

HOW SHOULD WE CRITIQUE RESEARCH?

tags: epistemology, Bayes, scientific bias, causality,
decision theory

Criticizing studies and statistics is hard in part because so many criticisms are possible, rendering them meaningless. What makes a good criticism is the chance of being a ‘difference which makes a difference’ to our ultimate actions.

2019-05-19_{2m} 2019-07-07 · *finished* · certainty: *highly likely* · importance: 8

Valley of Bad Statistics

All Things Large and Small

Relevant But Not Definitive

Bad Criticisms

Good Criticisms

Beliefs Are For Actions

Decision-Theoretic Criticisms

External Links

Appendix

Teaching Statistics

Footnotes

Backlinks

Scientific and statistical research must be read with a critical eye to understand how credible the claims are. The Reproducibility Crisis and the growth of meta-science have demonstrated that much research is of low quality and often false.

But there are so many possible things any given study could be criticized for, falling short of an unobtainable ideal, that it becomes unclear which possible criticism is important, and they may degen-

erate into mere rhetoric. How do we separate fatal flaws from unfortunate caveats from specious quibbling?

I offer a pragmatic criterion: what makes a criticism important is how much it could change a result if corrected and how much that would then change our decisions or actions: to what extent it is a “difference which makes a difference”.

This is why issues of research fraud, causal inference, or biases yielding overestimates are universally important: because a ‘causal’ effect turning out to be zero effect or grossly overestimated will change almost all decisions based on such research; while on the other hand, other issues like measurement error or distributional assumptions, which are equally common, are often *not* important: because they typically yield much smaller changes in conclusions, and hence decisions.

If we regularly ask whether a criticism would make this kind of difference, it will be clearer which ones are important criticisms, and which ones risk being rhetorical distractions and obstructing meaningful evaluation of research.



LEARNING STATISTICS IS GREAT. IF YOU WANT TO READ AND understand scientific papers in general, there’s little better to learn than statistics because *everything* these days touches on statistical issues and draws on increasingly powerful statistical methods and large datasets, whether flashy like machine learning or mundane like geneticists drawing on biobanks of millions of people, and if you don’t have at least some grasp of statistics, you will be increasingly left out of scientific and technological progress and unable to meaningfully discuss their application to society, so you must have a good grounding in statistics if you are at all interested in these topics—or so I want to say. The problem is... learning statistics can be dangerous.

VALLEY OF BAD STATISTICS

Like learning some formal logic or about cognitive biases, statistics seems like the sort of thing one might say “A little learning is a dangerous thing / Drink deep, or taste not the Pierian spring / There shallow draughts intoxicate the brain, / And drinking largely sobers us again.”

When you first learn some formal logic and about fallacies, it's hard to not use the shiny new hammer to go around playing ‘fallacy bingo’ (to mix metaphors): “aha! that is an *ad hominem*, my good sir, and a logically invalid objection.” The problem, of course, is that many fallacies are perfectly good as a matter of inductive logic: *ad hominem*s are often highly relevant (eg. if the person is being bribed). A rigorous insistence on formal syllogisms will at best waste a lot of time, and at worst becomes a tool for self-delusion by selective application of rigor.

Similarly, cognitive biases are hard to use effectively (because they are informative priors in some cases, and in common harmful cases, one will have already learned better), but are easy to abuse—it's always easiest to see how someone *else* is sadly falling prey to confirmation bias.

ALL THINGS LARGE AND SMALL

With statistics, a little reading and self-education will quickly lead to learning about a universe of ways for a study to screw up statistically, and as skeptical as one quickly becomes, as Ioannidis and Gelman and the Replicability Crisis and far too many examples of scientific findings completely collapsing show, one probably isn't skeptical enough because there are in fact an awful lot of screwed up studies out there. Here are a few potential issues, in deliberately no particular order:

- ✦ “spurious correlations” caused by data processing (such as ratio/percentage data, or normalizing to a common time-series)
- ✦ multiplicity: many subgroups or hypotheses tested, with only statistically-significant ones reported and no control of the overall

false detection rate

- ✦ missingness not modeled
- ✦ in vivo animal results applied to humans
- ✦ experiment run on regular children rather than identical twins
- ✦ publication bias detected in meta-analysis
- ✦ a failure to reject the null or a positive point-estimate being interpreted as evidence for the null hypothesis
- ✦ “the difference between statistically-significant and non-statistically-significant is not statistically-significant”
- ✦ choice of an inappropriate distribution, like modeling a log-normal variable by a normal variable (“they strain at the gnat of the prior who swallow the camel of the likelihood”)
- ✦ no use of single or double-blinding or placebos
- ✦ a genetic study testing correlation between 1 gene and a trait
- ✦ an IQ experiment finding an intervention increased before/after scores on some IQ subtests and thus increased IQ
- ✦ cross-sectional rather than longitudinal study
- ✦ ignoring multilevel structure (like data being collected from sub-units of schools, countries, families, companies, websites, individual fishing vessels, WP editors etc)
- ✦ reporting performance of GWAS polygenic scores using only SNPs which reach genome-wide statistical-significance
- ✦ nonzero attrition but no use of intent-to-treat analysis
- ✦ use of a fixed alpha threshold like 0.05
- ✦ correlational data interpreted as causation
- ✦ use of an “unidentified” model, requiring additional constraints or priors

- ✦ non-preregistered analyses done after looking at data; *p*-hacking of every shade
- ✦ use of cause-specific mortality vs all-cause mortality as a measurement
- ✦ use of measurements with high levels of measurement error (such as dietary questionnaires)
 - ✦ ceiling/floor effects (particularly IQ tests)
- ✦ claims about latent variables made on the basis of measurements of greatly differing quality
 - ✦ or after “controlling for” intermediate variables, comparing total effects of one variable to solely indirect effects of another
 - ✦ or that one variable mediates an effect without actually setting up a mediation SEM
- ✦ studies radically underpowered to detect a plausible effect
- ✦ the “statistical-significance filter” inflating effects
- ✦ base rate fallacy
- ✦ self-selected survey respondents; convenience samples from Mechanical Turk or Google Surveys or similar services
- ✦ animal experiments with randomization not blocked by litter/cage/room
- ✦ using a large dataset and obtaining many statistically-significant results
- ✦ factor analysis without establishing measurement invariance
- ✦ experimenter demand effects
- ✦ using a SVM/NN/RF without crossvalidation or heldout sample
 - ✦ using them, but with data preprocessing done or hyperparameters selected based on the whole dataset

- ✦ passive control groups
- ✦ not doing a factorial experiment but testing one intervention on each group
- ✦ flat priors overestimating effects
- ✦ reporting of relative risk increase without absolute risk increase
- ✦ a genetic study testing correlation between 500,000 genes and a trait
- ✦ conflicts of interest by the researchers/funders
- ✦ lack of power analysis to design experiment
- ✦ analyzing Likert scales as a simple continuous cardinal variable
- ✦ animal results in a single inbred or clonal strain, with the goal of reducing variance/increased power (Michie 1955)
- ✦ right-censored data
- ✦ temporal autocorrelation of measurements
- ✦ genetic confounding
- ✦ reliance on interaction terms

Some of these issues are big issues—even fatal, to the point where the study is not just meaningless but the world would be a better place if the researchers in question had never published. Others are serious but while regrettable, a study afflicted by it is still useful and perhaps the best that can reasonably be done. And some flaws are usually minor, almost certain not to matter, possibly to the point of being misleading to bring up at all as a ‘criticism’ as it implies that the flaw is worth discussing. And many are completely context-dependent, and could be anything from instantly fatal to minor nuisance.

But which are which? You can probably guess at where a few of them fall, but I would be surprised if you knew what I meant by all of them, or had well-justified beliefs about how important each is, because I don’t, and I suspect few people do. Nor can anyone tell you how

important each one is. One just has to learn by experience, it seems, watching things replicate or diminish in meta-analyses or get debunked over the years, to gradually get a feel of what is important. There are checklists and professional manuals¹ which one can read and employ, and they at least have the virtue of checklists in being systematic reminders of things to check, reducing the temptation to cherry-pick criticism, and I recommend their use, but they are not a complete solution. (In some cases, they recommend quite bad things, and none can be considered *complete*.)

No wonder that statistical criticism can feel like a blood-sport, or feel like learning statistical-significance statistics: a long list of special-case tests with little rhyme or reason, making up a “cookbook” of arbitrary formulas and rituals, useful largely for “middlebrow dismissals”.

After a while, you have learned enough to throw a long list of criticisms at any study regardless of whether they are relevant or not, engaging in “pseudo-analysis”², which devalues criticism (surely studies can’t *all* be equally worthless) and risks the same problem as with formal logic or cognitive biases—of merely weaponizing it and having laboured solely to make yourself more wrong, and defend your errors in more elaborate ways. (I have over the years criticized many studies and while for many of them my criticisms were much less than they deserved and have since been borne out, I could not honestly say that I have always been right or that I did not occasionally ‘gild the lily’ a little.)

RELEVANT BUT NOT DEFINITIVE

So, what do we mean by statistical criticism? what makes a good or bad statistical objection?

BAD CRITICISMS

“ Here I should like to say: a wheel that can be turned though nothing else moves with it, is not part of the mechanism.

Ludwig Wittgenstein, §271, Philosophical Investigations

”

It can't just be that a criticism is boring and provokes eye-rolling—someone who in every genetics discussion from ~2000₁₀–2010_{14ya} harped on statistical power & polygenicity and stated that all these exciting new candidate-gene & gene-environment interaction results were so much hogwash and the entire literature garbage would have been deeply irritating to read, wear out their welcome fast, and have been absolutely right. (Or for nutrition research, or for social psychology, or for...) As provoking as it may be to read yet another person sloganize “correlation \neq causation” or “yeah, in mice!”, unfortunately, for much research *that is all that should ever be said* about it, no matter how much we weary of it.

It can't be that some assumption is violated (or unproven or unprovable), or that some aspect of the real world is left out, because all statistical models are massively abstract, gross simplifications. Because it is *always* possible to identify some issue of inappropriate assumption of normality, or some autocorrelation which is not modeled, or some nonlinear term not included, or prior information left out, or data lacking in some respect. Checklists and preregistrations and other techniques can help improve the quality considerably, but will never solve this problem. Short of tautological analysis of a computer simulation, there is not and never has been a perfect statistical analysis, and if there was, it would be too complicated for anyone to understand (which is a criticism as well). All of our models are false, but some may be useful, and a good statistical analysis is merely ‘good enough’.

It can't be that results “replicate” or not. Replicability doesn't say much other than if further data were collected the same way, the results would stay the same. While a result which doesn't replicate is of questionable value at best (it most likely wasn't real to begin with³), a

result being replicable is no guarantee of quality either. One may have a consistent GIGO process, but replicable garbage is still garbage. To collect more data may be to simply more precisely estimate the process's systematic error and biases. (No matter how many published homeopathy papers you can find showing homeopathy works, it doesn't.)

It certainly has little to do with p -values, either in a study or in its replications (because nothing of interest has to do with p -values); if we correct an error and change a specific p -value from $p = 0.05$ to $p = 0.06$, so what? ("Surely, God loves the 0.06 nearly as much as the 0.05...") Posterior probabilities, while meaningful and important, also are no criterion: is it important if a study has a posterior probability of a parameter being greater than zero of 95% rather than 94%? Or >99%? Or >50%? If a criticism, when corrected, reduces a posterior probability from 99% to 90%, is that what we mean by an important criticism? Probably (ahem) not.

It also doesn't have to do with any increase or decrease in effect sizes. If a study makes some errors which means that it produces an effect size twice as large as it should, this might be absolutely damning or it might be largely irrelevant. Perhaps the uncertainty was at least that large so no one took the point-estimate at face-value to begin with, or everyone understood the potential for errors and understood the point-estimate was an upper bound. Or perhaps the effect is so large that overestimation by a factor of 10 wouldn't be a problem.

It usually doesn't have to do with predictive power (whether quantified as R^2 or AUC etc); sheer prediction is the goal of a subset of research (although if one could show that a particular choice led to a lower predictive score, that would be a good critique), and in many contexts, the best model is not particularly predictive at all, and a model being *too* predictive is a red flag.

GOOD CRITICISMS

“ *The statistician is no longer an alchemist expected to produce gold from any worthless material offered him. He is more like a chemist capable of assaying exactly how much of value it contains, and capable also of extracting this amount, and no more. In these circumstances, it would be foolish to commend a statistician because his results are precise, or to reprove because they are not. If he is competent in his craft, the value of the result follows solely from the value of the material given him. It contains so much information and no more. His job is only to produce what it contains...Immensely laborious calculations on inferior data may increase the yield 95 → 100 per cent. A gain of 5 per cent, of perhaps a small total. A competent overhauling of the process of collection, or of the experimental design, may often increase the yield ten or twelve fold, for the same cost in time and labour. ...To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.*

R. A. Fisher, “Presidential address to the first Indian statistical congress”, 1938

”

What would count as a good criticism?

Well, if a draft of a study was found and the claims were based on a statistically-significant effect in one variable, but in the final published version, it omits that variable and talks only about a different variable, one would wonder. Discovering that authors of a study had been paid millions of dollars by a company benefiting from the study results would seriously shake one's confidence in the results. If a correlation didn't exist at all when we compared siblings within a family, or better yet, identical twins, or if the correlation didn't exist in other datasets, or other countries, then regardless of how strongly supported it is in that one dataset, it would be a concern. If a fancy new machine learning model outperformed SOTA by 2%, but turned out to not be us-

ing a heldout sample properly and actually performed the same, doubtless ML researchers would be less impressed. If someone showed an RCT reached the opposite effect size to a correlational analysis, that would strike most people as important. If a major new cancer drug was being touted as being as effective as the usual chemotherapy with fewer side-effects in the latest trial, and one sees that both were being compared to a null hypothesis of zero effect and the point-estimate for the new drug was lower than the usual chemotherapy, would patients want to use it? If a psychology experiment had different results with a passive control group and an active control group, or a surgery's results depend on whether the clinical trial used blinding, certainly an issue. And if data was fabricated entirely, that would certainly be worth mentioning.

These are all inherently different going by some of the conventional views outlined above. So what do they have in common that makes them good criticisms?

BELIEFS ARE FOR ACTIONS

“ *Results are only valuable when the amount by which they probably differ from the truth is so small as to be insignificant for the purposes of the experiment. What the odds should be depends:*

1. *On the degree of accuracy which the nature of the experiment allows, and*
2. *On the importance of the issues at stake.*

Student, 1904⁴

”

“ *Moreover, the economic approach seems (if not rejected owing to aristocratic or puritanic taboos) the only device apt to distinguish neatly what is or is not contradictory in the logic of uncer-*

tainty (or probability theory). That is the fundamental lesson supplied by Wald's notion of 'admissibility'...probability theory and decision theory are but two versions (theoretical and practical) of the study of the same subject: uncertainty.

Bruno de Finetti, "Comment on Savage's 'On Rereading R. A. Fisher'", 1976

”

But what I think they share in common is this decision-theoretic justification which unifies criticisms (and would unify statistical pedagogy too):


The importance of a statistical criticism is the probability that it would change a hypothetical decision based on that research.

I would assert that p -values are not posterior probabilities are not effect sizes are not utilities are not profits are not decisions. Dichotomies come from decisions. All analyses are ultimately decision analyses: our beliefs and analyses may be continuous, but our actions are discrete.

When we critique a study, the standard we grope towards is one which ultimately terminates in real-world actions and decision-making, a standard which is inherently context-dependent, admits of no bright lines, and depends on the use and motivation for research, grounded in what is the right thing to do. ⁵

It doesn't have anything to do with attaining some arbitrary level of "significance" or being "well-powered" or having a certain k in a meta-analysis for estimating heterogeneity, or even any particular posterior probability, or effect size threshold; it doesn't have anything to do with violating a particular assumption, unless, by violating that assumption, the model is not 'good enough' and would lead to bad choices; and it is loosely tied to replication (because if a result doesn't replicate in the future situations in which actions will be taken, it's not useful for planning) but not defined by it (as a result could replicate fine while still being useless).

The importance of many of these criticisms can be made much more intuitive by asking what the research is for and how it would affect a downstream decision. We don't need to do a formal decision analysis going all the way from data through a Bayesian analysis to utilities and a causal model to compare (although this would be useful to do and might be necessary in edge cases), an informal consideration can be a good start, as one can intuitively guess at the downstream effects.

STUDY PROVES X!	
LOL: $n=12$ PROVES NOTHING	
But $D=big$: so, well-powered!	
Single studies prove nothing due to heterogeneity	
Multilevel Models + Informative Priors for decision-making.	
How should we evaluate a single small study?	

I think we can meaningfully apply this criterion even to 'pure' research questions where it is unclear how the research would ever be applied, specifically. We know a great deal about epistemology and scientific methodology and what practices tend to lead to reliable knowledge. (When people argue in favor of pure research because of its history of spinoffs like cryptography from number theory, that very argument implies that the spinoffs aren't *that* unpredictable & is a successful pragmatic defense! The fact that our evolved curiosity can be useful is surely no accident.)

For example, even without a specific purpose in mind for some research, we can see why forging fraudulent data is the worst possible criticism: because there is no decision whatsoever which is made better by using faked data. As Feynman put it: "For a successful technology, reality must take precedence over public relations, for nature cannot be fooled." Many assumptions or shortcuts will work in some cases, but there is no case where fake data, which is uncorrelated with reality, works; even in the case where the fake data is scrupulously forged to exactly replicate the best understanding of reality⁶, it damages deci-

sion-making by overstating the amount of evidence, leading to overconfidence and underexploration.

Similarly, careless data collection and measurement error. Microbiologists couldn't know about CRISPR in advance, before it was discovered by comparing odd entries in DNA databases, and it's a good example of how pure research can lead to tremendous gains. But how could you discover *anything* from DNA databases if they are incomplete, full of mislabeled/contaminated samples, or the sequencing was done sloppily & the sequences largely random garbage? If you're studying 'cancer cells' and they are a mislabeled cell line & actually liver cells, how could that possibly add to knowledge about cancer?

Or consider the placebo effect. If you learned that a particular study's result was driven entirely by a placebo effect and that using blinding would yield a null, I can safely predict that—regardless of field or topic or anything else—you will almost always be badly disappointed. If a study measures just a placebo effect (specifically, demand or expectancy effects), this is damning, because the placebo effect is already known to be universally applicable (so showing that it happened again is not interesting) through a narrow psychological causal mechanism which fades out over time & doesn't affect hard endpoints (like mortality), while it doesn't affect the countless causal mechanisms which placebo-biased studies *seem* to be manipulating (and whose manipulation would in fact be useful both immediately and for building theories). If, say, heart stents don't reduce actual chest pain except through the placebo effect, why would we want to use them? There are *some* exceptions where we would be indifferent after learning a result was just a placebo effect (chronic pain treatment? mild influenza?), but not many.

How about non-replicability? The simplest explanation for the Replicability Crisis in psychology is that most of the results aren't real and were random noise, *p*-hacked into publications. The most charitable interpretation offered by apologists is that the effects *were* real, but they are simply either small or so highly context-dependent on the ex-

act details (the precise location, color of the paper, experimenter, etc) to the point where even collaborating with the original researchers is not guaranteed to replicate an effect. Again, this presents a trilemma which is particularly damaging from a decision-theory point of view:

1. either the results aren't real (and are useless for decision-making),
2. they are much smaller than reported (and thus much less useful for any kind of application or theory-building),
3. or they are so fragile and in any future context almost as likely to be some other effect, in even the opposite direction, that their average effect is effectively zero (and thus useless).

Decisions precede beliefs. Our ontology and our epistemology flows from our decision theory, not vice-versa. This may appear to be logically backwards, but that is the situation we are in, as evolved embodied beings thinking & acting under uncertainty: like Otto Neurath's sailors on a raft—there is nowhere we can 'step aside' and construct all belief and knowledge up from scratch and logical metaphysics, instead, we examine and repair our raft as we stand on it, piece by piece. The naturalistic answer to the skeptic (like Plantinga) is that our beliefs are not unreliable because they are empirical or evolved or ultimately temporally begin in trial-and-error but they are reliable *because* they have been gradually evolved to pragmatically be correct for decision-making, and due to the constraints of evolution, developed reliable knowledge of the world and methods of science. (An example of reversing the flow would be the Deutsch-Wallace attempt to found the Born rule in MWI quantum mechanics on decision theory; earlier, statisticians such as Student, Frank P. Ramsey, Abraham Wald, Leonard J. Savage, Howard Raiffa & Robert Schlaifer etc showed that much of statistics could be grounded in decision-making instead of vice-versa, demonstrated by the subjective probability school and devices like the Dutch book enforcing coherency.)

DECISION-THEORETIC CRITICISMS

“ *The threat of decision analysis is more powerful than its execution.*

Andrew Gelman, 2019

”

“ *A good rule of thumb might be, ‘If I added a zero to this number, would the sentence containing it mean something different to me?’ If the answer is ‘no’, maybe the number has no business being in the sentence in the first place.*

Randall Munroe

”

Revisiting some of the example criticisms with more of a decision-theoretic view:

- ✦ A critique of **assuming correlation=causation** is a good one, because correlation is usually not causation, and going from an implicit ~100% certainty that it is to a more realistic 25% or less, would change many decisions as that observation alone reduces the expected value by >75%, which is a big enough penalty to eliminate many appealing-sounding things.

Because causal effects are such a central topic, any methodological errors which affect inference of correlation rather than causation are important errors.

- ✦ A critique of **distributional assumptions** (such as observing that a variable isn’t so much normal as Student’s *t*-distributed) isn’t *usually* an important one, because the change in the posterior distribution of any key variable will be minimal, and could change only decisions which are on a knife’s-edge to begin with (and thus, of little value).

- ✧ There are exceptions here, and in some areas, this can be critical. Distribution-wise, using a normal instead of a log-normal is often minor since they are so similar in the bulk of their distribution... unless we are talking about their *tails*, like in an order statistics context (common in any kind of selection or extremes analysis, such as employment or athletics or media or natural disasters, or in leaky pipeline processes), where the more extreme points out on the tail are the important thing; in which case, using a normal will lead to wild underestimates of how far out those outliers will be, which could be of great practical importance
- ✧ On the other hand, treating a Likert scale as a cardinal variable is a statistical sin... but only a peccadillo that everyone commits because Likert scales are so often equivalent to a (more noisy) normally-distributed variable that a fully-correct transformation to an ordinal scale with a latent variable winds up being a great deal more work while not actually changing any conclusions and thus actions.⁷

Similarly, temporal autocorrelation is often not as big a deal as it's made out to be.

- ✧ **Sociological/psychological correlations with genetic confounds** are vulnerable to critique, because controlling for genetics routinely shrinks the correlation by a large fraction, often to zero, and thus eliminates most of the causal expectations.
- ✧ **Overfitting to a training set** and actually being similar or worse than the current SOTA is one of the more serious criticisms in machine learning, because having better performance is typically why anyone would want to use a method. (But of course—if the new method is intriguingly novel, or has some other practical advantage, it would be entirely reasonable for someone to say that the overfitting is a minor critique as far as *they* are concerned, be-

cause they want it for that other reason and some loss of performance is minor.)

- ✦ use of a **strawman null hypothesis**: in a medical context too, what matters is the cost-benefit of a new treatment compared to the best existing one, and not whether it happens to work better than nothing at all; the important thing is being more cost-effective than the default action, so people will choose the new treatment over the old, and if the net estimate is that it is probably slightly worse, why would they choose it?
- ✦ Interpreting a **failure to reject the null as proof of null**: often a problem. The logic of significance-testing, such as it is, mandates agnosticism any time the null has not been rejected, but so intuitive is Bayesian reasoning—absence of evidence is evidence of absence—that if a significance-test does not vindicate a hypothesis, we naturally interpret it as evidence *against* the hypothesis. Yet really, it might well be evidence *for* the hypothesis, and simply not *enough* evidence: so a reasonable person might conclude the opposite of what they would if they looked at the actual data.

I'm reminded of the *Toxoplasma gondii* studies which use small samples to estimate the correlation between it and something like accidents, and upon getting a point-estimate almost identical to other larger studies (ie. infection predicts bad things) which happens to not be statistically-significant due to their sample size, conclude that they have found evidence *against* there being a correlation. One should conclude the opposite! (One heuristic for interpreting results is to ask: “if I entered this result into a meta-analysis of all results, would it strengthen or weaken the meta-analytic result?”)

- ✦ **inefficient experiment design**, like using between-subject rather than within-subject, or not using identical twins for twin experiments, can be practically important: as Student pointed out in his discussion of the Lanarkshire Milk Experiment, among other

problems with it, the use of random children who weren't matched or blocked in any way meant that statistical power was unnecessarily low, and the Lanarkshire Milk Experiment could have been done with a sample size *97% smaller* had it been better designed, which would have yielded a major savings in expense (which could have paid for many more experiments).

- ✦ lack of **measurement invariance**: questions of 'measurement invariance' in IQ experiments may sound deeply recondite and like statistical quibbling, but they boil down to the question of whether the test gain is a gain on *intelligence* or if it can be accounted for by gains solely on a subtest tapping into some much more specialized skill like English vocabulary; a gain on intelligence is far more valuable than some test-specific improvement, and if it is the latter, the experiment has found *a* real causal effect but that effect is fool's gold.
- ✦ **conflating measured with latent variables**: And in discussions of measurement of latent variables, the question may hinge critically on the *use*.

One example would be correlating SAT scores with college GPA: are you an individual-differences psychologist, or a university admissions office? The former probably wants to correct for measurement error as much as possible to get at the underlying psychology; the latter is doing a job and has to work with the raw scores it has on the actual students who apply, not the hypothetical true scores it would like to have but never will.

Or suppose one compares a noisy IQ test to a high-quality personality test (without incorporating correction for the differing measurement error of each one), and finds that the latter is more predictive of some life outcome; does this mean 'personality is more important than intelligence' to that trait? Well, it depends on use. If one is making a theoretical argument about the latent variables, this is a serious fallacy and correcting for measurement error may completely reverse the conclusion and show the oppo-

site; but if one is researching screening (for industrial/organization psychology), then it is irrelevant which latent variables is a better predictor, because the tests are what they are—unless, on the gripping hand, one is considering introducing a better more expensive IQ test, in which case the latent variables *are* important after all because, depending on how important the latent variables are (rather than the crude measured variables), the potential improvement from a better measurement may be enough to justify the better test...

Or consider heritability estimates, like SNP heritability estimates from GCTA. A GCTA estimate of, say, 25% for a trait measurement can be interpreted as an upper bound on a GWAS for the same measurement; this is useful to know, but it's not the same thing as an upper bound on a GWAS, or 'genetic influence' in some sense, of the true, latent, measured-without-error *variable*. Most such GCTAs use measurements with a great deal of measurement error, and if you correct for measurement error, the true GCTA could be much higher—for example, IQ GCTAs are typically ~25%, but most datasets trade quality for quantity and use poor IQ tests, and correcting that, the true GCTA is closer to 50%, which is quite different. Which number is 'right'? Well, if you are merely trying to understand how good a GWAS based on that particular dataset of measurements can be and what your statistical power is like, the former is the right interpretation, as it establishes your upper bound and you will need better methods or measurements to go beyond it; but if you are trying to make claims about a trait *per se* (as so many people do!), the latent variable is the relevant thing, and talking only about the measured variable is highly misleading and can result in totally mistaken conclusions (especially when comparing across datasets with different measurement errors).

- ✦ **Lack of blinding** poses a similar problem: its absence means that the effect being estimated is not necessarily the one we want to

estimate—but this is context-dependent. A psychology study typically uses measures where some degree of effort or control is possible, and the effects of research interest are typically so small (like dual n -back's supposed IQ improvement of a few points) that they can be inflated by a small amount of trying harder; on the other hand, a medical experiment of a cancer drug measuring all-cause mortality, if it works, can produce a dramatic difference in survival rates, cancer doesn't care whether a patient is optimistic, and it is difficult for the researchers to subtly skew the collected data like *all*-cause mortality (because a patient is either dead or not).

This definition is not a panacea since often it may not be clear what decisions are downstream, much less how much a criticism could quantitatively affect it. But it provides a clear starting point for understanding which ones are, or should be, important (meta-analyses being particularly useful for nailing down things like average effect size bias due to a particular flaw), and which ones are dubious or quibbling and are signs that you are stretching to come up with any criticisms; if you can't explain at least somewhat plausibly how a criticism (or a combination of criticisms) could lead to diametrically opposite conclusions or actions, perhaps they are best left out.

EXTERNAL LINKS

- ✦ [End-to-end principle](#)
- ✦ ["How We Know What Not To Think"](#), Phillips...2019
- ✦ ["What is Wrong with Our Thoughts? A Neo-Positivist Credo"](#), Stove₁₉₉₁
- ✦ ["The changing structure of American innovation: Some cautionary remarks for economic growth"](#), Arora...2020

- ✦ The uncanny valley of expertise: Seeing Like a State, Medical students' disease, "Why a little knowledge is a dangerous thing"
- ✦ "The causal foundations of applied probability and statistics", Greenland₂₀₂₀
- ✦ "There is still only one test"
- ⊕ **Discussion:** HN: 1, 2, 3

APPENDIX

TEACHING STATISTICS

“
“Would you tell me, please, which way I ought to go from here?”
“That depends a good deal on where you want to get to,” said the
Cat. “I don’t much care where—” said Alice.
“Then it doesn’t matter which way you go,” said the Cat.
*“—so long as I get **somewhere**”, Alice added as an explanation.*
“Oh, you’re sure to do that”, said the Cat, “if you only walk long
enough.”

Alice in Wonderland, Chapter 6

”

If decision theory is the end-all be-all, why is it so easy to take Statistics 101 or read a statistics textbook and come away with the attitude that statistics is nothing but a bag of tricks applied at the whim of the analyst, following rules written down nowhere, inscrutable to the uninitiated, who can only listen in bafflement to this or that piped piper of probabilities? (One uses a *t*-test unless one uses a Wilcoxon test, but of course, sometimes the *p*-value must be multiple-corrected, except when it’s fine not to, because you were using it as part of the main analysis or a component of a procedure like an ANOVA—not to be confused with an ANCOVA, MANCOVA, or linear model, which might really

be a generalized linear model, with clustered standard errors as relevant...)

One issue is that the field greatly dislikes presenting it in any of the unifications which are available. Because those paradigms are not universally accepted, the attitude seems to be that *no* paradigm should be taught; however, to refuse to make a choice is itself a choice, and what gets taught is the paradigm of statistics-as-grab-bag. As often taught or discussed, statistics is treated as a bag of tricks and *p*-values and problem-specific algorithms. But there are paradigms one *could* teach.

One simple desirable change would be to drop the whole menagerie of ‘tests’ and recast them as small variations on a linear model, of which some are simply common enough to be named as shortcuts—which is what they really are, and explains the underlying logic.

Another desirable change would be to make clear the grounding of statistics in *decision theory*. Around the 1940s, led by Abraham Wald and drawing on Fisher & Student, there was a huge paradigm shift towards the decision-theoretic interpretation of statistics, where all these Fisherian gizmos can be understood, justified, and criticized as being about minimizing loss given specific loss functions. Why do sufficient statistics work the way they do—what purposes or ‘statistics’ are they sufficient *for*? So, the mean is a good way to estimate your parameter (rather than the mode or median or a bazillion other univariate statistics one could invent) not because that particular function was handed down at Mount Sinai but because it does a good job of minimizing your loss under such-and-such conditions like having a squared error loss (because bigger errors hurt you much more); and if those conditions do not hold, *that* is why the, say, median is better, and you can say precisely how much better and when you’d go back to the mean (as opposed to rules of thumbs about standard deviations or arbitrary *p*-value thresholds testing normality).

Many issues in meta-science are much more transparent if you simply ask how they would affect decision-making (see the rest of this essay).

A third way to improve the motley menagerie that is the usual statistics education is Bayesianism. Bayesianism means you can just ‘turn the crank’ on many problems: define a model, your priors, and turn the MCMC crank, without all the fancy problem-specific derivations and special-cases. Instead of all these mysterious distributions and formulas and tests and likelihoods dropping out of the sky, you understand that you are just setting up equations (or even just writing a program) which reflect how you think something works in a sufficiently formalized way that you can run data through it and see how the prior updates into the posterior. The distributions & likelihoods then do not drop out of the sky but are pragmatic choices: what particular bits of mathematics are implemented in your MCMC library, and which match up well with how you think the problem works, without being too confusing or hard to work with or computationally-inefficient?

And causal modeling is a fourth good example of a paradigm that unifies education: there is an endless zoo of biases and problems in fields like epidemiology which look like a mess of special cases you just have to memorize, but they all reduce to straightforward issues if you draw out a DAG of a causal graph of how things might work (eg. most of these biases are just a collider bias, but on some specific variable & given a fancy name).

What happens in the absence of explicit use of these paradigms is an implicit use of them. Much of the ‘experience’ that statisticians or analysts rely on when they apply the bag of tricks is actually a hidden theory learned from experience & osmosis, used to reach the correct results while ostensibly using the bag of tricks: the analyst knows he ought to use a median *here* because he has a vaguely defined loss in mind for the downstream experiment, and he knows the data sometimes throws outliers which screwed up experiments in the past so the mean is a bad choice and he ought to use ‘robust statistics’; or he knows

from experience that most of the variables are irrelevant so it'd be good to get shrinkage by sleight of hand by picking a lasso regression instead of a regular OLS regression and if anyone asks, talk vaguely about 'regularization'; or he has a particular causal model of how enrollment in a group is a collider so he knows to ask about "Simpson's paradox". Thus, in the hands of an expert, the bag of tricks works out, even as the neophyte is mystified and wonders how the expert knew to pull this or that trick out of, seemingly, their nether regions.

Teachers don't like this because they don't want to defend the philosophies of things like Bayesianism, often aren't trained in them in the first place, and because teaching them is simultaneously too easy (the concepts are universal, straightforward, and can be one-liners) and too hard (reducing them to practice and actually computing anything—it's easy to write down Bayes's formula, not so easy to actually compute a real posterior, much less maximize over a decision tree).

There's a lot of criticisms that can be made of each paradigm, of course—none of them are universally assented to, to say the least—but I think it would generally be better to teach people in those principled approaches, and then later critique them, than to teach people in an entirely unprincipled fashion.



- 1 There's two main categories I know of, reporting checklists and quality-evaluation checklists (in addition to the guidelines/recommendations published by professional groups like the APA's manual based apparently on JARS or AERA's standards).

Some reporting checklists:

- ✦ STROBE (checklists; justification) for cohort, case-control, and cross-sectional studies

- ✦ TREND (checklist; justification) for non-randomized experiments
- ✦ CONSORT (checklist; justification) for RCTs
- ✦ QUOROM/PRISMA (checklist; justification) for reviews/meta-analyses

Some quality-evaluation scales:

- ✦ Jadad scale
- ✦ NOS (scale; manual)
- ✦ CEBM



2

As pointed out by Jackson in a review of a similar book, the arguments used against heritability or IQ exemplified bad research critiques by making the perfect the enemy of better & selectively applying demands for rigor:

There is no question that here, as in many areas that depend on field studies, precise control of extraneous variables is less than perfect. For example, in studies of separated twins, the investigator must concede that the ideal of random assignment of twin pairs to separated foster homes is not likely to be fully achieved, that it will be difficult to find comparison or control groups perfectly matched on all variables, and so on. Short of abandoning field data in social science entirely, there is no alternative but to employ a variational approach, seeking to weigh admittedly fallible data to identify support for hypotheses by the preponderance of evidence. Most who have done this find support for the heritability of IQ. Kamin instead sees only flaws in the evidence...As the data stand, had the author been equally zealous in evaluating the null hypothesis that

such treatments make no difference he would have been hard pressed to fail to reject it.



3 Non-replication of a result puts the original result in an awkward trilemma: either the original result was spurious (the most *a priori* likely case), the non-replicator got it wrong or was unlucky (difficult since most are well-powered and following the original, so—one man's modus ponens is another man's modus tollens—it would be easier to argue the *original* result was unlucky), or the research claim is so fragile and context-specific that non-replication is just 'heterogeneity' (but then why should anyone believe the result in any substantive way, or act on it, if it's a coin-flip whether it even exists anywhere else?).



4 As quoted in Pearson₁₉₃₉.



5 It is interesting to note that the medieval origins of 'probability' were themselves inherently decision-based as focused on the question of what it was *moral* to believe & act upon, and the mathematical roots of probability theory were also pragmatic, based on gambling. Laplace, of course, took a similar perspective in his early Bayesianism (eg. Laplace on witness testimony or estimating the mass of Saturn). It was later statistical thinkers like Boole or Fisher who tried to expunge pragmatic interpretations in favor of purer definitions like limiting frequencies.



6 Which is usually not the case, and why fakers like Diederik Stapel can be detected by looking for 'too good to be true' sets of results, over-rounding or overly-smooth numbers, or sometimes just not even arithmetically correct! (Stroebe₂₀₁₉ notes, incidentally, that

while Stapel's effects individually *looked* plausible to other researchers, collectively they were too large and this led to anomalies when meta-analyzed—European, but not American, priming studies had an oddly larger average effect.) ↩

- 7 As formally incorrect as it may be, whenever I have done the work to treat ordinal variables correctly, it has typically merely tweaked the coefficients & standard error, and not actually *changed* anything. Knowing this, it would be dishonest of me to criticize any study which does likewise unless I have some good reason (like having reanalyzed the data and—for once—gotten a major change in results). ↩

BACKLINKS

- ✦ [Fake Journal Club: Teaching Critical Reading \(full context\):](#)

So, how do you learn good research criticisms If the end of the path of learning active reading is deep domain expertise, to the point of being able to know which papers to do forensic statistics on to detect fraud, what is the start of the path? What is the first and smallest possible step one can take?

- ✦ [Subscrips For Citations \(full context\):](#)

This is also why accounting systems are never feature-complete, and always keep sprouting new ad hoc features & flags & reports, eventually embodying Greenspun's tenth rule with their own, often ad hoc and extremely low-quality, built-in programming language. Accounting is not a simple problem of just tracking some movements of a few kinds of assets like 'dollars'

or ‘cash’ in and out of some ‘accounts’. Accounting is as complex as the world, because we always want more informative models of our financial position, and have more domain-specific knowledge to encode, and people will always differ about what they think the causal models are or what the future will be like, and this will cause conflicts over ‘accounting’. Accounting, like statistics is ultimately for *decision-making*: a failure to make this explicit will lead to many confusions & illusions in understanding the purpose of accounting or things like ‘GAAP’. Accounting systems should never delude themselves about this intrinsic complexity by pretending it doesn’t exist and trying to foist it onto the users, but accept that they have to handle it and provide real solutions, like being built with a properly supported language like a Lisp or Python, instead of fighting a rearguard action by adding on features ad hoc as users can ram them through to meet their specific needs.

✦ Does Mouse Utopia Exist? (full context):

A particularly serious criticism as I doubt that the urban planners or demographers or Democratic politicians who took an interest in Mouse Utopia would be as interested if the causal mechanism turned out to be “urban densities increase STDs or genetic mutations to the point of collapse”. And if it turned out that Mouse Utopia replicated in mice but never humans (early attempts to correlate population density with social decay in humans apparently did not do well, incidentally), I also doubt if most people citing it, aside from a few zoologists, ethologists, or mouse breeders, would be doing so.

✦ Dog Cloning For Special Forces: Breed All You Can Breed (full context):

However, ranking for selection is easier than prediction of all datapoints: only the ordering matters, and only the ordering in a particular region (near the threshold) matters. When considered in a real-world context, such predictive improvements do not need to be all that large (a point long made by psychometricians & industrial psychologists eg. Taylor & Russell¹⁹³⁹/Schmidt...¹⁹⁷⁹/Lubinski & Humphreys¹⁹⁹⁶); counterintuitively, a score or test which correlates, say, $r = 0.10$ with an outcome, which in many areas of science would be dismissed as a trivial correlation of no interest, can be quite useful in screening & should not be dismissed as 'small'—and the rarer the outcome, the larger the benefit.² In the case of dog cloning, our 'score' is the extent to which a donor's performance predicts the performance of its clones, through their shared genes.

✦ Why Tool AIs Want to Be Agent AIs (full context):

There is no hard Cartesian boundary between an algorithm & its environment such that control of the environment is irrelevant to the algorithm and vice-versa and its computation can be carried out without regard to the environment—there are simply many layers between the core of the algorithm and the furthest part of the environment, and the more layers that the algorithm can model & control, the more it can do. Consider Google Maps/Waze⁷. On the surface they are 'merely' Tool AIs which produce lists of possible routes which would optimize certain requirements; but the entire point of such Tool AIs—and all large-scale Tool AIs and research in general—is that countless drivers will *act on* them (what's the point of getting driving directions if you don't then drive?), and this will greatly change traffic patterns as drivers become appendages of the 'Tool' AI, potentially making driving in an area much worse by their errors or myopic per-driver optimization causing Braess's

paradox (and far from being a theoretical curiosity, GPS, Google Maps, and Waze are regularly accused of that in many places, especially Los Angeles).

✦ Embryo Selection For Intelligence (full context):

[backlink context]

✦ Resorting Media Ratings (full context):

So, I think our target distribution ought to maximize *usefulness*: it is just a summary of an unknowable underlying distribution, which we summarize pragmatically as ratings, and so statistics like ratings can only be right or wrong as defined by their intended use. On a rating website, we are not interested in making fine distinctions among mediocre or trash. We are looking for interesting new candidates to consider, we are looking for the best. A skew maximizes the provided information to the reader in the region of interest of likely recommendations. So our distribution ought to throw most of our ratings into an uninformative ‘meh’ bucket, and spend more time on the right-tail extreme: do we think a given work an all-time masterpiece, or exceptional, or merely good?

✦ Why Correlation Usually \neq Causation (full context):

The Replication Crisis: a shockingly large fraction of psychological research and other fields is simple random noise which cannot be replicated, and due to p-hacking, low statistical power, publication bias, and other sources of systematic error. The RC can manufacture arbitrarily many spurious results by data-

mining, and this alone ensures a high error rate: random noise is good for nothing

✦ The Replication Crisis: Flaws in Mainstream Science (full context):

None of these systematic problems should be considered minor or methodological quibbling or foolish idealism they are systematic biases and as such, they force an upper bound on how accurate a corpus of studies can be even if there were thousands upon thousands of studies, because the total error in the results is made up of random error and systematic error, but while random error shrinks as more studies are done, systematic error remains the same.