{Data})$ — Chow Liu algorithm (limited to tree structures with minimum one parent for each node)

CORRELATION *DOES* 'IMPLY' CAUSATION (SOMETIMES)

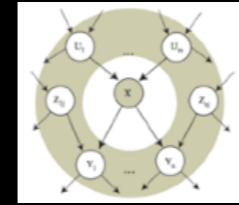
STRUCTURE LEARNING

This generalizes



from MIT OCW, Tenenbaum Computational Cog Sci Class

Markov Property



from MIT OCW, Tenenbaum Computational Cog Sci Class

Markov Blanket

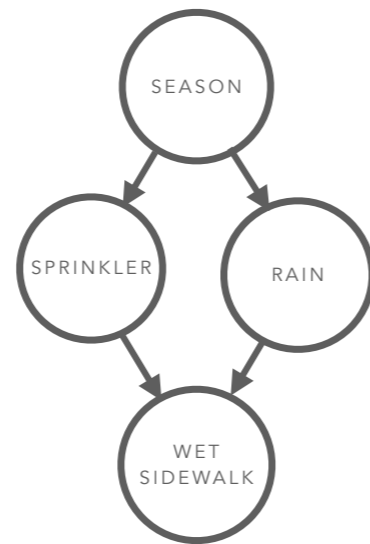
Can factorize global graph to assess a subgraph. We can do science on even very complicated graphs by factoring them!

“Markov property”: Each node is conditionally independent of its **non-descendants** given **its parents**.

“Markov blanket”: Each node is conditionally independent of **all other nodes** given **its parents, children, children’s parents**.

This is called constraint based structure learning, but there are score based $P(S|Data)$ typically computationally intensive, and greedy search methods as well.

DO sprinkler = on vs. SEE sprinkler = on



$p(\text{Season} \mid \text{Sprinkler} = 1)$

vs

$p(\text{Season} \mid \text{DO}(\text{Sprinkler} = 1))$

— Judea Pearl

This kind of explicit causal reasoning has not been in favor.

Just to drive the point home that our tools are impoverished; stuck in a frequentist world where explicit causal reasoning isn't possible, even though it's easy and obvious to us.

Probabilistic Programming Languages

Model building — tradeoff between data and inductive bias on predictive power.

PPLs — all languages we use are turing complete. PPLs privilege probability distributions and inference.

Generative Models.

Start with strong inferences; a model that simulates the process and generates lots of plausible data; then infer the parameters that fit.

PPLs are awesome and you should check them out.

Probabilistic Programming Languages

Pyro.

Gen.

STAN.

Edward.

Pymc3.

Here are some to look for.

Make your Scientific Model Explicit.

Part I: Why?

Part II: How?

Part III: Where does it breaks down?

Why most psychological research findings are not even wrong

Anne M. Scheel

Human-Technology Interaction Group,
Eindhoven University of Technology
Eindhoven, The Netherlands

ORCID: <https://orcid.org/0000-0002-6627-0746>

Correspondence: anne-scheel@tue.nl

Psychology's replication crisis is typically conceptualised as the insight that the published literature contains a worrying amount of unreplicable, false-positive findings. At the same time, meta-scientific attempts to assess the crisis in more detail have reported substantial difficulties to identify unambiguous definitions of the scientific claims in published articles and to determine how they are connected to the presented evidence. I argue that most claims in the literature are so critically underspecified that attempts to empirically evaluate them are doomed to failure—they are *not even wrong*. Meta-scientists should beware of the flawed assumption that the psychological literature is a collection of well-defined claims. To move beyond the crisis, psychologists must reconsider and rebuild the conceptual basis of their hypotheses before trying to test them.

no process models in psychology, where they are needed the most.

Ontological Entities and their Relations.

step through how we get this wrong.

What Entities?

**Human Behavior, Beliefs, and
Mental Experience** are shaped by
internal Personal factors like **Personality/
Traits and Mood/Disposition**
as well as **external Environmental/
Situational factors.**

The subjective, latent, personal, unobservable quantities matter a lot.

9 Attention and Control: Have We Been Asking the Wrong Questions? A Critical Review of Twenty-Five Years

Alan Allport

9.1 A CRITICAL LOOK BACK

The twenty-five years since the first Attention and Performance symposium (Sutton 1981) have been a period of extraordinarily rapid growth and transformation within the cognitive sciences. Indeed, if some Big van Winkle of experimental psychology were to return to the land of AM after a quarter century of sleep sleep, he could scarcely fail to be astonished at the extent to which the scientific landscape had expanded compared to what he had known in the early 1980s. Transformation has occurred on almost all fronts in cognitive psychology and neuropsychology in the neurosciences, in cognitive and artificial intelligence, and in connectionist modeling. It would be struck first, by the growing recognition of conceptual as well as empirical interdependence among these very diverse disciplines in the movement toward an integrated cognitive neuroscience. In short, over this twenty-five-year period, the surrounding scientific landscape, the alternatives being withheld, and the conceptual and empirical frameworks underlying them would appear changed almost beyond recognition.

Nonetheless, in contrast to so much change, if one rereads Big van Winkle were to turn to the psychology of attention, he might be surprised to find certain features of the contemporary landscape remarkably familiar. To a remarkable extent, the same issues and controversies that dominated the psychology of attention twenty-five years ago have continued to preoccupy students of attention throughout this period. And, underlying these issues and controversies, certain enduring assumptions about the nature of attention and about the functional architecture of cognitive and perceptual-motor processes continue to exist from field, even a quarter century later. Two (interrelated) controversies in particular are still widely perceived as central, and the most fundamental, issues to be resolved. These are (i) the controversy over "early" versus "late" perceptual selection, where the question at issue is whether attentional selection occurs before or after the encoding of abstracted stimulus identity; and (ii) the question of which cognitive processes "require attention" and are therefore "attended", and which processes can be performed "without attention" (or "automatically"). Both issues have had extensive

No one knows what attention is

Bertrand Russell¹, Craig S. Chapman², Paul Cook³, Heather T. Reynolds⁴, Jonathon King⁵, Timothy N. Wills⁶

© The Authors 2011

Abstract
 In this article, we challenge the definition of "attention" as a unitary construct with a neural basis. We point out that the concept has no clear meaning in both experiments and the "literature" in which it is used to describe the use of attention in neural implementation and to explain the use of processes during the experiment. To illustrate these points, we focus on discussion on visual selective attention. It is argued that attention is processing, but processed through evolution as a single feature of a complex multi-channel sensorimotor system, which generates selective phenomena of "attention" as one of many by-products. Instead of the traditional analytic approach to attention, we suggest a synthetic approach that starts with well-established mechanisms that do not need to be defined as attention, and are instead for the selective phenomena under investigation. We conclude that what would serve scientific progress best would be to stop the term "attention" as a label for a specific, localized neural system and instead focus on hierarchical sensorimotor processes and the many systems that implement them.

Keywords: Attention • Motor control • Selection • Sensorimotor • Decision making • Phenomena • Attention • Evolution • Perceptual selection • Cognitive architecture

Introduction
 "Everyone knows what attention is" (Sutton, 1980) is one of the most popular quotes from William James and conveys the most basic consensus about human attention.¹ We argue, however, that the meaning and popularity of this statement in cognitive research has been diminished in progress – the is but, as one leaves what attention is. More specifically, we argue that the concept of "attention" is one of the most misleading and abused terms in the cognitive sciences. In the present paper, we state the position that the term "attention" should be abandoned and the nature of the research in this area be reconceptualized to focus on the subjects of processes and mechanisms that lead to task-specific performance. Further positions have been proposed and discussed previously (see Anderson, 2011; De Gaille, 2010; Fehrer & Collins, 2011; Kluwe, Bollenkamp, Isenack, & Wang, 2011; Miall, 2011). The present paper reflects and expands discussion by stating particular and new emphasis on the interconnected and integrative nature of the human sensorimotor information processing system. This emphasis on integrative sensorimotor information processing distinguishes from the traditional approach to understanding "attention" (Anderson & Collins, 2011) and a phenomenological refinement of the current approach to understanding behavior (Cook, 2011) (see below).

¹ Timothy N. Wills is now at the University of Queensland, Australia.
² Institute of Psychology, University of Cambridge, United Kingdom.
³ Faculty of Knowledge, Space, Architecture, University of Ottawa, Ottawa, Canada.
⁴ Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada.
⁵ School of Health and Behavior, Performance, Deakin University, Sheffield, Great Britain, United Kingdom.
⁶ Department of Psychology, Linguistics and Psychological Sciences, Brunel University, Uxbridge, UK.
⁷ Centre for Human Factors, Faculty of Knowledge and Applied Sciences, University of Derby, Derby, UK.

See this kind of critique for most psychological concepts.

The idea of a unitary notion of attention is flawed in the first place. attention is too loose (attention is our bottleneck, or it's our way of dealing with having a bottleneck, which is it)? and analytical approach isn't useful for it, need a synthetic approach.

"there can be no such thing as attention. There is no one uniform computational function, or mental operation (in general, no one causal mechanism), to which all so-called attentional phenomena can be attributed. On the contrary, there is a rich diversity of neuropsychological control mechanisms of many different kinds (and no doubt many yet to be discovered), from whose cooperative and competitive interactions emerge the behavioral manifestations of attention."

Why the Most Important Idea in Behavioral Decision-Making Is a Fallacy

The popular idea that avoiding losses is a bigger motivator than achieving gains is not supported by the evidence

By David Gal on July 31, 2018

“To be sure it is true that big financial losses can be more impactful than big financial gains, but this is not a cognitive bias that requires a loss aversion explanation, but perfectly rational behavior....Moreover, belief in loss aversion has meant that phenomena that have nothing to do with loss aversion have nonetheless been interpreted to reflect loss aversion. For example, the *sunk cost effect*, the finding that people are more likely to continue an endeavor once an investment in it has been made, has been attributed to loss aversion.”

Loss aversion dispute— about what entities to use. Is it ‘loss aversion’ or a collection of other more fundamental concepts like ‘sunk cost effect’?

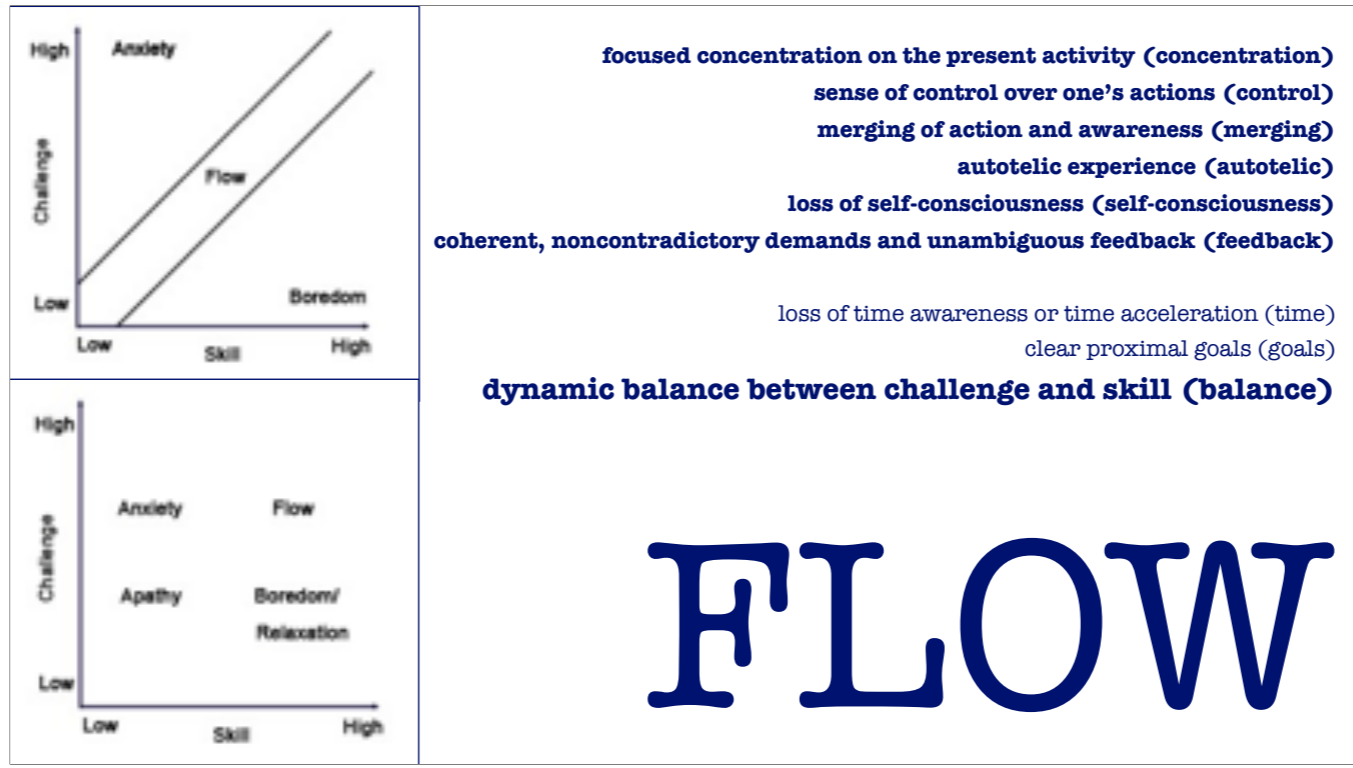
[Kahneman] agreed with Gal and others, he said, that evidence of loss aversion only appears in certain situations. "It's not a law of human nature that you have to find it in every context." "There are experiments where people don't find loss aversion," he added later. "And, again, there's an explanation for every one of them. That doesn't violate loss aversion, because there are exceptions to loss aversion."

What, then, could ever disprove the idea? Is loss aversion falsifiable? Probably not, said Kahneman. "I don't know, maybe it's falsifiable, it's hard to imagine. There are alternative explanations for just about any experiment," he said. But, he suggested, that didn't matter: In the science of decision making, the theory had established its place.

"Having a principle that helps understand a wide body of phenomena — that's considered useful," Kahneman said. "That doesn't mean that loss aversion's true. It means that it's useful."

<https://undark.org/2021/11/15/among-social-scientists-a-vigorous-debate-over-loss-aversion/>

Kahneman's response to the critique.



Issues with Flow definition; 'skill' discounts reading, movies, casinos, social media, conversation, music listening, etc. Not ideal. Other issues as well.

Similar issue with Dark Patterns vs Nudges; moral appraisal conflated with cognitive state.

Attention
Arousal
Mental Effort
Engagement
Task Engagement
Working Memory
Cognitive Load

How to separate it from other concepts?



why does the definition matter? Really easy to get confused and study the wrong thing.

'cognitive load' — mentally taxed and working really hard to manage a complex situation, like driving under challenging conditions or playing a hard video game. The ideas are obviously related, but cognitive load is a different thing— you can have deep effortless attention doing these things, but you can also lead to lability.

Typically easy to induce a state of high stress and attention, cognitive load and flow. Hard to introduce a state of low stress and unconscious deep attention in the lab (like reading or writing or playing music). Weak evidence for a low stress, highly alert physiological state when browsing social media.

hard to isolate flow from cognitive load; few try.

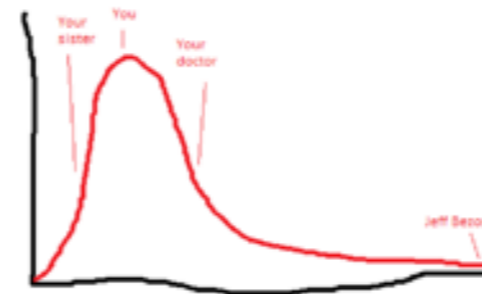
Ontology Of Psychiatric Conditions: Taxometrics

Is mental illness a thing? What kind of thing is it?

Jan 27, 2021 232 251



Wealth as a taxon (not realistic)



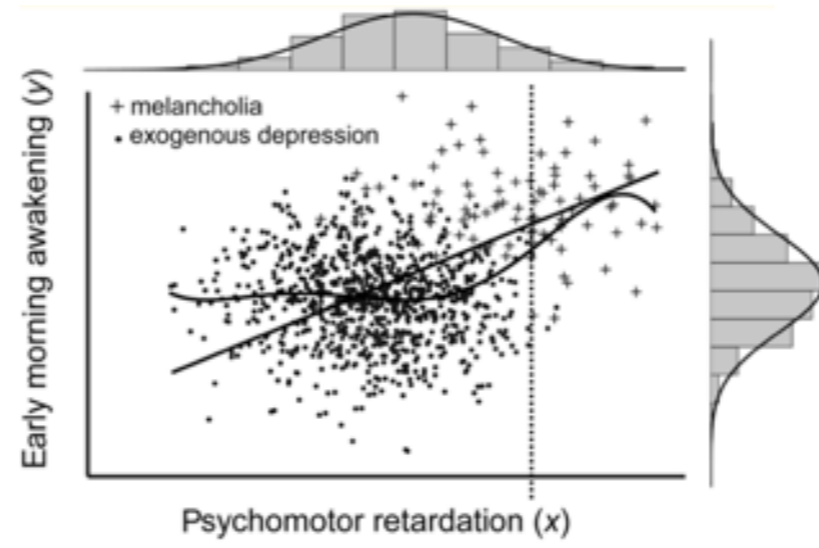
Wealth as a long distribution with some extreme values (realistic)

Haslam, Nick, et al. "Dimensions over categories: A meta-analysis of taxometric research." *Psychological Medicine* 50.9 (2020): 1418-1432.

<https://astralcodexten.substack.com/p/ontology-of-psychiatric-conditions>

Are we categorizing truly different things, or not? Science of Taxometrics.
Depression, ADD vs Schizophrenia. Categorical 'taxons' or tails of distributions?

MAXSLOPE, MAXCOV, MAMBAC



Beauchaine, Theodore P. "Methodological article: A brief taxometrics primer." *Journal of Clinical Child and Adolescent Psychology* 36.4 (2007): 654-676.

Example of computational tool to try and do this rigorously; check out this paper.

**it's hard to define
concepts; read and select
concepts critically.**

How To Measure these Entities?

Reliability and Validity

reliability: precision (how consistent) an instrument is

validity: accuracy (properly represent what you care about)

Reliability

reliability: precision (how consistent) an instrument is

validity: accuracy (properly represent what you care about)

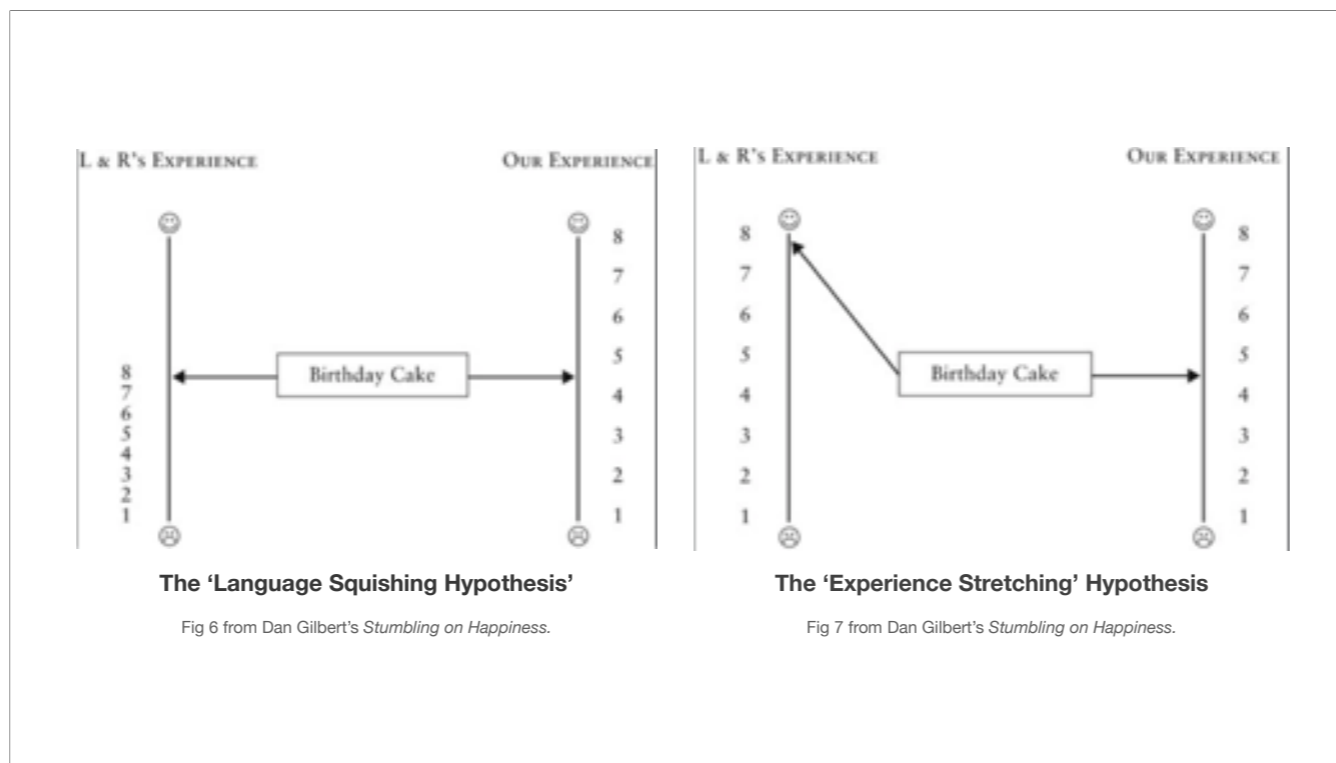
**Inter-Rater Reliability (self-report/informant,
informant/informant agreement)**

Test-Retest Reliability

**Internal Consistency (Cronbach's alpha, split
half reliability)**

Validity

especially difficult, focus on unobservables in psychology. The state of the art is self-report. BUT....



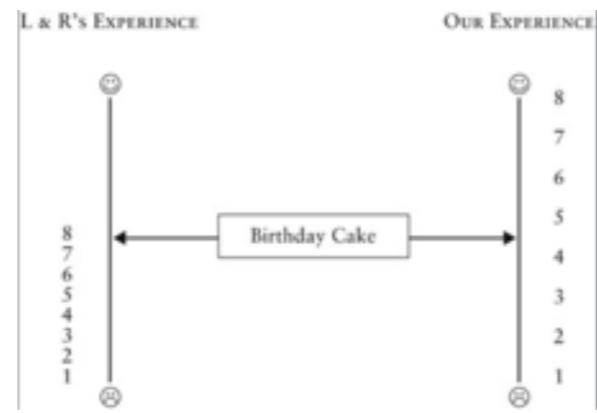
Dan Gilbert talks about Reba and Lori Schappell are conjoined twins attached at the forehead, when asked about surgical separation, they say “Our point of view is no, straight out no. Why would you want to do that? For all the money in China, why? You’d be ruining two lives in the process.”

Fig 1 we have the same subjective experience (more or less) of the cake; just use different words. *“They label their happiest experience with the happiest word in the eight-word language, naturally, but this should not cause us to overlook the fact that the experience they call eight is an experience that we might call four and a half. In short, they don’t mean happy the way we mean happy.”* Impoverished experiential background means they don’t understand the full range of experiences.

Fig 2 an impoverished experiential background shapes our experience in a different way; their experience of an 8 is the same as our experience of an 8. Joy of M&Ms.

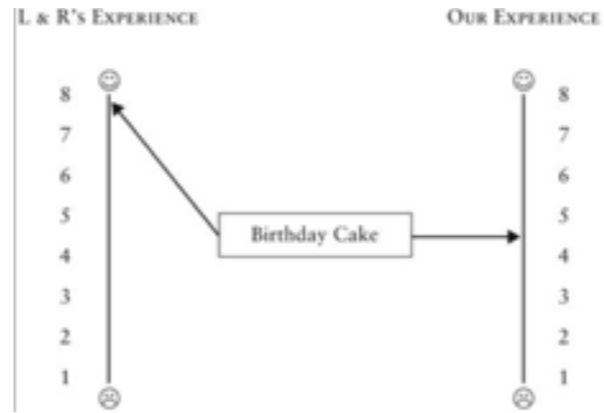


Day 86 of south pole expedition; He forgot that he had buried some cheese puffs and other candy for himself for on the way back. Real and deep joy for cheese puffs.



The 'Language Squishing Hypothesis'

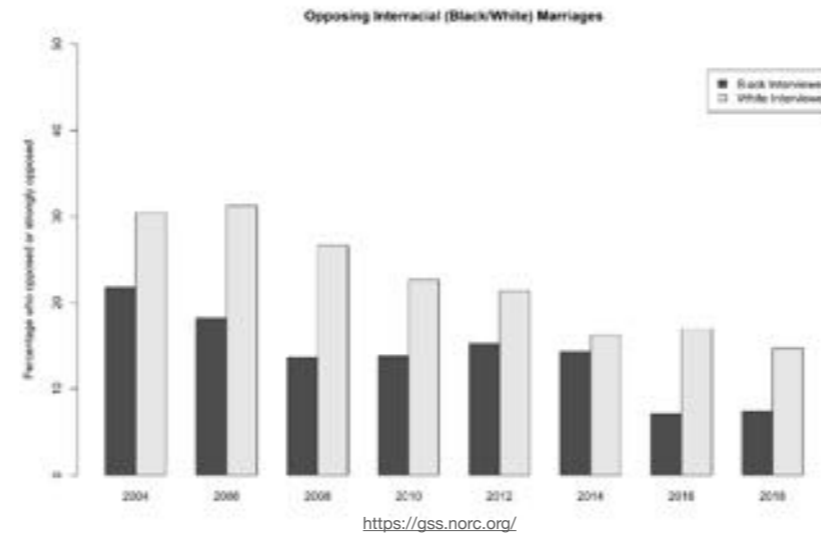
Fig 6 from Dan Gilbert's *Stumbling on Happiness*.



The 'Experience Stretching' Hypothesis

Fig 7 from Dan Gilbert's *Stumbling on Happiness*.

Social Desirability Bias



clearly happens.

Social Desirability Bias

Journal of Consulting Psychology
1960, Vol. 28, No. 4, 347-354

A NEW SCALE OF SOCIAL DESIRABILITY INDEPENDENT OF PSYCHOPATHOLOGY

DOUGLAS P. CROWNE

Ohio State University

AND DAVID MARLOWE

College of Medicine, University of Kentucky

- 21. T F I am always courteous, even to people who are disagreeable.
- 30. T F I am sometimes irritated by people who ask favors of me.
- 33. T F I have never deliberately said something that hurt someone's feelings.

1960 paper.

"Marlowe and Crowne added the word "always" to the desirable question and the word "at least one occasion" to the undesirable item. The idea is that even the most moral person would sometimes fail to do the right thing. As a result, if they answer "True" to the desirable item and "False" to the undesirable item, they are lying. The more respondents answer this way, the less we can trust their responses to be truthful.

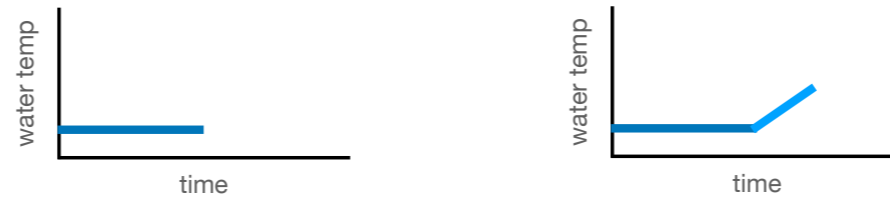
Only four years later, Crowne and Marlowe (1964) wrote a book in which they questioned the usefulness of their measure to detect socially desirable responding. Rather than measuring a response style to questionnaire items, the measure seemed to identify individuals who were actually striving to do the right thing most of the time. This means, it measured real personality traits rather than a tendency to provide deceptive answers on a personality questionnaire. As a result, numerous articles concluded that socially desirability scales are invalid; that is, they cannot be used to detect deliberate faking of responses in personality questionnaires (McCrae & Costa, 1983; Ones, Viswesvaran, & Reiss, 1996).

People don't interpret the questions *literally*.

The failure of creating valid measures of socially desirable responding creates a problem for personality psychology. We still do not know how much self-ratings are distorted because respondents are unwilling to report the truth."

-Ulrich Schimmack's Textbook

Memory - Peak-End Rule (and Recall in General)



WHEN MORE PAIN IS PREFERRED TO LESS:
Adding a Better End

Daniel Kahneman,¹ Barbara L. Fredrickson,² Charles A. Schreiber,¹ and
Donald A. Redelmeier³

¹University of California, ²Duke University, and ³University of Toronto

...so use *the* Experience Sampling Method

ordinal vs cardinal

order vs how much. Are these linear quantities? How the heck would you even know?

Likert Scales

 **medical education**

[Full Access](#)

Resolving the 50-year debate around using and misusing Likert scales

James Carifio, Rocco Perla

First published: 19 November 2008 | <https://doi.org/10.1111/j.1365-2923.2008.03172.x> | Citations: 382

Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes

¹James Carifio and ²Rocco J. Perla
¹University of Massachusetts Lowell, One University Ave, Lowell, MA 01854
²HealthAlliance Hospital, 60 Hospital Road, Leominster, MA 01453

Can Likert Scales be Treated as Interval Scales?—A Simulation Study

Huiping Wu^a and Shing-On Leung^b
^aCollege of Mathematics and Computer Science, Fujian Normal University, Fujian, China; ^bFaculty of Education, University of Macau, Macau, China

Website User Survey



ordinal vs cardinal debates! Seems to hold up empirically with cardinal assumption if there are a lot of points and we treat it as a scale (never analyze individual items). lots of pro-parametric papers.

Demand Characteristics



<https://www.markstivers.com/wordpress/?p=67>

Demand Characteristics



[PLoS One](#). 2012; 7(1): e29081.

Published online 2012 Jan 18. doi: [10.1371/journal.pone.0029081](https://doi.org/10.1371/journal.pone.0029081)

PMCID: PMC3261136

PMID: [22279526](https://pubmed.ncbi.nlm.nih.gov/22279526/)

Behavioral Priming: It's All in the Mind, but Whose Mind?

[Stéphane Doyen](#),^{1,2,3,*} [Olivier Klein](#),² [Cora-Lise Pichon](#),¹ and [Axel Cleeremans](#)¹

also apply to not-surveys! Bargh timer example for people primed with words associated with old age walking slower (which is not true). Experimenters are the participants in this study, and they are subtly told what result they should expect. Their timing is unreliable, and generous in the direction of the desired result. The effect is reproducible in line with the original experiment.

blind your experimenters! and data analysts!

The Story of SAM

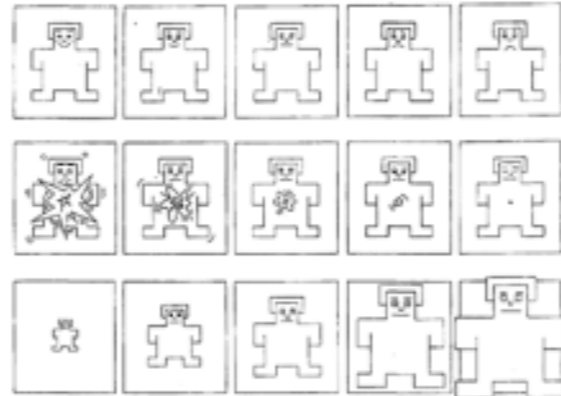


Figure 1. The Self-Assessment Manikin (SAM) used to rate the affective dimension of valence (top panel), arousal (middle panel), and dominance (bottom panel).

(Self Assessment Mannequin) Bradley and Lang 1994, cited ~9k

de facto standard tool for emotion measurement.

The Story of SAM

Osgood (1957/1963)

Gives people lots of word pairs across cultures because interested in simplified semantics/meaning, sees they cluster into three.

Mehrabian and Russell (1974)

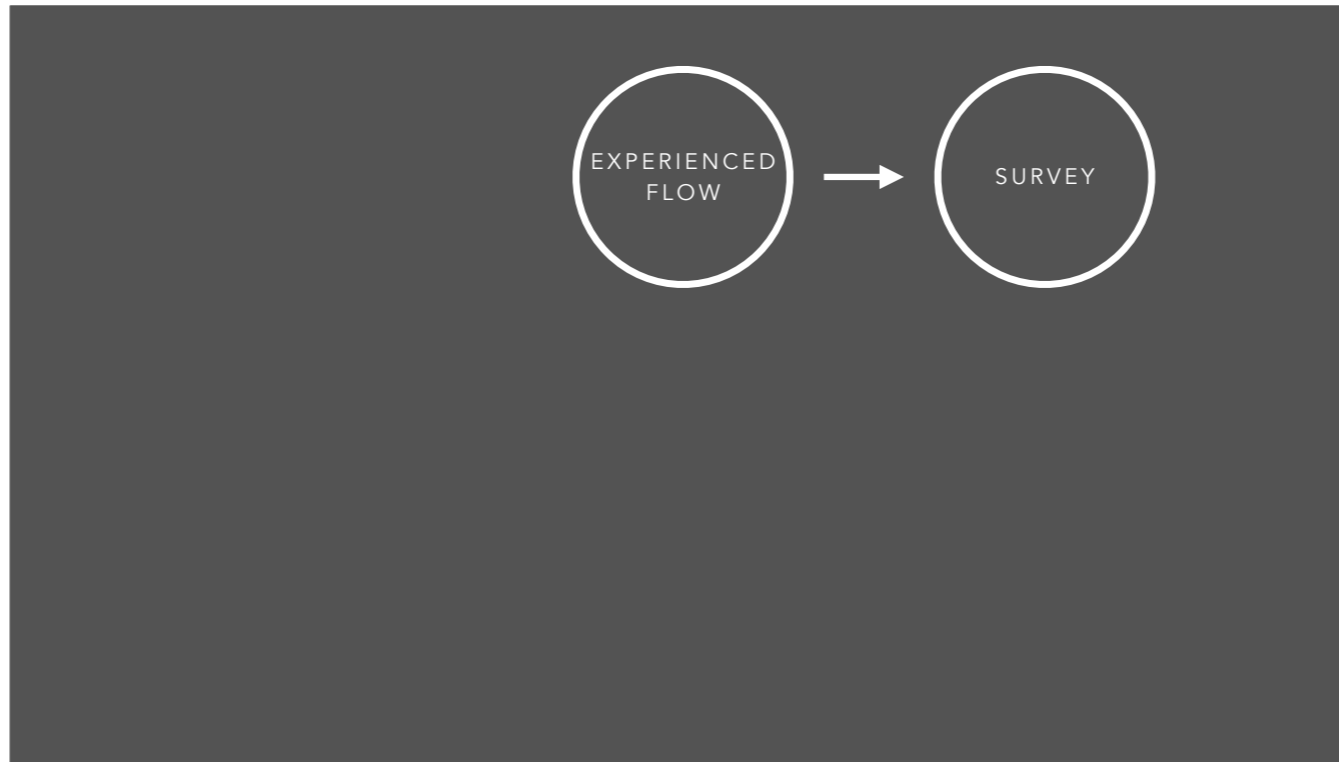
Decide three factors are pleasure, arousal, and dominance and pick 28 word pairs that represent them.
Based on 134 undergrads rating 40 stories, they narrow these word pairs down to 18.

Bradley and Lang (1994)

Compare Mehrabian and Russell ratings to SAM ratings for 21 pictures using 71 college students.

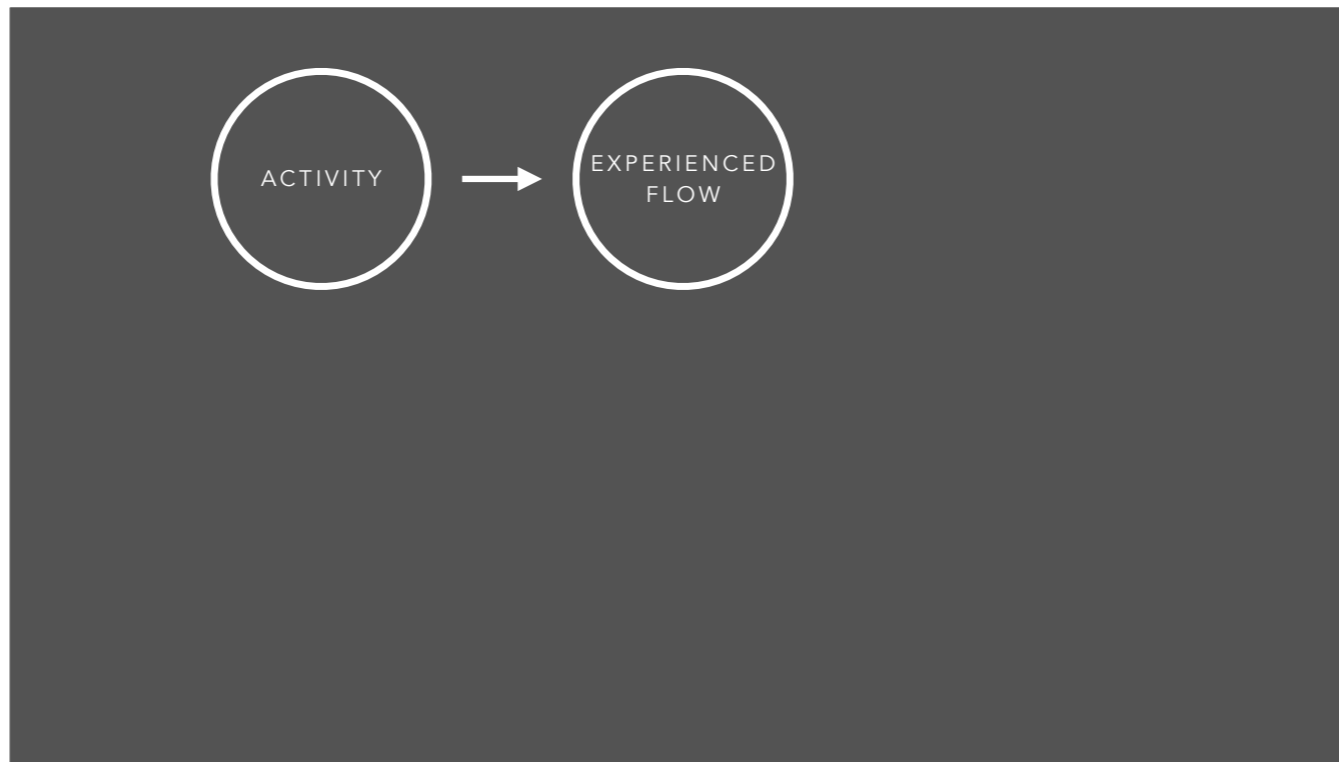
based on semantic meaning of word clusters, in the 50s which showed three main clusters; adapted to environmental psychology and turned into a scale of emotion based on 134 undergraduate emotion ratings of 40 situations across three axes they came up with; SAM is compared against these ratings based on 21 pictures/71 college students to validate it; it does not replicated one of the three axes of the Semantic Difference ratings create by M&R well at all.

Not to say SAM is bad or hasn't been corroborated otherwise, but the history of how it came into being is surprising.



SURVEY: BAD

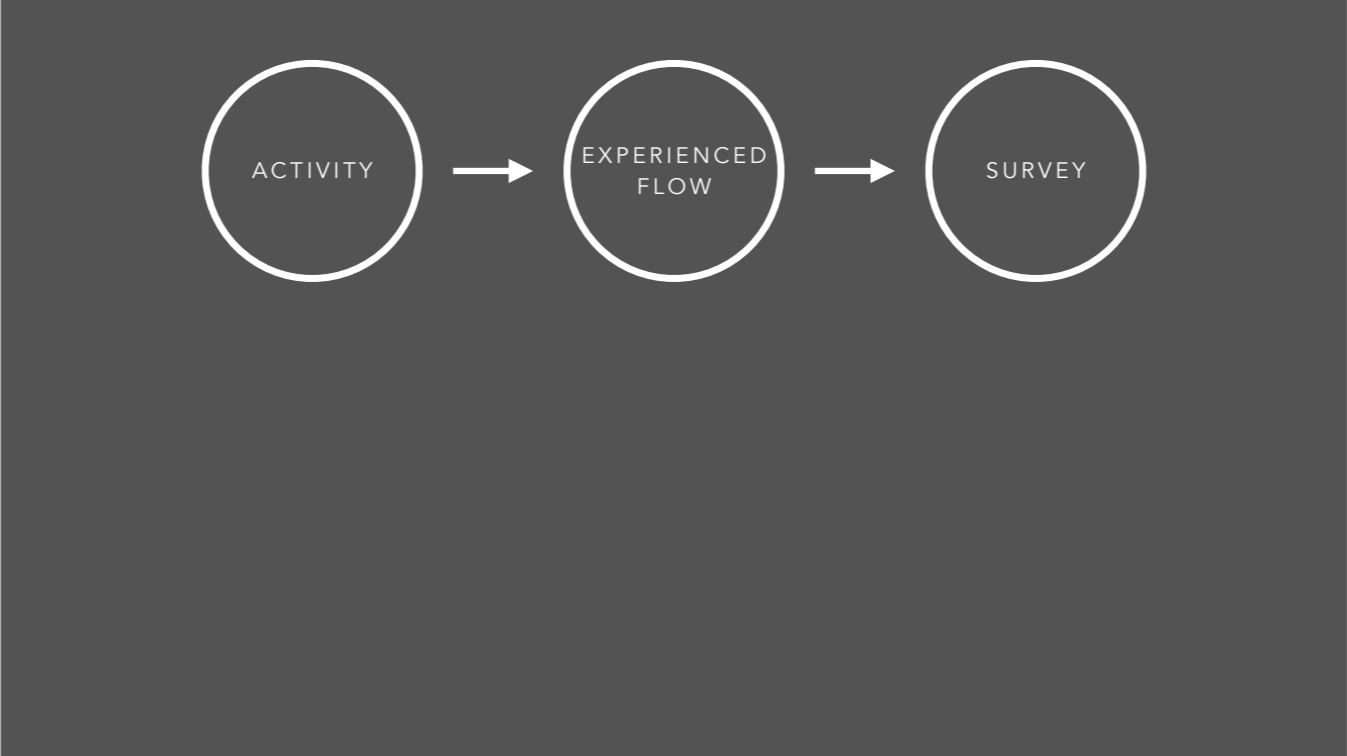
- (1) one number, uncertain, for entire period
- (2) introspecting about a state of lack of ability to introspect
- (3) peak-end rule. the way you remember something is biased



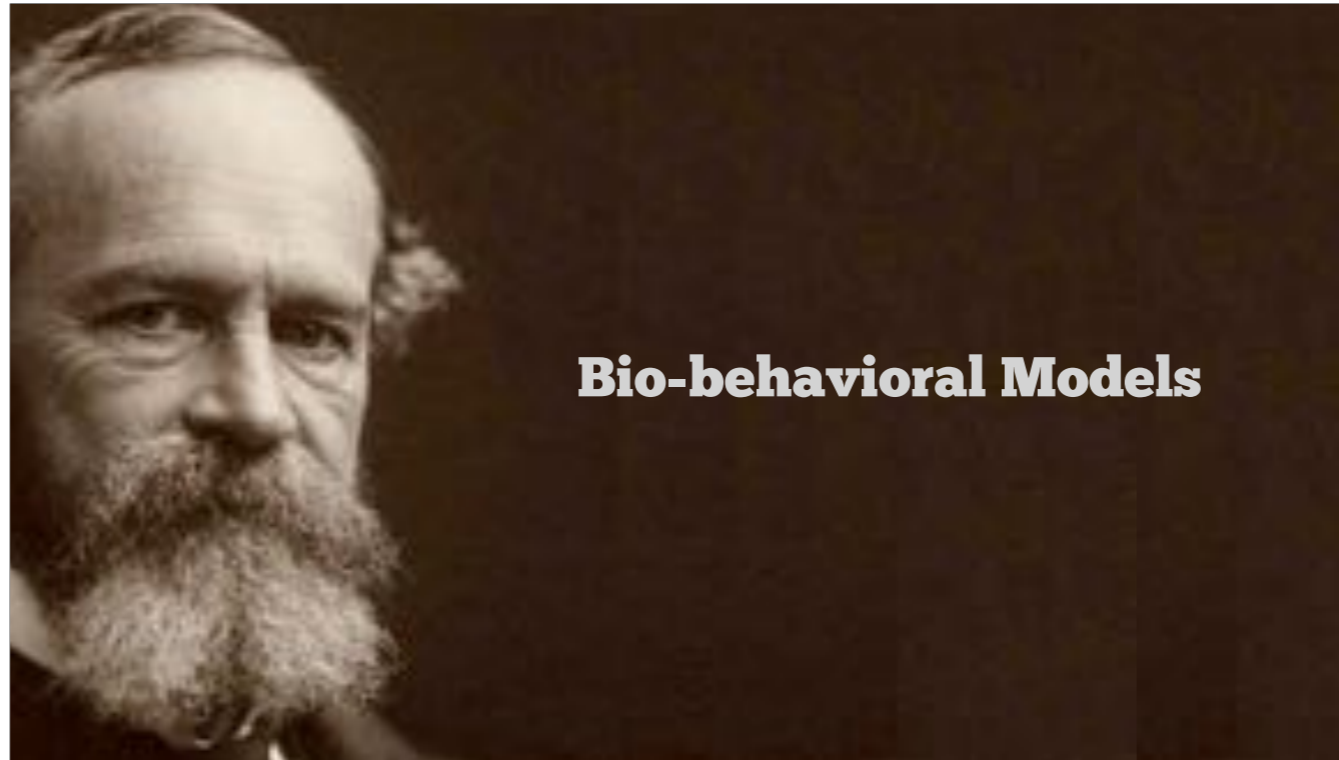
better, but not great.

Binary conception of flow, over an entire period. 'In' when playing; 'out' when not.

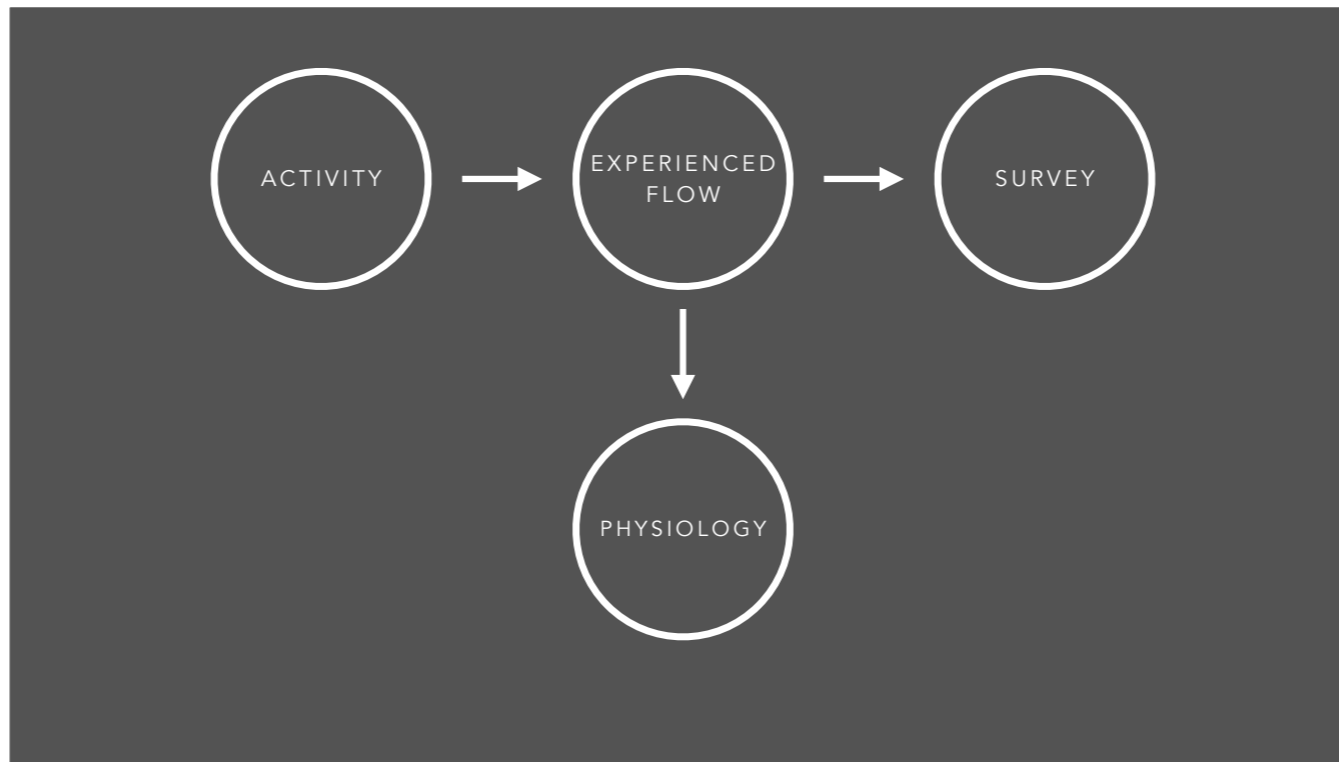
Cognitive load? Have we isolated the concept? Does playing a game in a lab generalize to flow states in your life?



these two are really it.



probabilistic, bio-behavioral models (suggested by William James) are the way forward.

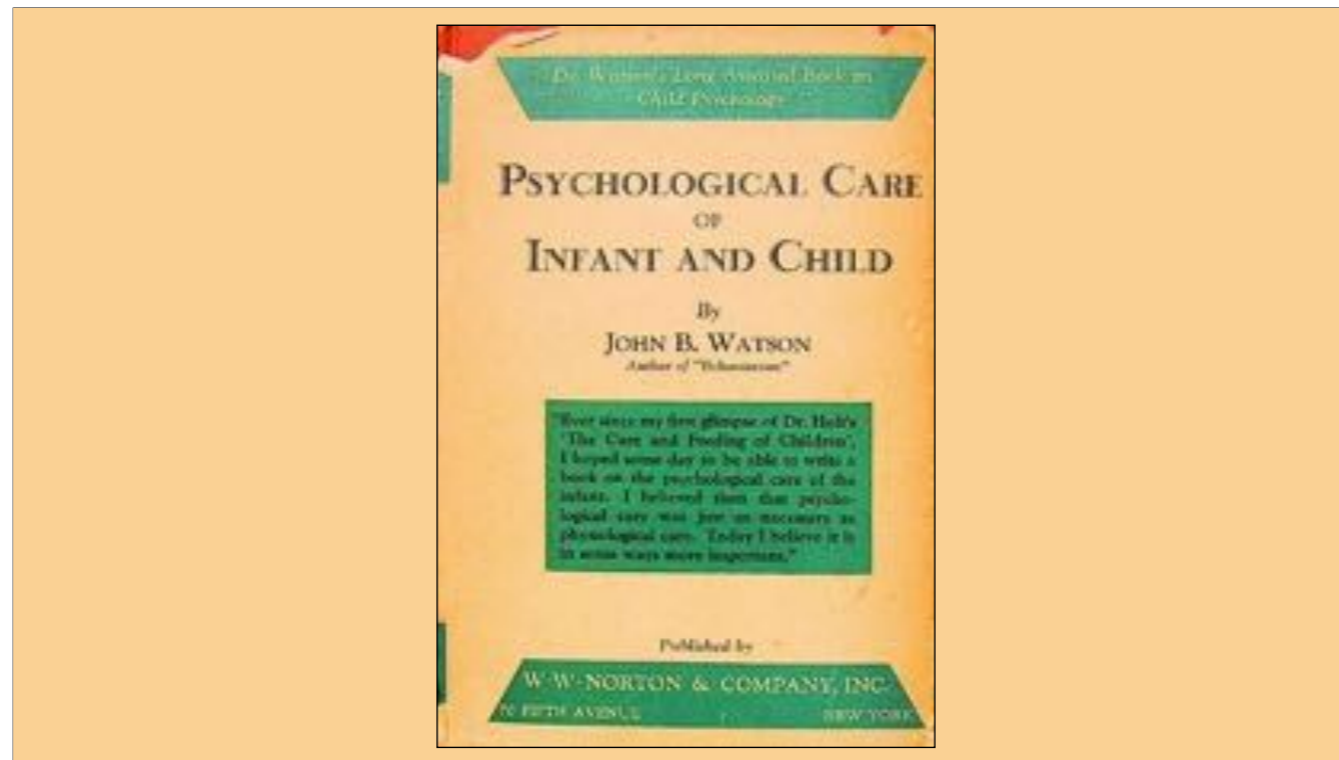


instead of having to use survey or activity as a bad ground truth to compare bio-behavioral signals against to prove they're correct, we should just trust our intuition that they capture something useful and use all methods to create our estimate of flow experience.

**bio-behavioral,
probabilistic measures
with explicit causal
assumptions.**

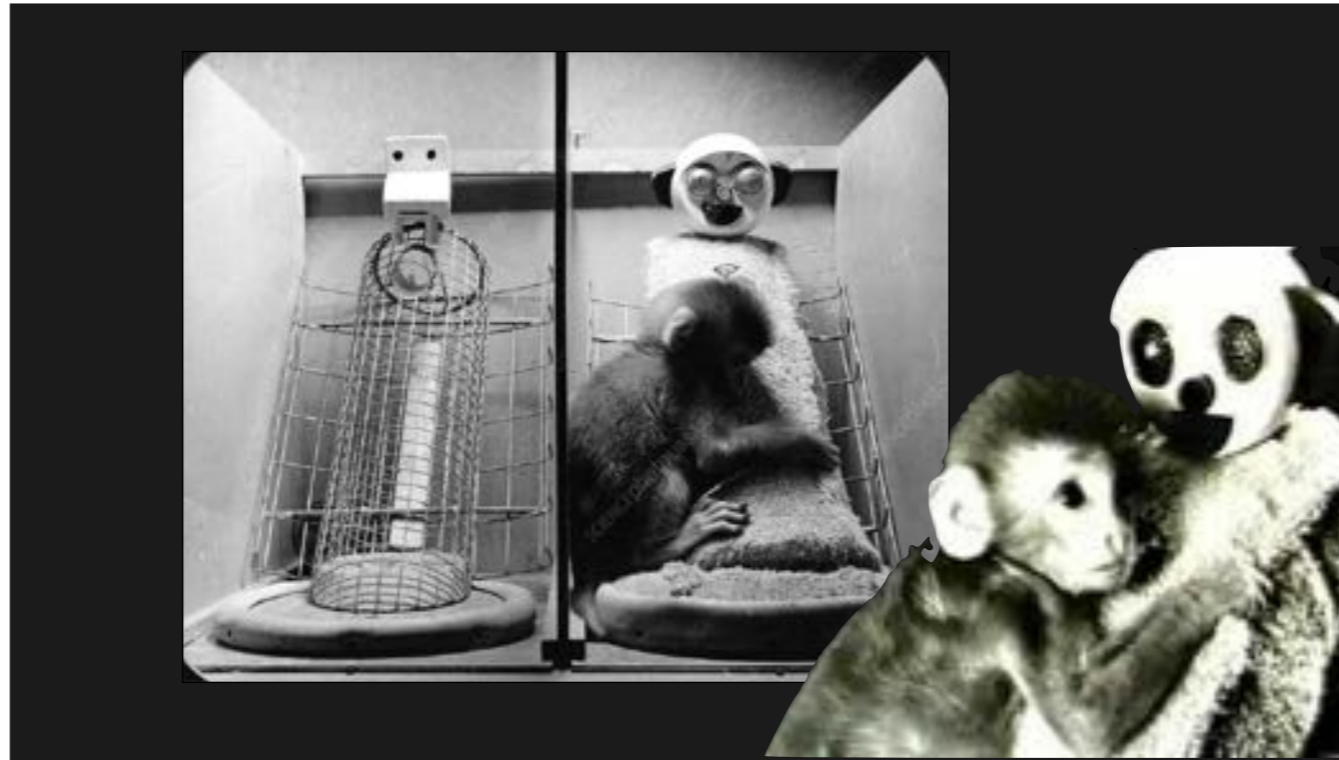
Which Entities to Use?

start with a couple examples from the psychology world getting things really wrong.



1928 Book by John Watson.

“John Watson, held that young children should never be caressed, held or physically comforted by parents. Watson and later behaviorists like B. F. Skinner claimed that a baby reaching for Mom is simply reflecting an association between Mom and food. Early psychologists said that mothers who responded warmly to a baby's cries would produce excessively dependent adults, unable to function in American society.” <https://www.nytimes.com/2003/02/02/books/no-more-wire-mothers-ever.html>



human experience *matters* and is different than *behavior*.

Harry Harlow experiment — baby monkeys prefer the felt monkey mom despite the wire monkey mom delivering the milk. They don't just respond to rewards, they have an innate need to be held.

Crazy that we had an era of behaviorism that shifted focus so completely toward behavior and missed so much about internal mental experience.

$$\mathbf{B} = \mathbf{f}(\mathbf{P}, \mathbf{E})$$

Kurt Lewin

behavior is a function of the person and environment 1936.



1930s— Harvard's Gordon Allport. Extracted 4,500 adjectives from the dictionary and had people answer how applicable all of them were to them. Clustered/Factor analyzed. Viewed as one of the triumphs of psychology.

Clinical Trial > J Pers Soc Psychol. 2002 Aug;83(2):380-93.

Double dissociation between implicit and explicit personality self-concept: the case of shy behavior

Jens B Asendorpf¹, Rainer Banse, Daniel Mücke

Affiliations + expand

PMID: 12150235

Abstract

Using the trait of shyness as an example, the authors showed that (a) it is possible to reliably assess individual differences in the implicitly measured self-concept of personality that (b) are not accessible through traditional explicit self-ratings and (c) increase significantly the prediction of spontaneous behavior in realistic social situations. A total of 139 participants were observed in a shyness-inducing laboratory situation, and they completed an Implicit Association Test (IAT) and explicit self-ratings of shyness. The IAT correlated moderately with the explicit self-ratings and uniquely predicted spontaneous (but not controlled) shy behavior, whereas the explicit ratings uniquely predicted controlled (but not spontaneous) shy behavior (double dissociation). The distinction between spontaneous and controlled behavior was validated in a 2nd study.



<https://replicationindex.com/2020/08/19/personality-science-the-science-of-human-diversity/>

Clinical Trial > J Pers Soc Psychol. 2002 Aug;83(2):380-93.

Double dissociation between implicit and explicit personality self-concept: the case of shy behavior

Jens B Asendorpf¹, Rainer Banse, Daniel Mücke

**r rarely exceeds 0.4 for personality;
explains 16% of the variance**

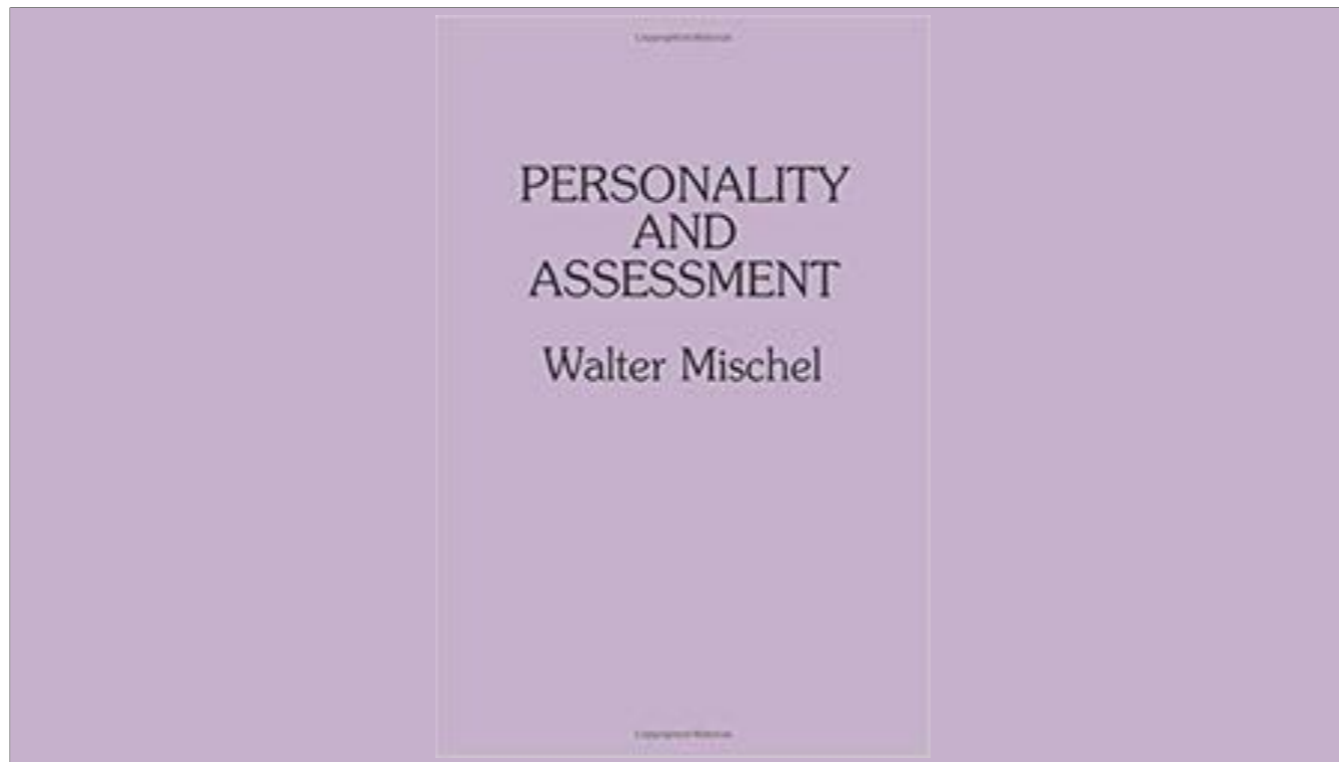


spontaneous behavior in realistic social situations. A total of 139 participants were observed in a shyness-inducing laboratory situation, and they completed an Implicit Association Test (IAT) and explicit self-ratings of shyness. The IAT correlated moderately with the explicit self-ratings and uniquely predicted spontaneous (but not controlled) shy behavior, whereas the explicit ratings uniquely predicted controlled (but not spontaneous) shy behavior (double dissociation). The distinction between spontaneous and controlled behavior was validated in a 2nd study.

<https://replicationindex.com/2020/08/19/personality-science-the-science-of-human-diversity/>

“Self-report measure of shyness was a significant predictor of gaze aversion, $r = .18$ and speech duration, $r = -.31$. It also correlated $r = .40$ with observer ratings of shyness.

A correlation of $r = .40$ means that somebody who scored above average on the shyness measure still has a 30% probability to act less shy than the average participant in this specific situation. Furthermore, if we square this correlation, we find that trait-shyness explains only 16% of the variance in a specific behaviour. This suggests that most of the variance is explained by other factors, including the situation.” Ulrich Schimmack’s Textbook



Argument in Mischel: those correlations are so small for personality cross-situationally, all that is left is the environment. We should focus on the environment. Destroyed Personality Psychology; Social Psychology took over.

1968.

One Hundred Years of Social Psychology Quantitatively Described

F. D. Richard, Charles F. Bond, Jr., Juli J. Stokes-Zoota

First Published December 1, 2003 | Research Article

<https://doi.org/10.1037/1089-2680.7.4.331>

Article information ▾

Altmetric 51



Abstract

This article compiles results from a century of social psychological research, more than 25,000 studies of 8 million people. A large number of social psychological conclusions are listed alongside meta-analytic information about the magnitude and variability of the corresponding effects. References to 322 meta-analyses of social psychological phenomena are presented, as well as statistical effect-size summaries. Analyses reveal that social psychological effects typically yield a value of r equal to .21 and that, in the typical research literature, effects vary from study to study in ways that produce a standard deviation in r of .15. Uses, limitations, and implications of this large-scale compilation are noted.



<https://replicationindex.com/2020/08/19/personality-science-the-science-of-human-diversity/>

One Hundred Years of Social Psychology Quantitatively Described

F. D. Richard, Charles F. Bond, Jr., Juli J. Stokes-Zoota

First Published December 1, 2003 | Research Article

<https://doi.org/10.1037/1089-2680.7.4.331>

r averages to 0.2 for situation; explains 4% of the variance

analyses of social psychological phenomena are presented, as well as statistical effect-size summaries.

Analyses reveal that social psychological effects typically yield a value of r equal to .21 and that, in the typical research literature, effects vary from study to study in ways that produce a standard deviation in r of .15.

Uses, limitations, and implications of this large-scale compilation are noted.



<https://replicationindex.com/2020/08/19/personality-science-the-science-of-human-diversity/>

100 years/25000 studies of 8 million people in social psychology give an average effect size of $r \sim 0.2$.

Social psych is overestimating effect sizes, as we've shown; it also has no measurement error re: the independent variable compared with personality psychology where there is a lot.

Environments

Personality taxonomy seen as a triumph. What about environments?

You might think with all the effort on social psychology we'd have some good theories here, but no.



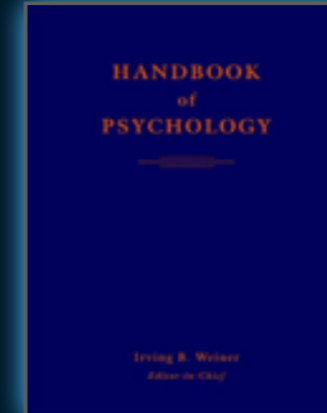
Stimulation - arousal theory, adaptation level theory

Agency - control over environment mediates stress

Barker's Behavioral Settings

Gibson's Affordances

Carter's Place Identity and Attachment



Taxonomies for specific areas— nature, architecture, home, work.

Generalizations— top are deterministic. bottom are relational/phenomenological.

Stimulation — arousal theory, adaptation level theory. Stimulation level of environment has behavioral effects; individuals have different personal preferences

Agency — control of environment mediates stress.

Barker's behavioral settings. Worshipping at churches. Settings and behavior/function co-evolve, exert pressure towards function.

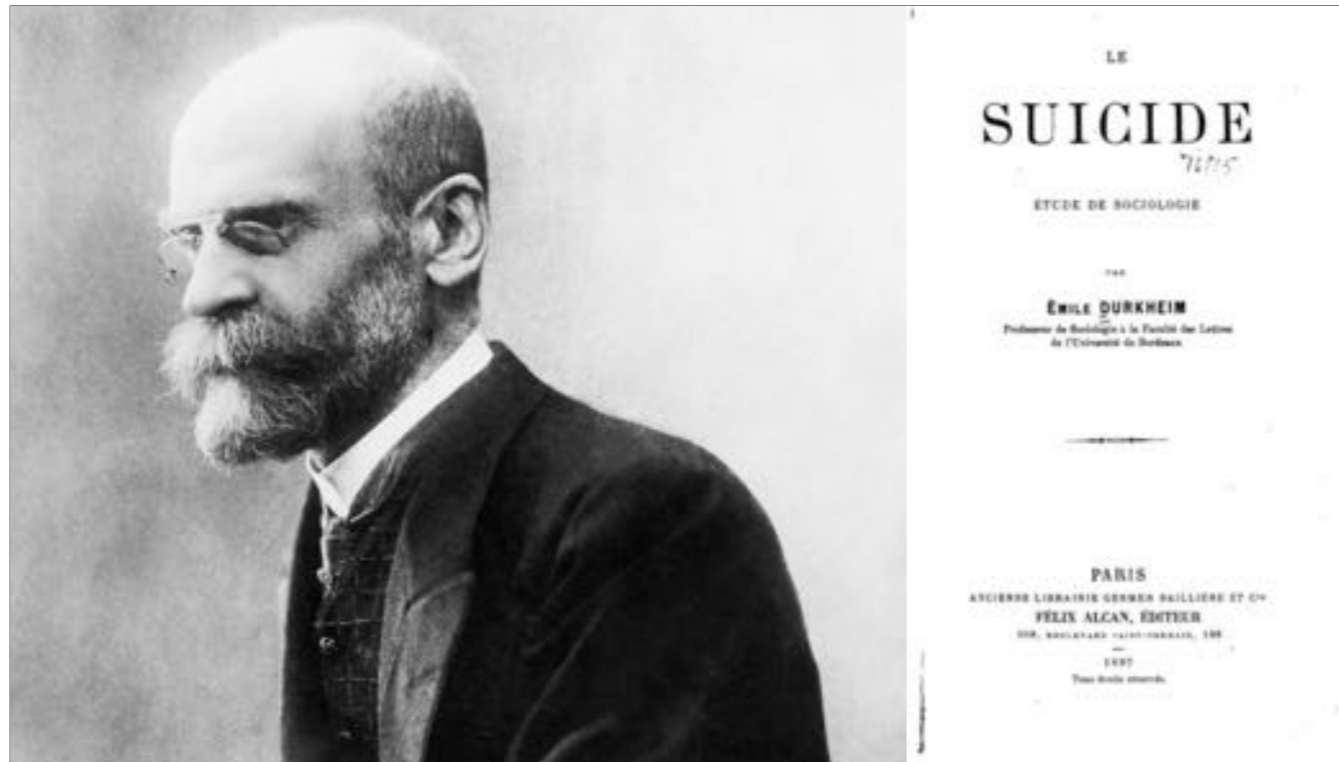
Gibson's Affordances. Think of environment through obvious physical interactions available.

Carter's place identity/attachment. social imageability. Place has conferred meaning by people.



privilege some aspects of the environment as it shapes human experience, belief, and behavior over others. Consider all of them! All of these people are right, but they all should exist in conversation.

Levels of Analysis



social facts and macro scale analysis. I like individual level, but there are some fascinating things to talk about at the sociological level too.

RCTs to Scale: Comprehensive Evidence from Two Nudge Units*

Stefano DellaVigna Elizabeth Linos
UC Berkeley and NBER UC Berkeley

July 2020

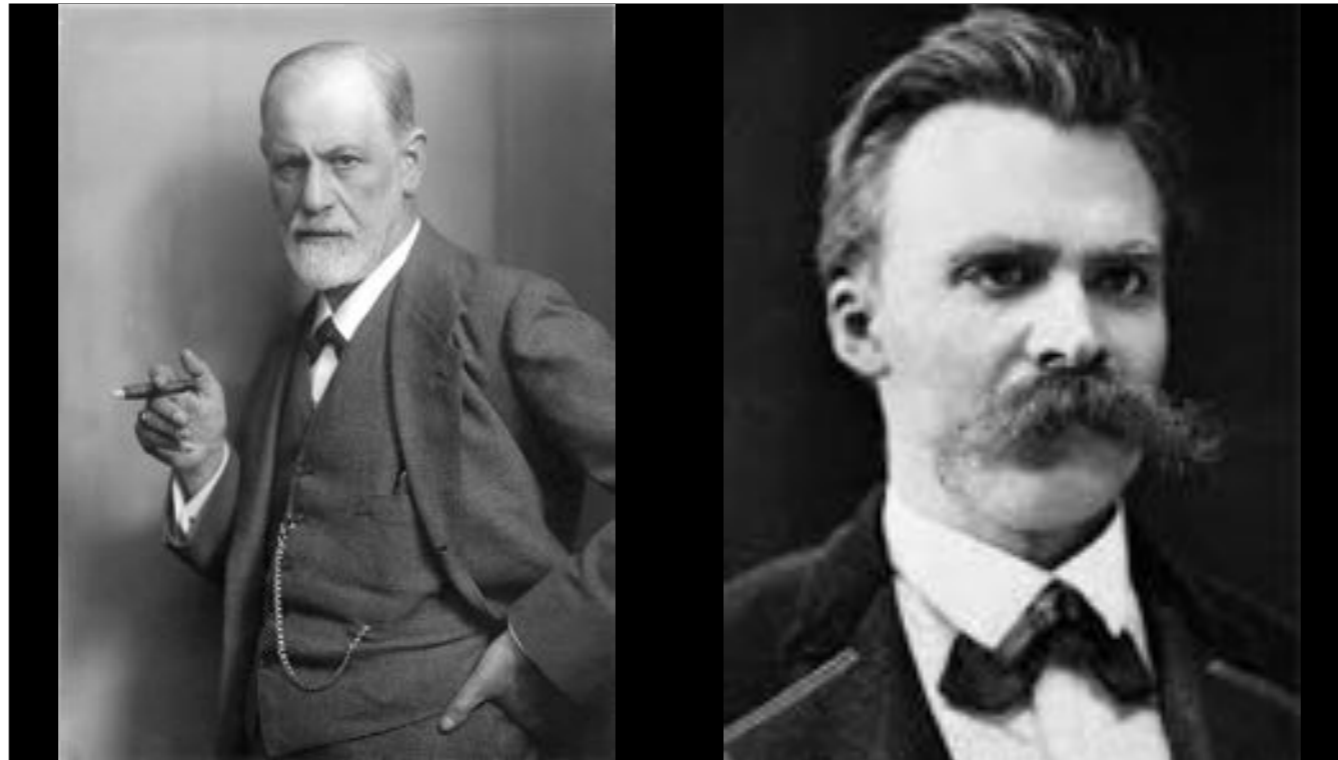
Abstract

Nudge interventions – behaviorally-motivated design changes with no financial incentives – have quickly expanded from academic studies to larger implementation in so-called Nudge Units in governments. This provides an opportunity to compare interventions in research studies, versus at scale. We assemble a unique data set of 126 RCTs covering over 23 million individuals, including all trials run by two of the largest Nudge Units in the United States. We compare these trials to a separate sample of nudge trials published in academic journals from two recent meta-analyses. In papers published in academic journals, the average impact of a nudge is very large – an 8.7 percentage point take-up effect, a 33.5% increase over the average control. In the Nudge Unit trials, the average impact is still sizable and highly statistically significant, but smaller at 1.4 percentage points, an 8.1% increase. We consider five potential channels for this gap: statistical power, selective publication, academic involvement, differences in trial features and in nudge features. Publication bias in the academic journals, exacerbated by

My favorite example of mis-understanding levels of analysis. Nudges do something meaningful at population levels; do they work at the individual level? Probably too weakly to notice in most cases. If you're going to apply these in individual level experiments, you should be very explicit about your reasoning and assumptions.

Inversions of Perspective

3 examples



psychopathology vs existentialism. You're normal unless something traumatizes you, or you're traumatized (like a child left by themselves as a mall) unless we structure a lot around you to make you feel secure?

Totally different approaches— one you look for a cause of trauma, one you look for the degradation of a previous sense of security.

Pathology vs. Flourishing

DSM vs. CSV

DSM vs. CSV (character strength and virtues).

humanistic and positive psychology claims vs old days of psychopathology research.



addiction caused by introduction of addictive things vs. removal of healthy context?

**survey the landscape
and consider all causal
forces before pruning
your scientific model.**

Do these relationships generalize?

RatSWD
Working Paper Series

Working Paper No. 139

The Weirdest People in the World?

Joseph Henrich, Steven J. Heine and Ara Norenzayan

April 2010



Get the obvious ones out of the way first.

Western, Educated, Industrialized, Rich, and Democratic.

73% American first authors, ~ 70% of all participants are intro psychology college students. Outliers in many ways. conformity, rationalization, moral reasoning, social networks.



lab studies to the real world. setting is contrived, pressure, observed; unavoidable, major differences relative to real life and a very different mental state elicited in this situation.

$$\mathbf{B} = \mathbf{f}(\mathbf{P}, \mathbf{E})$$

Kurt Lewin

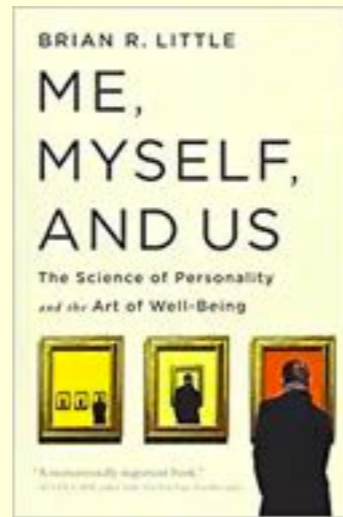
a bigger one is what we talked about before



Figure 1: An interpretation of Herbert Simon's 'behavioural scissors'

Lockton, Dan. "Simon's scissors and ecological psychology in design for behaviour change." Available at SSRN 2125405 (2012).

you can't necessarily separate environment and personality



Mark Snyder — High vs. Low Self Monitoring.

High = environment guides behavior. test food, then salt. 'be more socially appropriate.'
Low = trait guides behavior. salt before trying food. 'don't be fake.'

Biogenic, Idiogenic, and Sociogenic Structure of 'Personality'

Act against biological disposition based on social or personal values

Big 5

Conscientiousness mediates response to ordered vs chaotic environments
Neuroticism mediates response to negative external cues

Introversion vs Extraversion

'Introverts' are over-stimulated and enjoy calm
'Extraverts' are under-stimulated and enjoy excitement

Meyers Briggs: these are normally distributed traits- not binary- and perhaps not a good model at all (situationally contingent, more fundamental descriptors)

interactions of environment and personality illicit different responses from people. We know this. Here are some theories.

For example, introverts and extroverts will feel/respond/act very differently to environments with different levels of stimulation. You can't simply average personality out.

Strong vs. Weak Situations

interactions of environment and personality illicit different responses from people. This is obvious; some situations produce far less variability in behavior than others. Obviously there is a complex interaction between situation and personality, especially as situations get 'weaker'.

Idiographic vs. Nomothetic

Gestalt vs Analytical



experience of chills and immersion at a concert cannot be studied *analytically; cannot be decomposed*. It happens or it doesn't; if you take away one aspect of the experience it all falls apart. Not decomposable. 'The whole is greater than the sum of its parts.'



Hot Milks @ MFA by media lab students.

Do all the things that have been reported in the affective computing literature as good; some you'll recognize. Make your smile muscles engage. watch images of butterflies and clowns and kittens. Have someone stroke your face or massage you. Placebo pills that will make them happy, tickling, sounds of kids laughing.

Created a truly creepy experience that is very unpleasant using only 'good and positive' psychology stimuli.

**measure real-world data
from a representative
group; consider focusing
on individuals.**

Make your Scientific Model Explicit.

That was a review of ways we mess up choosing entities or misconstruing relationships in the underlying process model of our approach to psychology. There are many, many examples. They lead to really bad science.

Draw an explicit process model or data generating model (DGM) or scientific model; be thorough and wise in what you choose.

Resources

- The Book of Why (Judea Pearl) and other of his talks (https://www.youtube.com/watch?v=mfh4fp_8oPg).
- Statistical Rethinking (Richard McElreath), also available on youtube.
- MIT Computational Cognitive Science Course (Josh Tenenbaum).
- Ulrich Schimmack's Personality Textbook
- Me, Myself, and Us by Brian Little