Paths in strange spaces: A comment on preregistration

Danielle J. Navarro

Original: November 2019; PsyArXiv: September 2020

Important note:

This essay is not a traditional academic output. It has not been peer-reviewed or submitted to a journal, nor indeed will it be. It was originally published as a two part blog post that no longer exists. Nevertheless, I have noticed that this essay has been cited in other academic works and I am loathe to leave holes in the academic citation trail. To remedy this I've reposted this as a preprint on PsyArXiv to ensure the document has a DOI and remains publicly available (the original URL now redirects to PsyArXiv to minimise any discontinuities). However, though it has been posted as a preprint it should not be mistaken for a peer-reviewed paper; it is merely an opinion piece. It has been very lightly edited for clarity but in most respects it is identical to the original version posted to my blog in November 2019. Correspondence concerning my ramblings, should such be necessary for some unfathomable reason, can be addressed to Danielle Navarro (School of Psychology, University of New South Wales, d.navarro@unsw.edu.au). Not surprisingly, as I wrote it originally as a personal reflection on the scientific process, the author asserts no relevant conflicts of interest and did not receive funding for this work. Duh.

Part 1

Here we go again, around and round
Here we go again, around and round
We're babies passing for adults
Who've loaded up their catapults
And can't believe the end results
So here we go again
– Aimee Mann[1]

A little over a week ago some colleagues and I uploaded a preprint to PsyArXiv entitled *"Preregistration is redundant, at best"*[2]. The manuscript is very brief, and is a response to another brief paper entitled *"Preregistration is hard, and worthwhile"*[3] published in *Trends in Cognitive Sciences*. We didn't advertise the paper on twitter. Personally I would have preferred not to do so at all, but the psyarxiv-bot[4] took matters out of our hands and posted it for us.

The firestorm that erupted was almost instantaneous.

I'm only a minor author on the paper – the hard work was done by Aba Szollosi and Chris Donkin, with commentary and contributions added by myself, Rich Shiffrin, Trish Van Zandt, Iris Van Rooij and David Kellen – but I'm terribly visible on twitter. So I had the privilege – or misfortune, depending on your point of view – to be tagged in tweets before I even knew the preprint was up. The experience was unpleasant, and not one that I'm in a hurry to repeat. In the last week there have been many responses to our paper, some positive and others negative, some thoughtful and others cruel. In this deluge of commentary I found two responses that I found particularly valuable: a blog post[5] by E.J. Wagenmakers and a twitter thread[6] by Christina Bergmann. What I like about both responses is that while both authors end up arguing that preregistration *is* worthwhile – and can be viewed as critiques of our paper – neither is dismissive and neither is insulting. Better yet, both critiques are thoughtful have made me think a lot about my own views, reconsidering some and reframing others. This kind of critique is one of my favourite things in science. I only wish I worked in an environment in which critiques were common and insults rare, rather than the real world in which these quantifiers are reversed.

Inspired by E.J. and Christina, this post is an attempt to articulate my own opinions on preregistration. Curious as this may sound given the high drama of the last week or so, I have not actually said much about my own views about preregistration. The paper – of which I am rather fond – was mostly Aba and Chris' work. While I'm glad I was able to contribute to their work, the only time I've ever expressed my own views about preregistration at any length was when I wrote *"Prediction, prespecification and transparency"*[7] in an invited blog post for the Psychonomic Society back in January.

I mention this because my post here is not really intended to add more fuel to the conflagration that has engulfed so much of open science twitter lately. Now that the reviews of our manuscript have been returned (with a revise and resubmit decision) I thought I might instead to pick up where I left off back in January when I started trying to put own my thoughts together.

Prelude I: Defining one's terms

One property I have observed in the online discussions of preregistration is that people are often unclear about what they take the term *preregistration* to mean. Does it refer specifically to the registrations toolkit within OSF?[8] Does it cover the less polished system offered by "aspredicted.org"? Does a non-public statement

[1] https://genius.com/Aimee-mann-simple-fix-lyrics
[2] https://psyarxiv.com/x36pz
[3] https://psyarxiv.com/wu3vs
[4] https://twitter.com/PsyArXivBot/status/1190014222526234624
[5] https://www.bayesianspectacles.org/a-breakdown-of-preregistration-is-redundant-at-best/
[6] https://twitter.com/chbergma/status/1190538529073508352
[7] https://featuredcontent.psychonomic.org/prediction-pre-specification-and-transparency/
[8] https://osf.io/

count as a preregistration? Is it a preregistration if the author can edit it after the fact but does not in fact do so? Does a precise mathematical model stated in paper X count as a preregistration for its subsequent use in paper Y? There is a distinct lack of clarity about these questions, and this lack of precision makes it difficult to have a sensible conversation on the topic. So for the purposes of the post I will assume that a preregistration system has the following properties:

- It allows authors to state data collection and analysis plans prior to running an experiment
- The statement must pertain to a specific experiment, not a general class of future experiments
- The statement is public, not private, and easily accessible by anyone
- Once the statement is lodged, the author cannot modify it

According to this construction, the OSF registration system and the aspredicted.org system both constitute preregistration tools; GitHub repositories do not. Similarly, open lab notebooks do not constitute preregistations; previously published computational models are not preregistrations; and so on.

This is a narrow definition of preregistration, admittedly, but I think it is an appropriate one. Almost without exception the public discussion around preregistration focuses on systems that are akin to the OSF registration system, and the only occasions on which I have seen anyone assert that "but a formal model *is* a preregistration..." are those in which the speaker has found themselves trying to move the goalposts. When one reads the *"Preregistration is hard, and worthwhile"* paper, it is very clearly making claims about an OSF-like system, and so too was our response paper. Neither paper makes claims that pertain to (say) the appendices of Tversky's (1977) paper[9] on *"Features of similarity"* in which he provides an axiomatic definition of what his featural account of stimulus similarity asserts. It is pure sophistry to pretend that the latter is part of the scope of what we mean in everyday parlance (it violates the first and second criteria above), and I will not treat this as a form of preregistration.

Prelude II: Defining the scope of one's argument

The second area in which I believe clarity is important at the outset regards the scope of the argument being made. The discussion on the vices and virtues of pregregistration has been far-ranging, and I have neither the desire nor the ability to capture all of the relevant issues in a blog post. As such, I will restrict myself to two specific claims about preregistration, namely:

- Preregistration prevents p-hacking
- Preregistration provides transparency

I chose these two quite deliberately, because I think they are the two most commonly advanced justifications for preregistration, but I think they are quite different from one another. In the first half of this post I'll talk about why I do not take the former claim seriously in the context of psychological research; and in the second half I'll talk about why I think the latter claim is accurate but misleading.

A second point about the scope of the piece. I'm not going to offer *empirical* evidence for my assertions here. Although I am quite firmly of the view that many questions about the efficacy and utility of preregistration are ultimately empirical claims rather than logical or philosophical ones, it is the latter that I am addressing in this post. That is to say, I am sketching out a case for why a rational person might have some justifiable skepticism about some of the claims that have been made about preregistration. Whether such skepticism turns out to be warranted by the data about how science progresses (in the presence or absence of preregistration tools) is a different question than the one I'm considering here. It is, nevertheless, an important topic and the fact that I do not discuss it here should not be taken as evidence that I think such matters are irrelevant.

With these preliminaries taken care of, I'll get to work, and the place for me to begin is to comment a little on how I think about the process of scientific discovery.

[9] https://doi.org/10.1037/0033-295X.84.4.327

On scientific discovery

> *Our conclusions must be warranted by the whole of the data, since less than the whole may be to any degree misleading. This, of course, is no reason against the use of absolutely precise forms of statement when these are available. It is only a warning to those who may be tempted to think that the particular precise code of mathematical statements in which they have been drilled at College is a substitute for the use of reasoning powers ... in which, as the history of the theory of probability shows, the process of codification is still incomplete.*
>
> – Sir Ronald Fisher, *The logic of inductive inference*[10] (1935)

Very often, when people discuss preregistration as a tool that scientists can adopt, it is introduced as a tool for improving our statistical inferences. Specifically, preregistration is advocated as a tool we can use to minimise the risk of "p-hacking" in the scientific literature. Although this is so often the go-to justification for preregistration, in my opinion it is easily the *least* valuable way in which scientists can use preregistration. I say this because I am increasingly of the view that hypothesis testing – be that in the form of an orthodox null hypothesis test or a Bayes factor analysis – is almost never relevant to the practice of scientific discovery. To put it crudely: I am quite unconcerned with p-hacking because in most instances I have little interest in p-values specifically nor in hypothesis testing generally.

The story of why I'm so uninterested in the process of "testing" a scientific hypothesis – in the usual sense of seeking to falsify that hypothesis – begins with a paper I published with my colleague Andy Perfors in *Psychological Review* back in 2011. The paper is entitled *[Hypothesis generation, sparse categories and the positive test strategy*[11]and it is not a statistics paper or indeed any kind of methodological paper. Rather, it is a cognitive science paper, and its goal was to discuss some of the reasons why confirmation bias is so prevalent in human reasoning. The key idea in our paper is that in the kinds of environments people tend to inhabit, if you have only a single plausible-sounding hypothesis that might account for some observed phenomenon and are afforded only the opportunity to pose yes/no "questions" of the world, on average you will tend to learn more (in the sense of expected information gain) by asking questions that are designed to *confirm* your hypothesis, rather than falsify it. Looking for "yes" is more informative – in the typical case – than looking for "no".

This result is surprising and might sound suspicious to some readers. Indeed, some suspicion *is* deserved, because the result does not hold generally and only applies to some kinds of worlds that the reasoner might find themselves in. Moreover, there's quite the literature on this topic and I'm not going to delve into it deeply here (except to note that our paper is merely one of many that have discussed this topic, and arguable not even one of the best papers that have done so!) but the thing I want to point out is this: if you have a hypothesis and your goal is to learn *the truth about the world*, attempting to falsify your hypothesis is in many cases a very inefficient learning strategy.

To anyone with a background in the philosophy of science literature these psychological findings would come as little surprise. The view of hypothesis testing I articulated earlier (and is popular among many scientists) is usually referred to by philosophers of sceince as the "naive falsificationist" view, and is a simplified version of Karl Popper's work. From my – admittedly superficial – reading of the history, Karl Popper's views on falsification stemmed from the logical asymmetry between verification and falsification; if a theoretical proposition P entails that we should observe the quantity Q and we do in fact observe Q we cannot logically conclude P to be true. Affirmation of the consequent is not a deductively valid form of reasoning. In contrast, if we observe Q to be *false* we can indeed reject proposition P; denial of the consequent is our modus tollens, and is deductively valid. Viewed this way, falsificationism offers us the tantalising possibility of placing science on a deductively solid foundation.

This is an appealing idea, as it would give is some of the inferential certainties we as scientist craves, but on closer inspection this foundation is decidedly unstable. Even if we set aside the nuances of what falsificationism looks like in practice – e.g., I'm going to pretend for the purposes of this post that I have never heard of the problem of ancillary assumptions – Popper's argument pertains to the testing of a single hypothesis. He says nothing about what process a scientist should follow in order to construct this hypothesis,

[10] https://www.jstor.org/stable/2342435
[11] https://psyarxiv.com/rj9kt/

nor does he tell us what we should do if that hypothesis is falsified. Worse yet, the falsificationist perspective provides no guarantee at all that this process of sequentially testing and falsifying hypothesis will get us any closer to the truth, or indeed any truth-like theories. There are infinitely many hypotheses that one *might* entertain about the world and only finitely many tests that we as scientists have the ability to construct. Mere falsification does not get us very far, and if our goal is to discover truths, a stronger principle is required.

Another way of framing this is to note that we have two rather different problems to consider. The process by which scientists *search* the space of possible theories for plausible hypotheses, and the process by which a scientist *tests* these hypotheses once generated are largely unrelated. As we found in our *Psychological Review* paper, a strategy that is good for one need not be good for the other: a falsificationist perspective allows you to evaluate a single hypothesis (per the infamous "four card task"[12] proposed by Peter Wason in 1968) it says very little about how one should search a space of possible hypotheses (per the much more interesting "2-4-8" task[13] that Wason introduced in 1960).

Naturally, because philosophers of science are sneaky people, this issue has been discussed extensively over in their little corner of the world. In particular, Imre Lakatos raised exactly this counterargument against Karl Popper's falsificationism. Viewed from the perspective of a scientist whose goal is to make *discoveries*, rather than the perspective of an engineer who wishes to test a system, Popper's framing of the problem is a kind of stage magic. What he does, in effect, is sweep everything that makes science hard (i.e., search and discovery) off to the side, replace it with a much simpler problem of testing and falsification, and declare that the result of this procedure can be viewed as a normative theory of science. Fortunately for me, I don't have to call out this trickery myself, because Lakatos already did it! Having read only a small part of his work, I'll cheat slightly and rely on the *Stanford Encyclopedia of Philosophy*[14]:

> *But Lakatos points out a problem. There is now a disconnect between the game of science and the aim of science. The game of science consists in putting forward falsifiable, risky and problem-solving conjectures and sticking with the unrefuted and the well-corroborated ones. But the aim of science consists in developing true or truth-like theories about a largely mind-independent world. And Popper has given us no reason to suppose that by playing the game we are likely to achieve the aim. After all, a theory can be falsifiable, unfalsified, problem-solving and well-corroborated without being true.*

This is precisely my concern with equating scientific work with the "testing" of hypotheses. At a fundamental level I do not believe this is a sensible thing to do, and to focus so relentlessly on *testing* things the way we do is a very bad idea.

Switching focus slightly, it is worth noting that Lakatos' philosophical concerns can be found mirrored in the writings of scientists and statisticians. As an example, you can find a lot of the same concern expressed in the early writing of Sir Ronald Fisher. For example, his 1935 paper on the logic of inductive inference is a quite wonderful article. Not only does he cover traditional statistical topics such as consistency, sufficiency and ancillarity, he is quite careful to highlight the dangers of relying too much on deductive reasoning:

> *Although some uncertain inferences can be rigorously expressed in terms of mathematical probability, it does not follow that mathematical probability is an adequate concept for the rigorous expression of uncertain inferences of every kind. This was at first assumed; but once the distinction between the proposition and its converse is clearly stated, it is seen to be an assumption, and a hazardous one. The inferences of the classical theory of probability are all deductive in character. They are statements about the behaviour of individuals, or samples, or sequences of samples, drawn from populations which are fully known.*

In modern parlance, what Fisher is talking about here is the fact that every statistical inference that we make requires us to *assume* that some model $\mathcal{M}$ provides an adequate characterisation of the problem at hand. We cannot do statistics at all without constructing a model (or set of models), and all our inferences depend on a kind of parlour trick... we use our models as proxies for the world itself, consider what our

[12] https://doi.org/10.1080/14640746808400161
[13] https://doi.org/10.1080/17470216008416717
[14] https://plato.stanford.edu/entries/lakatos/

models assert about the data we have observed, and then use this consideration to licence conclusions *about the world*. As was his wont, Fisher pulls no punches. He goes on to note that Bayesianism (the theory of inverse probability) offers no defence against this vulnerability:

> *Even when the theory attempted inferences respecting populations, as in the theory of inverse probability, its method of doing so was to introduce an assumption, or postulate, concerning the population of populations from which the unknown population was supposed to have been drawn at random; and so to bring the problem within the domain of the theory of probability, by making it a deduction from the general to the particular*

Now, some care is needed here. Fisher's assertions about the limits of probabilistic models are not precisely the same as Lakatos' claims about the limits of falsificationism. They operate at different levels of abstraction, they have different historical contexts, and so on. The thing they have in common, however, is that they expose the danger of pretending that your elegant inferential tool – whether that be falsificationism or statistical hypothesis testing – can possibly serve as a good proxy for the scientific discovery process itself.

Though I am myself a Bayesian data analyst by preference, and as such I'm less opposed to inverse probability than Fisher was, I am very much in agreement with him that all statistics starts with an unverifiable *supposition* that a particular model class is applicable to the problem at hand, and because these models are – at best – a crude approximation to the world, it is an act of extreme recklessness to put too much faith in them. They are tools, nothing more. When it comes to doing science, the tools might be handy things to use from time to time, but when you want to work out what they are telling you about the world itself, you're on your own honey.

On hypothesis tests and mathematistry

It has not escaped my attention that nothing in my discussion of scientific discovery has, to this point, made any connection with the role of preregistration and the dreaded crime of "p-hacking". It is high time I remedied this oversight, and the place for my to start is by questioning the value of hypothesis tests – after all, if preregistration is purportedly of value *because it prevents us from distorting the results of a hypothesis test*, then it must follow that hypothesis tests are themselves of some value to us as scientists. There is little point in trying to fix a thing that is irredeemably broken, after all.

With that in mind, I want to turn to my very favourite article in statistics. It is a 1976 paper by George Box, simply entitled *"Science and statistics"*[15]. Box's paper is a reflection of the legacy of Sir Ronald Fisher as scientist and a statistician, and the manner in which Fisher's thinking was influenced by the fact that he was both.[16] It is a brief paper, and one I would recommend that every scientist and statistician read at least once in their career. In the later stages of the paper Box discussed Fisher's irritations with *mathematistry*, which is pertinent to my annoyance with the claim that "preregistration prevents p-hacking". Here's Box:

> *Mathematistry is characterized by development of theory for theory's sake, which since it seldom touches down with practice, has a tendency to redefine the problem rather than solve it. Typically, there has once been a statistical problem with scientific relevance but this has long since been lost sight of.*

To my mind, mathematistry is slightly more general and is concerned with what we might call "the illusion of rigour". It applies in any situation where we confine ourselves to an unnecessarily rigid system to provide ourselves with an impression of sureties where none in fact exist. In *"Between the devil and the deep blue sea"*[17], I suggested that mathematistry is the practice of

> *using formal tools to define a statistical problem that differs from the scientific one, solving the redefined problem, and declaring the scientific concern addressed.*

[15] https://www.jstor.org/stable/2286841

[16] Fisher's legacy as a eugenicist ought also be noted, and not in a positive light. In the real world, science is not politically neutral and as scientists we ought to have the honesty not to hide from the terrible things that have been justified in our name: https://www.adelaide.edu.au/library/special/exhibitions/significant-life-fisher/eugenics/

[17] https://psyarxiv.com/39q8y/

The practice of hypothesis testing in psychological science (via null hypothesis testing, Bayes factors, or what have you) is often has this character. Statistical hypothesis testing provides an elegant, clean formalism for connecting a statistical toy (your model) to a dubious measurement (your data)... and at the end you get a number, p, that is supposed to summarise what these two things have revealed to us about the world.

Is this practice of interest to the scientist? I think not.

Hypothesis tests are mostly an irrelevance to scientific process, in my view, and any attempt to "fix" them via preregistration or any other practice is misguided. I'm reminded of the following quote from the classic 1963 paper *"Bayesian statistical inference for psychological research"*[18] by Ward Edwards, Harold Lindman and Leonard Savage:

> *No aspect of classical statistics has been so popular with psychologists and other scientists as hypothesis testing, though some classical statisticians agree with us that the topic has been overemphasized. A statistician of great experience told us, "I don't know much about tests, because I have never had occasion to use one."*

I wish I could say the same. In my own research I have on many occasions been *required* to report the results of a hypothesis test, such being the expectations of editors, reviewers and readers, but I confess in no instance have I found these tests helpful. Were I given the freedom to do science the way I think best, I would not report a single p-value or Bayes factor in any of my papers. The fact that I do so is entirely because I am forced to do so by others.

The reason I have little time for hypothesis tests is that I have yet to encounter a scientific situation where I have a hypothesis that I would deem to be sufficiently precise that it *warrants* testing. And I am saying this as a mathematical psychologist! Compared to most psychologists my work is highly circumscribed, extremely formal, and precise in a fashion that we don't usually bother with in this discipline. Nevertheless, I do not believe my theories and my models are well-formed enough that I would attempt to "test" them in the manner that a null hypothesis test requires. There is too much uncertainty regarding my operationalisations and too much unclarity about what my models assert about the structure of human cognition for this to be wise. As Fisher noted in 1935:

> *Our conclusions must be warranted by the whole of the data, since less than the whole may be to any degree misleading.*

For the situation we find ourselves in as psychologists, this is a major concern. In almost all cases our measurement instruments are proxies – we are *never* measuring the actual thing we care about – and our models are fantasies. It is very rare that our data are tightly linked to the phenomenon of interest and even rarer for the model to bear any strong connection to the theoretical claim we wish to assert. In such a world, your p-value is a lie before you even compute it. We are not drawing conclusions from "the whole of the data" because most of the relevant facts are hidden from us in the first place. It makes little difference whether your p-value has been "hacked" because it wasn't telling you very much in the first place. Rigorously quantifying an inference using a terrible statistical tool simply to cling to the illusion that it is more "scientific" to attach a number to your guesswork is as pure an example of mathematistry as I can think of. As a discipline have become entranced by the mathematical elegance of statistical theory, and I think it is to our detriment. Box continues:

> *Furthermore, there is unhappy evidence that mathematistry is not harmless. In such areas as sociology, psychology, education, and even, I sadly say, engineering, investigators who are not themselves statisticians sometimes take mathematistry seriously. Overawed by what they do not understand, they mistakenly distrust their own common sense and adopt inappropriate procedures devised by mathematicians with no scientific experience*

Unless and until we reach the point where we could plausibly construct the required mapping between a theoretical claim and an observable measurement, there is no point in deluding ourselves into thinking that our hypothesis tests are fit for purpose, regardless of whether we have preregistered them. We are, as Aimee Mann notes, "babies passing for adults, who've loaded up their catapults, and can't believe the end results".

[18] https://psycnet.apa.org/record/1964-00040-001

Intermission: When your plan becomes a prison

Before I move on, it is important to recognise the narrowness of the claim I have made above. Specifically, the argument in the preceding section is *entirely* focused on the particular claim that "preregistration prevents p-hacking". It does not have anything to say about *other* possible virtues of preregistration. In respect of that claim, I argue that advocating preregistration as a solution to p-hacking (or its Bayesian equivalent) is deeply misguided because we should never have been relying on these tools as a proxy for scientific inference in the first place.

It is important to recognise this, I think, because there is a very real danger that preregistration systems will ossify scientific practices in an undesirable fashion. That is to say, while advocates of preregistration often claim that *"preregistration is a plan, not a prison"*, they will also claim that preregistration is necessary to prevent p-hacking. Unfortunately, you cannot have it both ways: in my opinion these two claims are in direct opposition to one another unless you *also* stipulate an unreasonable level of foreknowledge on behalf of the experimenter.

To see why I say this, I think it is helpful to formalise what we mean by "p-hacking" here. Under Fisher's informal view of hypothesis testing it's not entirely clear what would constitute p-hacking, so – adopting the same rhetorical position that advocates of preregistation use – I'll switch to using Neyman's more formal decision theoretic construction, and assert that p-hacking is any research practice that causes the "true" Type I error rate associated with a test to depart from its nominal rate. Can preregistration guard against this? Well, I think it depends on how strict you consider the preregistration to be. I can think of three cases:

- *Strict.* Researcher specifies the complete analysis plan in advance and does not deviate even if the data turn out to be highly surprising. This does indeed prevent p-hacking in the conventional sense but forces the researcher to use inappropriate statistics: under the strict interpretation, preregistration is in fact a prison, not a plan.
- *Flexible.* Researcher specifies the analysis plan, but is willing to deviate if in retrospect the planned analysis seems inappropriate for the data. This satisfies the claim that preregistration is a plan not a prison, but does so by opening the door to p-hacking once more. A researcher may unwittingly decide to apply greater scrutiny to data that they find disappointing, and thereby become more likely discover reasons to justify departures from the plan.
- *Oracular.* There is a third possibility: the researcher writes a preregistered plan that covers every possible eventuality, listing exactly how each case will be handled and then – because this composite if/then decision making procedure is no longer a specific hypothesis test – derives an appropriate decision policy for that plan which satisfies Neyman's admissability criteria. This would indeed allow us to have the best of both worlds: because the prespecified "deviation plan" is incorporated into the design of the subsequent decision policy, we can have the flexibility we desire (our preregistration is indeed a plan and not a prison) while ensuring that our overall Type I error rate remains bounded at its nominal level (no p-hacking allowed). Perfect... and all it requires is godlike planning abilities! I'll confess I don't personally have the intellect to construct that kind of analysis plan for the kinds of experiments I do, but perhaps someone smarter than me can figure it out.

There is a second, perhaps worse, sense in which preregistration can become an unintended prison. In practice, when you look at the kinds of plans that researchers are often expected to preregister – and the reasons given for why we must do so – they almost always related to the *preregistration of hypothesis tests*. Although it is entirely possible to preregister something else (e.g., that the researcher intends to undertake an exploratory data analysis focusing only on estimating unknown quantities and not running any tests), it is very clear from the relentless focus on p-values and p-hacking that this is NOT the use case that most people have in mind for preregistration. If anything, because we focus so much on p-hacking specifically when we talk about preregistration, we end up in a situation where we direct *even more* of the methodological discussion onto hypothesis tests.

In other words – as much as I love the idea of having a plan, and endorse unreservedly the suggestion that we all should do so when designing, conducting and analysing experiments – I believe that *any attempt to frame preregistration as a solution to the problem of p-hacking is a way of locking ourselves into a prison.*

When we do this we confine ourselves to a statistical practice (hypothesis testing) that I believe is deeply ill-suited to most of the scientific problems we face in psychology.

Registration, documentation, and transparency

If you get a feeling
Next time you see me
Do me a favour and let me know
Because it's hard to tell
It's hard to say
Oh well, okay
– Elliot Smith[19]

There is a second kind of justification that we often give when extolling the virtues of preregistration, namely that it provides *transparency* to the scientific process. To my way of viewing things, this justification needs to be taken much more seriously. Unlike the "problem" of p-hacking – which I view as a mere side effect of the more serious problem psychologists using trying to force a round peg (scientific discovery) into a square hole (hypothesis testing), and one that cannot itself be solved by preregistering an unwise practice – the lack of transparency around scientific processes *is* a real problem, and one that preregistration might plausibly help with.

In fact, as it will turn out, it is in this respect that I quite like preregistration and advocate that it – or some other documentation process that is *at least* as thorough and transparent – be standard practice in scientific work. This is the focus of part two.

Part 2

I'm caught up in something I don't get
And I don't understand how I got here
And I'm losing everything I knew
And it was all for you
Could you be a little easier?
Could you be a little easier on me?
– Leddra Chapman[20]

The first half of this post paints a bleak picture of the scientific process. Not for the first time, I argued that in most situations facing psychological scientists, there is little reason to be worried about "p-hacking" per se because the p-values we report in our papers were never fit for purpose in the first place. Nor do I let Bayesians such as myself off the hook. While discussion among psychological methods researchers tends to focus on the pathologies of p-values in both the presence or absence of preregistration, statisticians are quick to point out that in practice the Bayes factor – often touted as the Bayesian alternative to orthodox hypothesis tests – has pathologies of its own. If preregistration is unlikely to provide me with the oracular foreknowledge I need to construct a Neyman-admissable decision procedure, it it hardly any better equipped to provide me with the precise knowledge I require to specify priors, particularly not with respect to complicated models that require me to think about high-dimensional parameter spaces. Taken at face value, I seem to be arguing a rather nihilistic position. Nothing works. Everything is broken. Inference is a doomed enterprise. Perhaps we are deluding ourselves to think that there is any hope...

On epistemological anarchy

In my early 20s I read Paul Feyerabend's (1975) classic work on the philosophy of science, provocatively entitled *Against Method* and I hated it. Part of the reason I hated it so much is the way that his work was introduced to me in my undergraduate philosophy of science class (yes, I actually took one!) As it was

[19] https://genius.com/Elliott-smith-oh-well-okay-lyrics
[20] https://genius.com/Leddra-chapman-a-little-easier-lyrics

described to me, Feyerabend was arguing that there is nothing special that differentiates science from any other belief system, that scientific methodology adds nothing worthwhile, and that when it comes to making inferences about our world, there is only one principle: "anything goes". I approached the book with a very hostile mindset, and to me it seemed disorganised, unscientific, and riddled with logical errors. In retrospect, I suspect mine was an uncharitable reading. Because I've lost my copy of the book, I'll cheat again and use the *Stanford Encyclopedia of Philosophy*[21] to supply the context I was missing:

> *By the early 1970s Feyerabend had flown the falsificationist coop and was ready to expound his own perspective on scientific method. In 1970, he published a long article entitled "Against Method" in which he attacked several prominent accounts of scientific methodology. In their correspondence, he and Lakatos subsequently planned the construction of a debate volume, to be entitled For and Against Method, in which Lakatos would put forward the "rationalist" case that there was an identifiable set of rules of scientific method which make all good science science, and Feyerabend would attack it. Lakatos' unexpected death in February 1974, which seems to have shocked Feyerabend deeply, meant that the rationalist part of the joint work was never completed.*

In other words, the *intended* structure of the work was one that should be familiar to most of us as scientists: on the one side (Lakatos) we have theory building, and on the other (Feyerabend) we have theory criticism. These two components are supposed to work together, and as individual scientists we try (hope) to engage in both sides of this process iteratively. We build our theory, attack our own theory, when it fails we build a new one and so forth. Both are necessary, and in retrospect I think *Against Method* is a less impressive work than *For and Against Method* would have been. In any case, here's the summary of Feyerabend's argument:

> *Against Method explicitly drew the "epistemological anarchist" conclusion that there are no useful and exceptionless methodological rules governing the progress of science or the growth of knowledge. The history of science is so complex that if we insist on a general methodology which will not inhibit progress the only "rule" it will contain will be the useless suggestion: "anything goes". In particular, logical empiricist methodologies and Popper's Critical Rationalism would inhibit scientific progress by enforcing restrictive conditions on new theories. The more sophisticated "methodology of scientific research programmes" developed by Lakatos either contains ungrounded value-judgements about what constitutes good science, or is reasonable only because it is epistemological anarchism in disguise. The phenomenon of incommensurability renders the standards which these "rationalists" use for comparing theories inapplicable.*

This summary matches my recollection of the book rather well... and it mirrors my own experience as a scientist rather well too. For example, in my own area of research there is a degree of tension between the "Bayesian models of cognition" school of thought that views human inductive reasoning as a form of probablistic inference, the "heuristics and biases" school that assumes our reasoning is based on simple, error-prone approximations, and the "connectionist" school of thought that emphasises the importance of parallel distributed computation and the underlying cognitive architecture. These three different frameworks *are* largely incommensurate. I've used all three at different stages of my career, and while most of my work falls within the "Bayesian cognition" framework I don't necessarily think it is "better" than the other two.

I'm not even sure the question makes sense. Much as Kuhn points out in *The Structure of Scientific Revolutions*, each paradigm emphasises different empirical phenomena, selects different operationalisations, and produces formal models that have different intended scope. While in the long run – of course – we would hope to construct a single unifying framework that encompasses all of human cognition, we are not even remotely close to that point. Trying to decide on which of these three completely incommensurable paradigms is "least wrong" when the reality is almost certainly that all three are spectacularly wrong, is just silly. Right now, with the cognitive science literature being where it stands, all three paradigms offer useful insights, and it is a good thing that we as a discipline retain all three.

The Stanford Encyclopedia entry continues:

> *At a time when Kuhn was downplaying the "irrationalist" implications of his own book, Feyerabend was perceived to be casting himself in the role others already saw as his for the taking.*

[21] https://plato.stanford.edu/entries/feyerabend/

I respect Feyerabend a lot for this. I did read Kuhn's book too (though I don't remember it very well) and I did get the impression that he was a little nervous about the entailments of the "incommensurability" problem, and sought to hide or minimise them. Feyerabend did not shy away from it, and as a result *Against Method* is a very provocative and unsettling read for a scientist.

On modesty and the scientific bootstrap

Oh dear. I seem to be digging myself into a deeper and deeper hole. I started this post with some statistical concerns about p-values and apparently I'm now at the point of endorsing epistemological *anarchy*? Really? That's... not a good place for a scientist to be! Well... maybe it's not so bad. To me, the major, substantive point that Feyerabend made is this one:

> *The history of science is so complex that if we insist on a general methodology which will not inhibit progress the only "rule" it will contain will be the useless suggestion: "anything goes".*

I think this is entirely correct. There are no hard and fast rules for good science, no magical set of procedures that we can follow that will guarantee – not even to a known probability of error – discovery of truths. To me it seems logically incoherent even to imagine that such a set of rules *could* be proposed by humans. It would be a different story if we already knew the truth. If we already knew the truth about our world, and how observations can be made within that world, *then* we would be able to work out what inferential rules make sense for that world. Until that time comes that we have such an complete understanding, however, we are relying on our best guess about the structure of the world to work out what the rules for learning about the world should be! As scientists we are hoping to bootstrap our way to the truth.

To me that seems like a very reasonable strategy, and I can't think of anything better, but let's not pretend that we really know what we're doing here. We're all making it up as we go along, to the best of our ability, hoping not to make a mess of everything. Under the circumstances, I think a little modesty in our scientific and statistical claims would be in order, no?

In the garden of forking paths

Besides the importance of being modest, my little story about reading Feyerabend in the misspent years of my youth contains the kernel of a defence for preregistration. Why was I so hostile to Feyerabend the first time I read *Against Method*? Mostly it was because of my history: I brought my own preconceptions to the book that were based on someone else's reading of the book (i.e., my undergrad lecturer) and that led me to emphasise some things and not others. I chose those parts of the book that seemed most relevant to me at the time, and those choices shaped my conclusions. Not only that, I was unaware of Feyerabend's history. I did not know that the work was originally intended to be joint work with Lakatos. Had I read the intended work, *For and Against Method*, I suspect I'd have come to different conclusions.

Sure, when I sat down as an impressionable 20-something to read the book, I read it with what I thought of as an open mind, and the book itself is what it is, but this "state" is not sufficient to properly describe or make sense of the situation. The earlier states matter too. The book has a history, a path that brought this specific volume to me, and I had a path that brought me to read it. Both of those histories matter. Knowing those histories better, as I do now, leads to a rather different impression of the book (and the reader!) than one might have without knowing this history. If you want to understand my reading of *Against Method* you need need to know where it came from, and where I came from.

So it goes in scientific research also... if you want to understand any scientific work properly, you need to know its history. You need to know where the data are coming from, what inferences the experiment was designed to support (and what inferences it was not), and because as psychologists we have only limited knowledge of what is and is not *relevant* to the study of a phenomenon, the only plausible mechanism we have is to document as much as we possibly can, as carefully as we possibly can. Epistemic modesty requires that we acknowledge our own limitations as researchers; we all make mistakes, we all miss details that matter, and the best hope we have for making progress – in my view at least – is to document the path we took through "the garden of forking paths" in the hope anyone who wants to build from our work can "backtrack", auditing and retracing the process. It also, I think, entails a principle of **kind critique**. Acknowledging our

own flaws requires us to avoid harshness in how we evaluate the work of others; to the extent that we begin to endorse a culture of harsh criticism, we encourage others to be competitive, defensive and hostile. This is the antithesis of what we should desire in a scientific process, I think.

The parable Jasmine and Rosemary

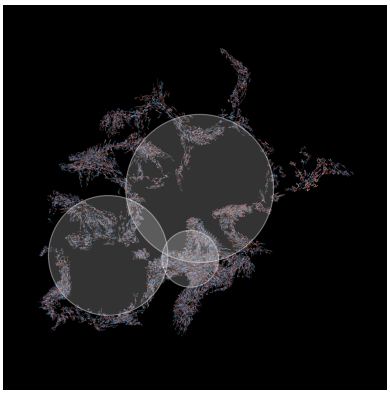
Throughout this post I've been inserting small snippets of code that construct generative artwork using the "jasmines" R package[22] that I've been slowly writing. The story behind the jasmines package is an interesting one, because I started it as an exercise in reverse engineering – I wanted to understand how Thomas Lin Pederson was creating beautiful pieces like this one:

[see genesis 112 piece at: https://www.data-imaginist.com/art]

As a matter of personal aesthetic principle Thomas doesn't release the source code to his artwork, which does make it tricky to work out what he's doing. Over time my goals with jasmines have shifted, but my initial intent was simple: work out how the magic of Thomas' art is performed and construct my own code that could approximate the behaviour of his system. This endeavour has a lot in common with how I do science: "out there" in the world there is a system (the mind) whose behaviour I don't understand, and my goal when developing computational models of cognition is to develop a formal system whose properties I *do* understand and whose behaviour approximates (in some limited way) that unknown system. Theoretical modelling in science is fundamentally an exercise in reverse engineering: trying to work out how system A works by building another system B that you control, and whose behaviour is the same as system A. The way I wrote the jasmines package provides a neat way of thinking about how I do science.

As is usually the case in science, my artistic project did not start as a blank slate. I began the project using some clues that the world had already given me. For instance, I knew that Thomas wrote the "ambient" package[23] that provides an R interface into the C++ FastNoise library, and I'd played with the ambient package before, enough to know you can use it to generate a variety of different textures. I'd also knew from twitter that Thomas had been playing around with curl noise slightly before these lovely pieces started to appear. So it was natural to expect that the `ambient::curl_noise()` function would be the place to start, and the `jasmines::unfold_tempest()` function is built on top of it.

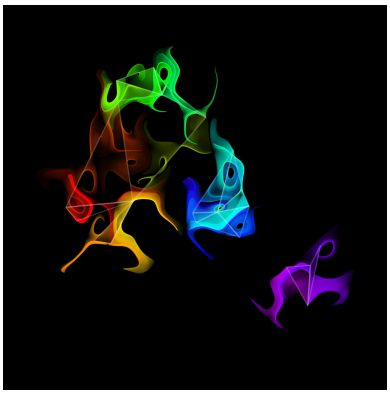
My first efforts did not resemble Thomas' in any meaningful sense:



This is very pretty, but it tells me very little about how Thomas performs his magic. If you squint hard enough you might suspect that there is *something* in common between the long flowing tendrils in Thomas' piece and those in my piece, but it's hardly very compelling. I retained the source code for this "constellations" piece, and continued tinkering. I rewrote various functions in the jasmines package, I tried new parameter settings and I started generating other pieces. Most of this process was exploratory. Sometimes I would create things that feel qualitatively similar to Thomas' pieces. For example, while there is almost no pixel-by-pixel match between Thomas' art and this piece...

[22] https://github.com/djnavarro/jasmines
[23] https://github.com/thomasp85/ambient

or this piece...



... it is hard to avoid the intuition that these pieces share something with Thomas' at a fundamental level, in a way that the constellations piece does not. As time went on, I created more and more pieces using the jasmines package, and from these creations I started to extract some general sense of *what* Thomas is doing when he creates his artwork. I liked these experiments in generative artwork enough that I decided to bundle them together into a separate R package that I called "rosemary"[24].

The two packages have a symbiotic relationship. It is immediately obvious that the rosemary package is deeply reliant on the toolkit provided by jasmines: it is a collection of experiments designed using jasmines, and cannot function without it. However, it is no less true to not that jasmines depends on rosemary. The experiments that I ran using rosemary are my only mechanism to link my artistic "theory" (jasmines) with the underlying phenomenon (Thomas' artwork) that I'm studying. Whenever I create a new artistic work with rosemary, it influences how I think about how the generative system needs to be structured, and guides the next step in the development of jasmines.

This kind of "virtuous cycle" is how we hope our scientific processes unfold. We rely on our theoretical insights (jasmines) to design experiments (rosemary) that allow us to modify our theories, design new experiments, and so on. The interplay between these two components is – we hope – the process that allows our theories to better approximate the truth, and our experimental results to better target the phenomenon of interest. However, the scientific bootstrap is not magic, and if we are not careful we may find ourselves wandering around at random, making little progress.

Unexpected fragility

The principle that underpins the scientific bootstrap, as I see it, is one of incremental improvement. Each new experiment that Rosemary conducts provides a constraint that Jasmine can incorporate into her theories, giving Rosemary an opportunity to design a better experiment, and so on. In an ideal world every time Jasmine adjusts her theory, she must do so in a way that is consistent with Rosemary's history. Jasmine is

[24] https://github.com/djnavarro/rosemary

very reactive, and her theory building is almost entirely post hoc, adapting to each new piece of evidence that Rosemary provides. However, she must operate within tight constraints: she must maintain backward compatible with previous experiments by Rosemary. As the volume of experiments increases, this is *much* harder to do than it looks. As an example, a couple of weeks ago I pushed an innocuous-looking update to the `jasmines` package that broke backward compatibility and rendered a few of my very favourite pieces irreproducible. The vast majority of the output was fine but a smallish number of the pieces had distortions introduced. The distortions were always in the colour palettes, and it took me several hours of work to figure out *what* had happened. Why did it distort such a small number of pieces, and why did it do so only in this specific way?

Path tracing in scientific hypothesis spaces

By the time I encountered this problem I'd created a great many pieces with rosemary, and the code base for the jasmines package had become a sprawling mess. This is of course to be expected in any exploratory process, be it scientific or artistic in nature. The underlying process is a *search* for something, and as you react to different cues in your environment you can end up tracing a very strange path and leaving quite a mess behind you. When something breaks your connection to your past, it is easy to find yourself completely, hopelessly lost.

What saved me is that I had *documentation*. Both of my packages were developed using git for version control, and I put a modest amount of work into ensuring that each change I made to the source code was added to the repositories with a somewhat informative commit message. It took me a while but I found the specific, utterly innocuous-looking change[25] that had broken my code. There was a line of code that I had used to count the *number* of colours needed to draw an image that was unnecessarily restrictive, so I modified it so that it would work for a broader range of possible scenarios. The old code only gave a meaningful answer in some cases, so I wrote a new version that always returned the same answer in those cases, but would *also* work in other cases.

You'd think that would ensure reproducibility of my code, right? The new version always returned the same number as the old one, for every situation of when the old one worked. So it's functionally "the same" as far as all retrospective cases are concerned and nothing should break. Unfortunately, this intuition is wrong. The two functions produce their answers by invoking the random number generator (RNG), and although both versions produce the same answer, they invoke the RNG a different number of times to do so. When executed within R sessions that have the same RNG state, they always produce the same output, and so one is tempted to conclude that they are equivalent, but because they leave the the RNG in *different* states, any subsequent computations that rely on the RNG will no longer yield the same output. A simplified illustration of the problem I encountered can be found on github.[26] [27]

Corrupting the state of the RNG is one particularly insidious way to break the reproducibility of one's code, but as any software developer will tell you, there's no shortage of frustrating ways to accidentally break code, even when you are being careful and using good coding practice, especially when you are involved in a collaborative project where you rely on other people's code and are not always aware of exactly what that code does and how it changes over time. As I said on twitter (rather melodramatically in retrospect), at the time I discovered my RNG state bug, when seeking to write reproducible code, *"dark shapes move beneath us"*.

Documentation in a time of anarchy

In keeping with my terribly meandering habits when writing blog posts, this discussion has wandered across a variety of topics. Even so, I think there is a single underlying theme to all this, and an important cautionary note for psychologists and other scientists. Whenever you are investigating a complicated system that you barely understand, there is a need for "epistemic modesty". We need to recognise that there the limits to

[25] https://github.com/djnavarro/jasmines/commit/695e773fb457b3c37780eefee094b081c1cdc509#diff-6ad78275fef5e660c04ab1601d8686d7L77

[26] https://gist.github.com/djnavarro/c992f93681f9b13141101d8a802ff53c

[27] https://gist.github.com/djnavarro/21525733f18f29f94613777b99ef8e5c

the diagnosticity of experiments, to the informativeness of our theories and to the relevance of our statistical tools. Very often we don't even know what is relevant to a phenomenon and what is irrelevant. We will – more often than not, I suspect – turn out to be wrong in what we infer from our data.

If this is the case, what hope do we have for *incremental* science? If it is in fact true (and I suspect that it is) that most of our empirical findings are wrong and our theoretical models poorly constructed, how will we ever "build" on previous findings? If you endorse a full blown "epistemological anarchist" view the way that Feyerabend does (and the older I get the more my view does start to look rather anarchic), what hope do we have?

The answer, I think, lies in meticulous documentation. In any given project I will always try my very best not to make mistakes, not to rely on foolish assumptions, and so on, but there are so very many ways to make a mistake that it is almost inevitable that I will slip up somewhere. This is exactly the kind of scenario I found myself in with the Rosemary/Jasmine code. I tried so hard to preserve reproducibility, to write good code, and I still made mistakes – but because I left behind a trail showing exactly what I had done and what decisions I had made at each step, I was able to work out what my error was and fix it. I think this principle holds more generally, and highlights the overwhelming importance of transparency and documentation. If someone else wants to rely on my work (even if that's just me a few months later), it's not sufficient for me to simply assert "I did X and found Y" the way that a brief report journal article often does. I need to give you more details than that. I need to leave behind this rich trail of breadcrumbs, exposing all the decisions I made and my reasons for making them. In an ideal world my work should "speak for itself", and it should be possible for anyone reading my 4000 word brief report to extract what they need. The real world, however, is less than ideal, and documentation is critical.

Preregistration as a documentation system

Returning at long last to preregistration, I hope it is clear that while I have been deeply skeptical of the idea that psychologists should be using preregistation to prevent p-hacking (see part 1 of this essay), I am *extremely* sympathetic when people advocate preregistration as a tool to improve documentation and the transparency of the research process. The only sense in which I have "reservations" about preregistration in this context is that I worry that it doesn't go far enough. Here's what I wrote in my original blog post[28]. In the original blog post I'm talking primarily about the kind of computational modelling work that I do, but I suspect it has value for other situations as well.

> *There are reasons why one might want to employ something akin to preregistration here: building a new computational model is a creative and iterative process of trying out different possible models, evaluating them against data, revising the model and so on. As a consequence, of course, there is quite rightly a concern that any model that emerges will be overfit to the data from my experiments. There are tools that I can use to minimize this concern (e.g., cross validation on a hold-out data set, evaluating the model on a replication experiment, and so on), but to employ them effectively I need to have alternatives to my model, and this is where an extremely high degree of transparency is important. Should someone else (even if that's just me a few months later) want to test this model properly at a later date, it helps to be able to follow my trail backwards through the "garden of forking paths" to see all the little decisions I made along the way, in order to ask what are the alternative models she didn't build? To my mind this really matters – it's very easy to make one's preferred model look good by pitting it against a few competitors that you aren't all that enthusiastic about. To develop a "severe test", a model should be evaluated against the best set of competitors you can think of, and that's the reason I want to be able to "go back in time" to various different decision points in my process and then try to work out what plausible alternatives might have emerged if I'd followed different paths.*
>
> *With this in mind, I don't think that (for example) the current OSF registration system provides the right toolkit. To produce the fine-grained document trail that shows precisely what I did, I would need to create a great many registrations for every project (dozens, at the very least). This is technically possible within the OSF system, of course, but there are much better ways to do it.*

[28] https://featuredcontent.psychonomic.org/prediction-pre-specification-and-transparency/

> *Because what I'm really talking about here is something closer to an "open notebook" approach to research, and there are other excellent tools that can support this. For my own part I try to use git repositories to leave an auditable trail of commit logs that can be archived on any number of public servers (e.g., GitHub, BitBucket, GitLab), and I use literate programming methods such as R Markdown and Jupyter notebooks to allow me to document my thinking on the fly during the model building process. Other researchers might have different approaches.*
>
> *Although the goals of the open notebook approach are not dissimilar to preregistration insofar as transparency is relevant to both, there are a lot of differences. The workflow around the OSF registration system makes it easy to lodge a small number of detailed registrations, whereas a notebook approach based around git repositories emphasizes many small registrations ("commits") and allows many paths to be explored in parallel in different branches of the repository. Neither workflow seems inherently better than the other in my view: they're useful for different things, and as such it is best not to conflate them. Trying to force an open notebook to fit within a framework designed for preregistrations seems likely to cause confusion rather than clarity, so I typically don't use preregistration as a tool for aiding transparency.*

My point when writing this was *not* to suggest that nobody should use preregistration as a tool for aiding transparency. On the contrary, I think it is a very useful tool for many people. However, it is not the only tool in our toolbox, and there will be many occasions when it isn't the right tool for the job. From my perspective, I'm in favour of any effort that a researcher can undertake that makes their research more transparent (within the limits of what is ethical), and any form of documentation of process they can provide.

Epilogue

The point I've been trying to make throughout this post (both halves), and elsewhere in my writings, is that scientific inferences are *hard*. Our statistical tools are rarely suited to the problem we're trying to solve, we as researchers don't understand the phenomenon we are studying so we often study it in less-than-ideal ways, and so on. Per Leddra Chapman, we are *almost always* trapped in a state of severe ambiguity and uncertainty

> *I'm caught up in something I don't get*
> *And I don't understand how I got here*

This makes science hard. Small missteps can make it hard to work out what we have done, and it is almost impossible to avoid making these missteps. As I see it, we have two options available to us: we can either lie to ourselves and pretend that these missteps are avoidable, or we can be honest about the fact that they happen and take sensible precautions. Speaking only for myself, I prefer the latter.

> *And I'm losing everything I knew*
> *And it was all for you.*

I don't want to risk losing *everything* I've learned from my research. If my work is deemed invalid because I made a small mistake, then I can guarantee that everything I have ever done is invalid. So is everything that you have done, dear reader. All research is flawed. We can – and should – do everything we can to minimise those imperfections, but we should never let perfect be the enemy of good. It is in *this* respect that I wholly endorse preregistration (or registrations, or documentation systems generally). In a complicated world filled with traps lying in wait for the unwary, anything that we can do to help us recover from our inevitable missteps is to be desired. Documentations systems are there, not to drown us in work, but to keep us afloat in a sea of uncertainty.

> *Could you be a little easier?*
> *Could you be a little easier on me?*