Error Correction, Severity, and Truth: What's Simplicity Got to Do With it?

Deborah Mayo

CMU: June 24, 2012

I always learn a lot from the challenge to see if I have anything to say about a topic I haven't written about—simplicity. (I have often said, on the other hand, that the account I favor is messy and complex and "it will not appeal to neatniks".)

I have thought a lot about some of the issues to which simplicity is often linked: induction, truth, reliability, parameter adjustment, selection effects, severe testing

A general simplicity slogan goes something like this:

If a simpler (method, model, theory) will suffice, then go with it?

(usually there's an addition, "all else being equal" **—**but that's a very big qualification!)

Suffice for what?

When I think about what I regard as key learning goals, it seems that simplicity criteria at most correlate with what's actually doing the work in knowledge promoting..
But, it seems to me always *second-hand* to what's responsible for:

> the goals of ensuring and appraising how well-tested claims are, how well methods control and avoid error.

It is like the "second hand emotion" in the lyrics to which my title alludes:
("*What's simplicity got to do with it?*

Tina Turner: "Oh what's love got to do, got to do with it
What's love but a *second hand emotion")*

I will discuss a number of different ways that simplicity recommendations may go hand in hand with, but also be in tension with, goals of truth, learning, and well-testedness.

I. Simplicity in Appraising Statistical Method and Principle
II. Simplicity, Severity, and Error Correction

I. **Appealing to Simplicity in Appraising Method and Statistical Principle**

Go with the simplest account, obey the simplest principles, test intuitions with the simplest examples

A. Bayesian inference is often promoted as the simplest, coherent account:

If the goal is to compute the degree of *evidential relationship* between given evidence statements, e, and a hypothesis H, Bayesians say, look to conditional probability or *Bayes's Theorem*:

$$P(H|e) = P(e|H)P(H)/P(e)$$

where $P(e) = P(e|H)P(H) + P(e|\text{not-}H)\, P(\text{not-}H)$.

"Any decision that depends on the data that is being used in making the inference only requires from the data the posterior distribution. Consequently the problem of inference is effectively solved by stating the posterior distribution". (Lindley 1971, 436)

But to suffice as an adequate account for inference, it

- must be *ascertainable* (must be able to apply it with the kind of information we tend to actually have in inquiry)
- it should be *relevant* and *applicable* to the tasks required of the inference tools

(Wesley Salmon's criteria)

These criteria are in tension with what is recommended by this kind of single unified evidential-relation account.

Computing P(H|e), the *posterior probability*, requires a probability assignment to all of the members of "not-H" (Bayesian *catchall hypothesis*)

Major source of difficulty: how to obtain and interpret these *prior probabilities*.

(a) If analytic and a priori, relevance for predicting and learning about empirical phenomena is problematic

(b) If they measure subjective degrees of belief, their relevance for giving objective guarantees of reliable inference is unclear.

More appeals to simplicity arise in obtaining priors

In statistics, (a) is analogous to “objective” Bayesianism (e.g., Jeffreys);

(b) to subjective Bayesianism

Even if the overall Bayesian logic is "simple,"Bayesians themselves (some of them) admit:

"to elicit all features of a subjective prior $\pi(\theta)$, one must infinitely accurately specify a (typically) infinite number of things. In practice, only a modest number of (never fully accurate) subjective elicitations are possible." (J. Berger 2006, p. 397)

One way of turning elicitations into full priors is to use conjugate priors, but as these are model-dependent (or at least likelihood-dependent) the subjective Bayesian following this "prior completion" strategy would be constructing different priors for the same θ, clearly incoherent. (J. Berger, 2006)

Some Bayesians are even saying aloud…
"Indeed, I cannot remember ever seeing a non-trivial Bayesian analysis which actually proceeded according to the usual Bayes formalism." (Goldstein, 2006)

[While "Bayesian inference is commonly associated with inductive reasoning and the idea that a model … can never be directly falsified by a significance test. [my goal] is to break these associations, which I think are incorrect and have been detrimental to statistical practice." (Gelman, 2011, p. 1)

B. Error probability methods (frequentist, sampling theory)

- Designed to reach statistical conclusions without invoking prior probabilities in hypotheses,
- Probability is used to quantify how *frequently* methods are capable of discriminating between alternative hypotheses and how *reliably* they facilitate the detection of error.

Error frequencies or *error probabilities* (e.g., significance levels, confidence levels).

These methods are often criticized as giving us a hodge-podge of tools and criteria: piecemeal, messy

But they may be unified in nifty ways for the goals of error probability control and piecemeal learning

In this account probabilities apply to events, and to rules or methods of inference:

Two construals within frequentist error statistics:

(i) *behavioristic*: to ensure low long run control of error *

(ii) *evidential*: to control and evaluate well-testedness of claims

Three reasons error statistical methods are complex

(1) Because the typical statistical hypothesis is rarely the final substantive inference, and (2) raw data has to be worked over to get them in shape for inference

Scientific questions	Statistical hypotheses	Statistical data	Actual data

- When it comes to piecemeal probes about parameters, directions of effect, observed correlations) the canonical null hypotheses are just the ticket! (Cox's taxonomy of several distinct types of null hypotheses)
- Scientific inquiry needs to be open-ended (unlike what probability theory requires)
- In setting sail to find things out, we do not have an exhaustive set of rival substantive hypotheses;
- Much less do we have what's needed by "full-dress Bayesians", as I.J. Good called them: an assignment of utilities or loss functions for decision making
- "Much like ready-to-wear [versus designer] clothes, these 'off the shelf' methods do not require collecting vast resources before you can get going with them" (1996, p. 100).

“It’s Complicated”

Aside from the fact that they are expected to perform different tasks at different stages of inquiry, the third reason error statistical methods are complicated is what at the heart of what makes them *error statistical.*

- Whatever base characterization of “fit” you like, we always want to know: what’s the probability you’d get so good a fit (between data $\mathbf{x}_0$ and H) if H is false
- If H is being declared the best of the lot (according to your favorite criterion) we always want to know if that’s something easy or difficult to achieve even if H is specifiably false

(Pearson’s steps: sample space, hypotheses, distance statistic, sampling distribution)

- For example, significance tests operate by computing the probability of different values of $d(\mathbf{X})$ —some distance measure between data and a test or null hypothesis $\boldsymbol{H_0}$,
- We can calculate P$(d(\mathbf{X}) > d(\mathbf{x}_0); H_0)$ – the p-value of the result—if it's very small infer evidence against the null (or evidence of a genuine discrepancy).
- The probability distribution of $d(\mathbf{X})$ *is* its *sampling distribution*
- It lets us calculate the probability of inferring evidence for H erroneously—an *error probability*.

As a result, aspects of the data and hypotheses generation may have to be taken account of: they may alter the error probabilities and thereby the probativeness of the test.

This introduces complications…

But that is the key to controlling and assess error probabilities.

(In my own revision of error statistical methods, I insist on an assessment that is relative to the actual outcome, as opposed to standard predesignated error probabilities, but existing methods can serve this role.)

C. Simplicity and Freedom (vs. control of error probabilities)

That error probabilistic properties may alter the construal of results gets a formal rendering: we violate the (strong) likelihood principle LP. (likelihoods aren’t enough)

Among aspects of the data generation that could alter error probabilities are stopping rules. By contrast

> “The irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design that had been lost by classical emphasis on significance levels (in the sense of Neyman and Pearson”. (“Savage Forum” 1963, p. 239)

We are prepared to exchange simplicity and freedom for controlling error probabilities

One way to illustrate the violation of the LP in error statistics is via the *"Optional Stopping Effect"*.

We have a random sample from a Normal distribution with mean μ and standard deviation σ, i.e.

$X_i \sim N(\mu,\sigma)$ and we test H_0: $\mu=0$, vs. H_1: $\mu\neq 0$.
stopping rule:

Keep sampling until H is rejected at the .05 level

(i.e., keep sampling until $|\bar{X}| \geq 1.96\ \sigma/\sqrt{n}$).

The rule is guaranteed to stop, it is assured of rejecting the null even if true.

More generally, actual significance level differs from, and will be greater than .05.

Violates the *weak repeated sampling rule* (Cox and Hinkley, 1974)

There are many equivalent ways to get this kind of violation of error probabilities (hinting for significance, selection effects)

It need not have anything to do with stopping rules, it can result from data-dependent selection of hypotheses for testing, or rejecting a null so long as any better fitting alternative exists

It is sometimes said that in requiring the actual type 1 error probability be small (i.e., requiring very small p-values) before the null is rejected in favor of the alternative), we are appealing to the simpler hypothesis (the null)

In a sense it is simpler but the actual rationale: if moderate p-values are taken as evidence of a genuine discrepancy from the null, then it will make it too easy to erroneously infer a real effect

We can't even assess whether an observed agreement (between data and a hypothesis) really is big or small without it the sampling distribution.

"<u>If we accept the criterion suggested by the method of likelihood it is still necessary to determine its sampling distribution in order to control the error involved in rejecting a true hypothesis,</u> because a knowledge of L [the likelihood ratio] alone is not adequate to insure control of this error. (Pearson and Neyman, 1967, 106).

D. Should we trust our intuitions in simple cases?

It is often noted that if the test is restricted to a comparative test, limited to simple or *point against point* hypotheses, then there is an upper error bound (so the problem with optional stopping is avoided).

(Savage switches to such cases, “Savage Forum” 1962)

But that’s a very different, very artificial example.

So, to the question, should we trust our intuitions about general principles from simple cases? The answer is no (we should look for exceptions)

It was the case of the complex (i.e., compound) alternative that led statistician George Barnard to reject the (strong) LP (surprising Savage).

The trouble with the so-called "Law" of Likelihood: (i.e., $\mathbf{x}_0$ support hypotheses H_1 more than H_2 if, $P(\mathbf{x}_0;H_1) > P(\mathbf{x}_0;H_2)$), Barnard notes, "there always is such a rival hypothesis: That things just had to turn out the way they actually did" .

Since such a maximally likelihood alternative H_2 can always be constructed, H_1 may always be found less well supported, even if H_1 is true—no error control.

Hacking soon rejected the likelihood approach on such grounds, likelihoodist accounts are advocated by others.

I turn now to more familiar appeals to simplicity (not for methods or principles, but for inference to hypotheses, models, theories)

II. Simplicity, Severity, Error Correction

A. Underdetermination and Simplicity

Clearly a big rationale for the appeal to simplicity is the supposition that we are otherwise stuck with terrible underdetermination

"But since there will always be an infinite number of theories which yield the same data with the same degree of inductive probability—but which make different predictions….without the criterion of simplicity we can make no step beyond the observable data. Without this all-important *a priori* criterion, we would be utterly lost." Swinburne ("Simplicity as Evidence of Truth" 1997, 15)

The problem might be seen as what more do we need to avoid underdetermination

(i) $\mathbf{x}_0$ agrees with or "fits" H

(ii) ______

Explanatory power, novelty, simplicity, well-testedness, *severity*

Popper

Mere fits are “too cheap to be worth having” (Popper)

> “In opposition to [the] inductivist attitude, I assert that C(H,**x**) must not be interpreted as the degree of corroboration of H by **x**, unless **x** reports the results of our sincere efforts to overthrow H. The requirement of sincerity cannot be formalized—no more than the inductivist requirement that e must represent our total observational knowledge. (Popper 1959, p. 418.)
>
> Observations or experiments can be accepted as supporting a theory (or a hypothesis, or a scientific assertion) *only if these observations or experiments are severe tests of the theory*—or in other words, only if they result from serious attempts to refute the theory, ….” (Popper, 1994, p. 89)

It is no wonder Popper is often compared to error statisticians (Fisher, and/or Neyman and Pearson)

True Popper was never able show “ qua pure deductivist…[that]we should expect the theory to fail if it is not true” (Guenbaum, 198, 130)

The best tested so far need not be well tested; his methods gave no way to assess error probabilities.

Complexity and Inseverity

The primary Popperian goal was always severity and avoidance of ad hoc strategems that would lower the testability of hypotheses

"From my point of view, a system must be described as complex in the highest degree if, …one holds fast to it as a system established forever which one is determined to rescue, whenever it is in danger, by the introduction of auxiliary hypotheses. For the degree of falsifiability of a system thus protected is equal to zero." (Popper LSD, 331).

Characteristic of pseudoscience.

Note, it's the system that is complex (I would say procedure or method).

Popper typically, misleadingly, suggests it is the hypothesis or theory that should be testable or simple.
The fact that he was unable to implement the idea using logical probability does not stop us from using contemporary statistical tools to do so.

Severity Principle

The Popperian intuition is right-headed:

If a procedure had little or no ability to find flaws in *H*,
then finding none scarcely counts in *H*'s favor.

Can put in terms of having evidence…

Severity Principle (Weak)**:** Data $\mathbf{x}_0$ provides poor evidence for *H* if it results from a method or procedure that has little or no ability of finding flaws in *H*, even if *H* is false.

- As weak as this is, it is stronger than a mere falsificationist requirement: it may be logically possible to falsify a hypothesis, while the procedure may make it virtually impossible for such falsifying evidence to be obtained.

- Although one can get considerable mileage even stopping with this negative conception (as perhaps Popperians would), I hold the further, positive conception:

Severity Principle (Full)**:** Data $\mathbf{x}_0$ provide a good indication of or evidence for hypothesis *H* (just) to the extent that test *T* severely passes *H* with $\mathbf{x}_0$.

I talk about SEV a lot elsewhere, and cannot get into qualifications here.

Severity has three arguments: a test, data, and an inference or a claim.

> ‘The severity with which H passes test T with outcome $\mathbf{x}_0$’ may be abbreviated by: SEV(Test T, outcome $\mathbf{x}_0$, claim H).

To say “H is severely tested” is an abbreviation of H has *passed* the severe or stringent probe, not, for example merely that H was subjected to one (corroboration)

- This contrasts with a common tendency to speak of “a severe test” divorced from the specific inference — leads to fallacies we need to avoid.

- A test may be made so sensitive (or powerful) that discrepancies from a hypothesis H are inferred too readily. (fallacy of rejection)

However, the severity associated with such an inference is *decreased*, the more sensitive the test (not the reverse).
One analogously avoids “fallacies of acceptance”, I’ll illustrate with a fanciful example.

My weight

If no change in weight registers on any of a series of well-calibrated and stable scales, both before leaving and upon my return from London, even though, say, they easily detect a difference when I lift a .1-pound potato, then we argue that the data warrant inferring that my weight gain is negligible within the limits of the sensitivity of the scales.

> *H*: my weight gain is no greater than δ, where $\delta > 0$ is an amount easily detected by these scales.

H, we would say, has passed a ***severe test*** were I to have gained δ pounds or more (i.e., were *H* false), then this method would almost certainly have detected this.

No Rigging!

Perhaps underdeterminationists would say I could insist all the scales are wrong—they work fine with weighing vegetables, etc.
(Cartesian demon of scales)

Rigged alternative *H**

H: H* is false but all data will be as if it is true.

All experiments systematically mask the falsity of H

(Gellerized hypothesis)

Are H and H empirically equivalent? If so, they are not testably equivalent on the severity account*

For any hypothesis H, one can always adduce a rigged H* (even if H is true and has passed highly severe tests!)

Were we to deny $\mathbf{x}_0$ is evidence for *H* because of the possibility of rigging, we would be prevented from correctly finding out about weight or whatever…

If the scales work reliably on test objects with known weight, what sort of *extraordinary circumstance* could cause them all to go astray just when we do not know the weight of the test object (can the scales read my mind? (C.S. Peirce)

It's simpler to assume the scales that work on my potato also work with unknown weights (in the intended range) but that's not why it is warranted.

It is the learning goal that precludes such *rigging, conspiracies, gellerization — highly unreliable strategy.*

Granted, this is a special case where there is knowledge of the probative capacities of the instrument, and this figures importantly in this account for justifying inductive (evidence-transcending) inferences

Central strategy for checking assumptions using known procedures:

- to ensure errors ramify: if we were wrong, we would find systematic departures from the known weight (likewise with the use of known probability models, e.g., Bernouilli model with p = .5 and coin tossing)
- To move from highly inaccurate measurements to far more accurate ones

Not an appeal to the uniformity of nature, but "*that the supernal powers withhold their hands and let me alone, <u>and that no mysterious uniformity… interferes with the action of chance".</u>*

The associated warrant for ampliative inference beyond today's paper ….

Sometimes it feels as if simplicity is appealed to in order to save some (flawed) accounts of inference from themselves

An account that regards *H and H* empirically (predictively) equivalent?*

By contrast, two pieces of data that equally well fit a hypothesis, may differ greatly in their evidential value due to differences in the probativeness of the tests from which they arose.

Empirical learning is complex in the sense of *piecemeal*

One way to distinguish the pieces is by considering what error of inference is of concern

Formal error statistical tests provide tools to ensure errors will be correctly detected (i.e., signaled) with high probabilities—but in scientific contexts, their role will be not to assure low long run error rates (behavioristic) but to learn about the source of the given data set.

Within the piece: we are not distinguishing a hypothesis or theory from its rivals, experiment sets out to distinguish and rule out a specific erroneous interpretation of the data from *this* experiment.

An error can concern any aspect of a model or hypothesis in the series of models, i.e., any mistaken understandings of an aspect of the phenomenon in question *these are errors about real vs. spurious effects, causes, parameters, model assumptions, links from statistical to substantive, classification errors, etc.—*

I don’t distinguish theoretical/observational

There is a corresponding localization of what one is entitled to infer severely:

“*H* is false” refers to a specific error that the hypothesis *H* is *denying*.

For example, we still need to distinguish the inference from *rejecting a null of “0 effect”* from theories to explain the effect
(they are on “different levels”)

Much less does evidence for a “real effect” warrant realism (entity realism, or other).

Discussions of error correcting or self-correcting methods often confuse two interpretations of the ‘long-run’ metaphor:

(a) **Asymptotic error-correction (as n→∞):** *I have a sample of 100 and I consider accumulating more and more data* as n increases the inference or estimate about m approaches the true value of m

(b) **Error probabilities of a test:** *I have a sample of 100 and I consider hypothetical replications of the experiment*—each with samples of 100
(the relative frequency with which a sample mean differs by more than 2 standard deviations from the true mean is .05).
So, I can use the observed mean to estimate how far off the “correct” value is.

The error probability tells me about the procedure underlying the actual 100-fold sample, e.g., that there’s good evidence it was not merely fortuitous or due to chance.

Sampling distribution supplies the counterfactual needed: the value of employing a sampling distribution to represent statistically what it would be like were one or another assumption of the data generating mechanism violated:

In one-sided Normal sampling with known σ, for example, an upper .975 CI_u

H: $\mu \leq \bar{x}_0 + 1.96\sigma_{\mathbf{x}}$

(i.e., $\mu \leq \mathrm{CI}_u$)

passes severely because were this inference false, and the true mean $\mu > \mathrm{CI}_u$

Simplicity and Economy

The idea that a central aim of statistical method is to speed things up in this way is at the heart of the rationale of error statistical methods:

The concern we might say is with making good on the long run claims in the short run, within the usual amount of time for a given research project.

“It changes a fortuitous event which may take weeks or may take many decennia into an operation governed by intelligence, which will be finished within a month. (Peirce 7.78)

Giving good leave:

An important consideration Peirce gives under economy is “that it may give good leave” as the billiard-players say. If it fails to fit the facts, the test may be instructive about the next hypothesis. Even if we wanted to know if a quadratic equation holds between quantities, we would do well to test a linear model first “because the residuals will be more readily interpretative.”

The residuals, differences between observed and predicted values, may teach more about the next hypothesis to try.

Models must not merely fit, but be statistically adequate: Studying the residuals we can probe if the statistical adequacy of a model, the residuals are like white noise

Again, one senses that simplicity is appealed to in order to save an inadequate account from itself

"Error fixing" gambits in model validation.
Example: A statistically significant difference from a null that asserts independence in a linear regression model, might be taken as warranting one of many alternatives that could explain non-independence:

H_1 :the errors are correlated with their past, expressed as a lag between trials.

H_1 now "fits" the data all right, but since this is just one of many ways to account for the lack of independence, alternative H_1 passes with low severity.

This method has little if any chance of discerning other hypotheses that could also "explain" the violation of independence.

It is one thing to arrive at such an alternative based on the observed discrepancy with the requirement that it be subjected to further tests.

Severity and Informativeness

(be stringent but learn something)

Severity is not the only goal, it has to be coupled with informativeness, with finding things out

Many of the accounts discussed here start out assuming restricted domains and goals that I do not, e.g., a set of models that “fit” the data (to which we assign model selection), certain machine learning contexts with training samples and the like, empirical fit or prediction, suffices, etc.

As a philosopher of science, I’m always looking for a very general account of learning, finding things out;

I’m interested in how we set sail to obtain the kind of knowledge we do, and how we can obtain more of it.

Scientific knowledge and understanding, as I see it, goes beyond predicting events

Statistical method provides methods and analogues to methods for the most general kind of inductive inference

To one who thinks fitting the facts and predictive accuracy is what is mainly wanted in inference, it must seem mystifying that scientists are not especially satisfied with that alone, they always want to push the boundaries to learn something new, to rock the boat.

There may be a tension between simplicity and breaking out of paradigms

Wouldn't it be simpler not to challenge adequate predicting theories?

Maybe, but scientists rock the boat to find out more, because they want understanding that goes beyond predicting

One may be entirely agnostic on realism; models are approximate and idealized, that doesn't prevent getting a correct understanding using them

Why did researchers deliberately construct rivals if General Theory of Relativity (GTR) was predicting adequately (maybe it had a high posterior)?

Some say severity is too tough to satisfy, but they overlook the value of recognizing inseverity.

Our severe tester sets about exploring just *why* we are *not* allowed to say that GTR is severely probed as a whole—why has it inseverely passed based on given tests

- *How could it be a mistake to regard the existing evidence as good evidence for GTR*?

(even in the regions probed by solar system tests)

- Parameterized Post Newtonian (PPN) framework: a list of parameters that allows a systematic articulation of violations of, or alternatives to, what GTR says about specific gravity effects (they want to avoid being biased toward GTR)

- Set up largely as straw men with which to set firmer constraints on these parameters, check which portions of GTR has and have been well-tested (Earman 1992)

Each PPN parameter is set as a null hypothesis of a test.
For example, λ, the deflection of light parameter, measures “spacial curvature”;

The GTR value for the PPN parameter under test serves as the null hypothesis from which discrepancies are sought (usually set at 1).

H_0: $\lambda = \lambda_{\text{GTR}}$

By identifying the null with the prediction from GTR, any discrepancies are given a very good chance to be detected, so if no significant departure is found, this constitutes evidence for the GTR prediction with respect to the effect under test, i.e., λ.

The tests rule out GTR violations exceeding the bounds for which the test had very high probative ability
(infer upper bounds to possible violations)

(could equivalently be viewed as inferring a confidence interval estimate $\lambda = L \pm e$)

Simplicity and Parameter Adjustment
(Our conference organizer is keen for me to touch on this, and GTR offers a good case)

Deliberately constructing viable rivals theories did not preclude "fixing arbitrary parameters" to ensure rivals yield correct predictions with regard to the severely affirmed effects

For example, the addition of a scalar field in Brans-Dicke theory depended on an adjustable constant w:

The smaller its value the larger the effect of the scalar field and thus the bigger the difference with GTR, but as w gets larger the two became indistinguishable.

(An interesting difference would have been with a small w like 40; its latest lower bound is pushing 20,000!)

The value for λ is fixed in GTR, but constraining a rival like the B-D theory to fit the GTR prediction involves adjusting a parameter w:

Several Bayesians (e.g., Berger, Rosenkrantz) maintain that a theory that is free of adjustable parameters is "simpler" and therefore enjoys a higher prior probability (Jefferys and J. Berger 1992, 72; "Ockham's razor and Bayesian analysis")

Here they are explicitly referring to adjustments of gravity theories.

Others maintain the opposite

"On the Bayesian analysis," this countenancing of parameter fixing "is not surprising, since it is not at all clear that GTR deserves a higher prior than the constrained Brans and Dicke theory" (Earman, 1992, p. 115).

"why should the prior likelihood of the evidence depend upon whether it was used in constructing *T?";* Earman, 1992, p. 116), As I’ve argued elsewhere, there are many cases where data are used to arrive at and support parameters that result in the fitted claim passing with high severity

To correctly diagnose the differential merit, the severe testing approach instructs us to consider the particular inference and the ways it can be in error in relation to the corresponding test procedure.

In adjusting *w,* thereby constraining Brans–Dicke theory to fit the estimated w, what is being learned regarding the Brans–Dicke theory is *how large would w need to be* to agree with the estimated λ.

In this second case, inferences that pass with high severity are of the form “w must be at least 500.”
(~confidence interval estimate)

The questions, hence the possible errors, hence the severity differs.

But the data-dependent GTR alternatives play a second role; namely to show that GTR has *not* passed severely as a whole: that were a rival account of the mechanism of gravity correct, the existing tests would not have detected this.

This was the major contribution provided by the rivals articulated within the PPN framework (of viable rivals to GTR).

The constrained GTR rivals successfully show the existing tests did not rule out, with severity, alternative explanations for the λ effect given in the viable rivals.

Some view their role as estimating cosmological constants, thus estimating violations that would be expected in strong gravity domains.

Discovering new things: Nordvedt Effect η

But what I really want to emphasize is the kind of strategy that enables finding a new effect.

Discovering new things is creative, but it's not the miracle Popper makes it out to be

In the 1960s Nordvedt identified a testable difference, that Brans–Dicke theory would conflict with

In the 1960s Nordvedt discovered in the 1960s that B-D theory would conflict with GTR by predicting a violation of the Strong Equivalence Principle

(basically the Weak Equivalence Principle for massive self-gravitating bodies, e.g., stars and planets, black holes);

a new parameter to describe this effect, the Nordvedt effect, was introduced into the PPN framework, i.e., $\boldsymbol{\eta}$.

$\boldsymbol{\eta}$ would be 0 for GTR, so the null hypothesis tested is

H$_0$: $\boldsymbol{\eta} = 0$ as against non-0 for rivals.

Measurements of the round trip travel times between the earth and moon (between 1969 and 1975) enabled the existence of such an anomaly for GTR to be probed severely (actually, the measurements continue today).

Because the tests are highly sensitive, these measurements provided evidence that the Nordvedt effect is absent, set upper bounds to the possible violations

I talk about experimental GTR elsewhere….

Unification May Grow Out of Testing Constraints

- many of the parameters are functions of the others—an extremely valuable source for cross-checking and fortifying inferences
  (e.g., λ measures the same thing as the so-called time delay, and the Nordevdt parameter η gives estimates of several others.)

- we may arrive at a unification, but note that the impetus was simultaneously (if not mainly), getting more constrained tests to learn more

- Combined interval estimates, constrains the values of the parameters, enabling entire chunks of theories to be ruled out at a time (i.e., all theories that predict the values of the parameter outside the interval estimates).

Concluding remarks

In the first part of this paper I considered appeals to simplicity in appraising inductive statistical accounts and principles, and denied it was a good guide to avoid too-easy confirmations and fits

In the rest, I considered how a desire for simplicity grows out of the desire for constraints against too easy inductive inferences
A general simplicity slogan goes something like this:

If a simpler (method, model, theory) will suffice, then go with it?

(usually there's an addition, "all else being equal" —but this enables certain simplicity positions to be retained.

Any example I might point to where something else is really operative could be dismissed by saying it violates the "all things are equal" requirement.

But then of course the simplicity position is itself maximally complicated (using Popper's notion)

That appraisals of inferences are altered by the overall error probing capacities of tests complicates the account, but in so doing enables it to avoid having to resort to familiar appeals to simplicity of other accounts.

It’s an appeal to well testedness, which gets at what is really at issue, or so I argue
but severity provides a general desideratum for when selection effects need to be taken account of.

The severity intuition: we have good evidence that we are correct about a claim or hypothesis just to the extent that we have ruled out the ways we can be wrong in taking the claim or hypothesis to be true.

Far from wishing to justify enumerative induction from all observed A’s have been B’s to an inference that all or most A’s are B’s in a given population, such a rule would license inferences that had not passed severe tests—highly *unreliable* rule.

An induction following this pattern is warranted only when the inference has passed a severe test

The goal of correct understanding, and learning more is not simple ("it will not appeal to neatniks")

1. But the piecemeal account enjoys the benefits of applicability of "ready to wear and easy to check" methods
2. the goal of attaining a more comprehensive understanding of phenomena
3. the exploitation of multiple linkages to constrain, cross-check, and subtract out, errors—higher severity
4. It is more difficult to explain things away within these interconnected checks
5. Enables the capacity to discover a new effect, entity, anomaly

   If a theory says nothing about a phenomenon, its tests generally have no chance of discerning how it may be wrong regarding that phenomenon

   (e.g., central dogma of molecular biology did not speak of prions)

Simplicity criteria correlate with what's actually doing the work in knowledge promoting—but the correlation is quite imperfect, and when it holds, it is only indirectly getting at the problem

It is always a *second-hand* emotion to what's responsible for:

> the goals of ensuring and appraising how well-tested claims are, how well methods control and avoid error.