OPINION ARTICLE

Three more steps toward better science [version 1; peer review: 1 approved, 1 approved with reservations]

Jose D. Perezgonzalez

Business School, Massey University, Palmerston North, 4442, New Zealand

v1 **First published:** 31 Oct 2018, **7**:1728 (https://doi.org/10.12688/f1000research.16358.1)

Latest published: 01 Apr 2019, **7**:1728 (https://doi.org/10.12688/f1000research.16358.2)

Abstract

Science has striven to do better since its inception and has given us good philosophies, methodologies and statistical tools that, in their own way, do reasonably well for purpose. Unfortunately, progress has also been marred by warring among different perspectives, leading to troubles such as the current reproducibility crises. Here I wish to propose that science could do better with more resilient structures, more useful methodological tutorials, and clearer signaling regarding how much we can trust what it produces.

Keywords

philosophy of science, methodology, statistics

Open Peer Review

Referee Status: ✔ ?

	Invited Referees 1	2
REVISED version 2 published 01 Apr 2019		
version 1 published 31 Oct 2018	✔ report	? report

1 **Ben van Calster**, KU Leuven, Belgium
Leiden University Medical Center, The Netherlands

2 **Lincoln J. Colling**, University of Cambridge, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Jose D. Perezgonzalez (j.d.perezgonzalez@massey.ac.nz)

Author roles: Perezgonzalez JD: Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Perezgonzalez JD. **Three more steps toward better science [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2018, **7**:1728 (https://doi.org/10.12688/f1000research.16358.1)

First published: 31 Oct 2018, **7**:1728 (https://doi.org/10.12688/f1000research.16358.1)

Science has striven to do better since its inception. For example, empiricism was sought as an alternative mode of learning as early as the XVI Century (Ball, 2012); XIX Century researchers sought a less subjective approach to learning from data via frequentist statistics, which progressively displaced Bayesian inference (Gigerenzer *et al.*, 1989); in the XX Century, seeking a better way of establishing causation, Fisher (e.g., 1954) popularized a consistent framework of experimental design and frequentist inference based on small samples; Neyman & Pearson (e.g., 1928) expanded on Fisher's statistical innovations to bring about more control of research power; Jeffreys (e.g., 1961) countered with a more nuanced approach toward evidential support for hypotheses via his Bayes factor; Cohen (1988) veered the focus away from significance testing and toward practical importance with his seminal work on effect sizes and power analyses; Mayo (e.g., 2018) is nowadays popularizing a framework based on severity testing for better frequentist inference; and computational advancements are giving full Bayesian inference a new opportunity to claw back the territory lost since the XX Century (McGrayne, 2012).

Such historical drive has given us good tools for purpose, including philosophies and methodologies, as well as statistical tools for exploratory data analyses, data testing, hypothesis testing, and replication research. The path has not been easy, with a lot of effort gone onto warring among different philosophies, methodologies, and statistical approaches, and leading to troubles such as the current reproducibility crises (e.g., Fanelli, 2018).

Still, most approaches have been put forth and defended on the common goal of bettering science and, in their own way, all do so reasonably well. For example, Table 1 summarizes results obtained using different testing approaches, all concluding with similar inferences. Therefore, the real "enemy" is not what makes for better science but what makes for worse science: namely, problems with methodological control, with the misunderstanding and misuse of statistics, and with unsupported conclusions (i.e., with ethical concerns and with the use of scientific methods in a pseudoscientific manner; Perezgonzalez & Frías-Navarro, 2018).

Such enemy will be difficult to defeat. On first impression, science seems to suffer the fate of the 'tragedy of the commons', the 'free-rider dilemma' being, perhaps, its most specific affliction (Fisher, 2008). A recent book by Taleb (2018) on asymmetry sheds some light on the gaming element of science, namely on its misuse of analytical models, agency problems, asymmetric information sharing, and the rationality of the enterprise. Taleb also proposes three solutions that we could expand upon to provide a synergic path for how to go about bettering science (Perezgonzalez, 2018).

Firstly, there is a need to make 'scientific structures' more resilient, for them to deliver the outcomes they were set up for: widespread accessibility and quality control. For example, open access publishing is nowadays countering the paywall limitations of traditional scientific publishing and its bias toward novel research with significant results, thus addressing important academic and social backlashes (Kelly, 2018; Schiltz, 2018). Unfortunately, it has also motivated the rise of predatory journals catering for the same pool of conscientious researchers. To counter the explosion of these predatory journals some idiosyncratic blacklists (e.g., the defunct Beall's list) and organizational whitelists (e.g., Directory of Open Access

Table 1. Reasonable conclusions based on frequentist and Bayesian results.

Case	Cohen's d	p	Decision	SEV	BF	Evidence
I (2t)	0.20	0.507	H_0	0.75	BF_{01} = 2.85	M_0 = anecdotal
II (1t)	0.80	0.995	H_0	0.99	BF_{01} = 10.96	M_0 = strong
III (2t)	0.80	0.010	noH_0	0.99	BF_{10} = 5.04	M_1 = moderate
IV (1t)	-0.67	0.965	H_0	0.99	BF_{01} = 7.20	M_0 = moderate
V (2t)	-0.67	0.071	H_0	0.51	BF_{10} = 1.25	M_1 = anecdotal
VI (1t)	-0.93	0.999	H_0	0.99	BF_{01} = 12.08	M_0 = strong
VII (2t)	-0.93	0.003	noH_0	0.99	BF_{10} = 12.70	M_1 = strong

Notes. Based on data from Vincent (2018; Perezgonzalez & Vincent, 2018). **Case**: tests are one-tailed (1t) or two-tailed (2t). **Cohen's *d***: exploratory tests assessing effect sizes against Cohen *d* = 0.5 (i.e., the sample size—n1= 23; n2 = 23—was sensitive to *d* ≥ 0.5; Perezgonzalez, 2017). **p**: *p*-values from independent *t*-tests (Fisher's approach, e.g., 1954). **Decision**: frequentist decision—noH_0 = reject H_0; H_0 = no decision—based on level of significance = 0.05 (e.g., Perezgonzalez, 2015). **SEV**: severity tests (e.g., Mayo, 1996). **BF**: Bayes Factors with alternative model based on a Cauchy distribution (e.g., Rouder *et al.*, 2009). **Evidence**: Bayesian evidence in favor of the null model (M_0) or the alternative model (M_1; e.g., Wagenmakers *et al.*, 2018). The effect sizes of Cases II, IV, and VI had signs opposite to those expected (therefore, the high *p*'s); Cases III, V, and VII are two-tailed tests of Cases II, IV, and VI (thus, the similar *d*'s). Only Case V may lead a Jeffreysian to an inference contrary to those of frequentists; most likely, they would refrain from inferring support based on anecdotal posterior probabilities (e.g., Jarosz & Wiley, 2014).

Journals) have been created, albeit with mix success. Meanwhile, pre-print servers are challenging the entry costs of open access journals, thus making widespread communication more resilient but with the drawback of lacking good quality control.

Quality control itself has received more attention of lately, with some journals becoming more transparent about who peer-reviews, while platforms such as Publons.com provide peer-review services and credit, including access to peer-reviews when allowed. Among quality-control structures is worth mentioning F1000Research, a publication platform that sits at the fringe of a paid preprint and a fully transparent peer-reviewed open access journal. This seems a more resilient structure worthy of emulation and improvement.

Perhaps more importantly, a new need is becoming imperative: To find an effective solution to the indexing and curation of the ever expanding universe of research outputs. We do have, for example, Altmetric.com, albeit it is too geared toward scoring research outputs. Instead, what we need is an integrated solution to the indexing of both an output and all related content relevant to it, including post-publication reviews, comments in blogs and preprint servers, retraction notices, and the like. We also need a good solution to curating the entire spectrum of research outputs, moving from a plethora of stand-alone manuscripts toward mega-content organized as, for example, research topics.

Secondly, 'minority movements' do have an impact on science via creating the above new structures (e.g., open access, pre-print archives…), but also by improving on legacy ones (e.g., post-publication review sites such as PubPeer.com). Paramount among such movements have been those calling for Open Science (e.g., Banks *et al.*, 2018) and research ethics (e.g., Committee on Publications Ethics, RetractionWatch.com).

Minority movements also have an impact on other aspects of science, from calls toward a better use of frequentist statistics (Perezgonzalez, 2015) to the outright banning of *p*-values (Trafimow & Marks, 2015), to the alternative use of Bayesian statistics (Wagenmakers *et al.*, 2018b) or mixed approaches (Perezgonzalez & Frías-Navarro, 2018). Because of the intrinsic social dynamics of minority groups, the polarization of inter-group attitudes and consequential external warring are not only unsurprising but also expected. Yet, as alternative scientific approaches mostly have a different research focus, science has been less productive than it could be because more effort has been put into warring between factions than into clearly explaining what each provides to the advancement of science. This has allowed specific methodological knowledge to be too much textbook-based, thereby more aligned with editorial concerns than with the advancement of science (Gigerenzer, 2004), or to be polarized by the intrinsic dynamics of minority groups. Thus, what we presently need are good tutorials on the purpose of each approach and on how to effectively use them for such purpose; preferably, tutorials which are independently created rather than centrally edited, so as to provide a diversity of options able to address the same topic from different perspectives and to cater to different stakeholders (e.g., researchers, reviewers, and readers; novice and experts; technically-focused, philosophically-aware, as well as practitioners; etc). Such diversity will also allow for progressively developing optimal tutorials that minimize steep learning curves, capture methodological errors, and avoid philosophical and interpretive misconceptions.

Finally, there is the need to signal how much 'soul is in the game' in each piece of published research. The pre-registration movement is achieving this via badges; most journals require authors to signal adherence to ethical principles via the corresponding disclaimers; some journals actively signal their peer-reviewing by naming peer-reviewers—e.g., Frontiersin.com, F1000Research.com—or by allowing open access to peer-reviews—e.g., via Publons.com. What we are presently lacking is good signaling to address methodological concerns and the avoidance of pseudoscience. That is, for authors to signal that they have followed, for example, Fisher's approach to data testing, or Neyman-Pearson's approach, or Mayo's severity approach, or Jeffreys's approach, or a full Bayesian approach; in brief, for them to signal when their research is compliant with the requisites of any of those approaches. The purpose of this signaling is to prevent what Farrington (1961, p. 311) already denounced, that ". . . there is no human knowledge which cannot lose its scientific character when [we] forget the conditions under which it originated, the questions which it answered, and the function it was created to serve". This signaling could work in a manner similar to when authors specify a creative commons license for an open-access document: for a particular manuscript researchers could signal the specific methodological approach followed. This, of course, calls for negotiating the appropriate standards and for hosting them for quick referencing both by prospective authors and their peers.

In brief, following from the ideas of Taleb (2018), science could do better with more resilient structures, with more useful methodological tutorials, and with good signaling regarding how much we can trust what it produces. Thus my overall recommendation: let's veer the focus from warring and onto improving our structures, tutorials and signals.

Data availability

No data is associated with this article.

Grant information

The author(s) declared that no grant(s) were involved in supporting this work.

References

Ball P: **Curiosity: How Science became interested in everything.** Chicago, IL: The University of Chicago Press. 2012.
Reference Source

Banks GC, Field JG, Oswald FL, *et al.*: **Answers to 18 questions about Open Science practices.** *J Bus Psychol.* 2018; 1–14.
Publisher Full Text

Cohen J: **Statistical power analysis for the behavioral sciences**. (2nd ed.). New York, NY: Psychology Press. 1988.
Publisher Full Text

Fanelli D: **Opinion: Is science really facing a reproducibility crisis, and do we need it to?** *Proc Natl Acad Sci U S A.* 2018; **115**(11): 2628–2631.
PubMed Abstract | Publisher Full Text | Free Full Text

Farrington B: **Greek Science: Its meaning for us.** (Rev. Ed). Middlesex, UK: Penguin Books. 1961.
Reference Source

Fisher L: **Rock, paper, scissors: Game theory in everyday life.** New York, NY: Basic Books. 2008.
Reference Source

Fisher RA: **Statistical methods for research workers.** (12th ed.). Edinburgh, UK: Oliver and Boyd. 1954.
Reference Source

Gigerenzer G: **Mindless statistics.** *J Soc Econ.* 2004; **33**(5): 587–606.
Publisher Full Text

Gigerenzer G, Swijtink Z, Porter T, *et al.*: **The empire of chance: How probability changed science and everyday life.** Cambridge, UK: Cambridge University Press. 1989.
Publisher Full Text

Jarosz AF, Wiley J: **What are the odds? A practical guide to computing and reporting Bayes Factors.** *J Problem Solving.* 2014; **7**(1): 2.
Publisher Full Text

Jeffreys H: **Theory of probability.** (3rd ed.). Oxford, UK: Oxford University Press. 1961.
Reference Source

Kelly E: **EU research chief's next act: changing the future of academic publishing.** Science Business. 2018.
Reference Source

Mayo DG: **Error and the growth of experimental knowledge.** Chicago, IL: The University of Chicago Press. 1996.
Reference Source

Mayo DG: **Statistical inference as severe testing: How to get beyond the statistics wars.** Cambridge, UK: Cambridge University Press. 2018.
Publisher Full Text

McGrayne SB: **The theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy.** New Haven, CT: Yale University Press. 2012.
Reference Source

Neyman J, Pearson ES: **On the use and interpretation of certain test criteria for purposes of statistical inference: part I.** *Biometrika.* 1928; **20A**(1/2): 175–240.
Publisher Full Text

Perezgonzalez JD: **Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing.** *Front Psychol.* 2015; **6**: 223.
PubMed Abstract | Publisher Full Text | Free Full Text

Perezgonzalez JD: **Statistical sensitiveness for the behavioural sciences.** *PsyArxiv.* 2017.
Publisher Full Text

Perezgonzalez JD: **Book review: Skin in the game.** *Front Psychol.* 2018; **9**: 1640.
Publisher Full Text | Free Full Text

Perezgonzalez JD, Frías-Navarro MD: **Retract p < 0.005 and propose using JASP, instead [version 2; referees: 3 approved].** *F1000Res.* 2018; **6**: 2122.
PubMed Abstract | Publisher Full Text | Free Full Text

Perezgonzalez JD, Vincent N: **Análisis paralelos frecuentista-bayesianos con JASP [Parallel frecuentist-Bayesian data analysis with JASP].** IIn Frías-Navarro D, García-Pérez F, Pascual-Llobell J. (Eds.), Investigación en Psicología: diseño y análisis (ch.5). Valencia, Spain: Palmero ediciones. 2018.

Rouder JN, Speckman PL, Sun D, *et al.*: **Bayesian t tests for accepting and rejecting the null hypothesis.** *Psychon Bull Rev.* 2009; **16**(2): 225–237.
PubMed Abstract | Publisher Full Text

Schiltz M: **Science without publication paywalls: Coalition S for the realisation of full and immediate Open Access.** *Front Neurosci.* 2018; **12**: 656.
Publisher Full Text

Taleb NN: **Skin in the game: Hidden asymmetries in daily life.** New York, NY: Random House. 2018.
Reference Source

Trafimow D, Marks M: **Editorial.** *Basic Appl Soc Psych.* 2015; **37**(1): 1–2.
Publisher Full Text

Vincent N: **Situational awareness of pilots in the cruise.** Unpublished Master's thesis, Massey University, New Zealand. 2018.

Wagenmakers EJ, Love J, Marsman M, *et al.*: **Bayesian inference for psychology. Part II: Example applications with JASP.** *Psychon Bull Rev.* 2018; **25**(1): 58–76.
PubMed Abstract | Publisher Full Text | Free Full Text

Wagenmakers EJ, Marsman M, Jamil T, *et al.*: **Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications.** *Psychon Bull Rev.* 2018b; **25**(1): 35–57.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔ ?

Version 1

Referee Report 14 March 2019

https://doi.org/10.5256/f1000research.17868.r44358

? **Lincoln J. Colling**
Department of Psychology, University of Cambridge, Cambridge, UK

I think this paper could benefit from a minor revision.

Some of the statements in the paper are a little vague. Most of the "meat" of the paper comes in the penultimate paper. It might be worth separating out and clearly stating the specific recommendations.

Table 1 presented SEV values. However, SEV is always with respect to a **specific inference** and an observation (an observation). For example, for an observation of e.g., mu = 17, the inference mu1 = 12 and the inference mu1 = 14 would have different SEV values, but Table 1 gives no indication of what inference is being made (I assume the inference is the same as the observed result, but this should at least be indicated somewhere). It also appears that Table 1 presented the results of t-tests. I would be inclined to include a test statistic (or an N or both - with an N the reader could calculate the test stat, or with a test stat the reader could calculate the N). I think it is almost essential, because the sample size is one of the key determinants of when the various inferential frameworks actually come apart. I think in its current form Table 1 paints a slightly misleading picture.

On Page 3, Para 1: It may be worth discussing "overlay journals". These exist in physics and computing, but recently an overlay journal in neuroscience has also been launch (Neuroscience, Behaviour, Data and Theory).

Page 3, Para 5: There are some examples of these tutorial style papers: 1. Four reasons to prefer Bayesian analyses over significance testing, Dienes et al.[1] and 2. Statistical Inference and the Replication Crisis, Colling et al.[2].

References

1. Dienes Z, Mclatchie N: Four reasons to prefer Bayesian analyses over significance testing.*Psychon Bull Rev*. **25** (1): 207-218 PubMed Abstract | Publisher Full Text
2. Colling L, Szűcs D: Statistical Inference and the Replication Crisis. *Review of Philosophy and Psychology*. 2018. Publisher Full Text

Is the topic of the opinion article discussed accurately in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?

Yes

Are arguments sufficiently supported by evidence from the published literature?
Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Philosophy of Statistics, Philosophy of Cognitive Science, Neuroscience

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 25 Mar 2019

Jose Perezgonzalez, Massey University, New Zealand

Thank you very much for your review and useful pointers.

- *I think this paper could benefit from a minor revision. Some of the statements in the paper are a little vague. Most of the "meat" of the paper comes in the penultimate paper. It might be worth separating out and clearly stating the specific recommendations.*

I agree that the text is a little vague at times, albeit this is namely for the reason of lack of effective control on how the recommendations will eventually be implemented. For example, I can foresee the benefit of tutorials, in general (recommendation 2), but I cannot phantom whether there is a perfect tutorial to satisfy everyone (multiple and diverse tutorials may be needed for a "market-type" selection to kick-in, thus signalling helpful from less helpful ones). Actually, it is the third recommendation (in the penultimate paragraph) the one I see affected the least by variability in our imagination, as it calls for specific standards similar to standards already existing elsewhere (which is also why I placed it last, so to somewhat finish the commentary with a clearer, less vague, recommendation). It is for those reasons that I find it difficult to make the remaining recommendations more specific, and had to resort to the use of the conventional keywords "firstly", "secondly", and "finally" to somewhat constrain them to particular sections in the manuscript.

Still so, I also aimed to put the recommendations more clearly towards the end of their sections: better indexing and curation as Recommendation 1 (this implies software-based indexing and curation, but I have little idea whether such software will work well); tutorials (but, again, I am not sure how well the idea will work; nonetheless, I added a correction linking to the STRATOS Initiative, at www.stratos-initiative.org, which may be one of the ways forward; other could be a methodology-based overlay journal, as you mentioned).

- *Table 1 presented SEV values. However, SEV is always with respect to a **specific inference** and an observation (an observation). For example, for an observation of e.g., mu = 17, the inference mu1 = 12 and the inference mu1 = 14 would have different SEV values, but Table 1 gives no indication of what inference is being made (I assume the inference is the same as the observed result, but this should at least be indicated somewhere). It also appears that Table 1 presented the results of t-tests. I would be inclined to include a test*

statistic (or an N or both - with an N the reader could calculate the test stat, or with a test stat the reader could calculate the N). I think it is almost essential, because the sample size is one of the key determinants of when the various inferential frameworks actually come apart. I think in its current form Table 1 paints a slightly misleading picture.

I have added a new column with the *t*-test statistics and corresponding degrees of freedom. I have also extended the note on Severity to clarify it and also give a quick pointer for assessment, as *"SEV: severity tests based on the observed effects (severity is strong if greater than 0.80; e.g., Mayo, 1996)"*

- *On Page 3, Para 1: It may be worth discussing "overlay journals". These exist in physics and computing, but recently an overlay journal in neuroscience has also been launch (Neuroscience, Behaviour, Data and Theory).*

Thanks for the pointer. I wasn't aware of overlay journals. But they certainly are a pretty good solution. I have added the following statement to paragraph 1, page 3: *"*—although overlay journals are taking care of the later drawback"*.*

- *Page 3, Para 5: There are some examples of these tutorial style papers: 1. Four reasons to prefer Bayesian analyses over significance testing, Dienes et al. and 2. Statistical Inference and the Replication Crisis, Colling et al.*

I am aware of several tutorials; I even have one of my own. However, I thought it would be not too good an idea to name them as the recommendation focuses on the idea of generating them; thus pointing to some (which may or may not be useful, in hindsight) seems distracting. I have, however, added a new entry to an initiative that seems to be working on the same idea, the STRATOS Initiative).

Competing Interests: No competing interests.

Referee Report 18 February 2019

https://doi.org/10.5256/f1000research.17868.r44361

Ben van Calster [1,2]

[1] Department of Development and Regeneration, KU Leuven, Leuven, Belgium

[2] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

This is an opinion piece, so I reviewed it as such. The paper discusses a few possible directions for improving the scientific process. I largely agree with the expressed opinions. That said, I think that the text remains quite general and vague at times. For example, with respect to the second issue, what do you mean with 'tutorials which are independently created rather than centrally edited'? About the third issue: what is your specific suggestion? That researchers should better frame their work before starting it (as to what methodology and statistical approach they will use), and adhere to that when writing the report? This is unclear.

Further comments:

- In the abstract, the statement 'warring among different perspective' may need a bit more clarity.
- The paragraph where the second issue is explained, could start with 'Secondly'. (The first issue is introduced with 'firstly', the third issue with 'finally'.)

- Table 1: too concise in my opinion. On what N are the p-valus based? Are the Cohen's d values observed ones? A bit more information on severity tests might be useful. Perhaps columns that belong together (p and Decision; BF and Evidence; if I'm correct) may be put together more clearly.
- The paragraph starting with 'such enemy will be difficult to defeat' is quite vague. E.g. what do you mean with terms like 'tragedy of the commons', 'asymmetric information sharing', 'rationality of the enterprise', and others?
- Pre-print servers (line 2 of p3): do you refer to repositories like arXiv?
- Regarding scientific structures and peer-review: my issue is that peer review is very important (I am afraid that reviewers are sometimes the only people who check the contents of a study report), but it remains a largely uncredited – almost secret – endeavor. Don't you agree that it still needs to be made more official, e.g. that universities do not only demand their researchers to publish papers, but also to deliver decent peer review reports? For example, when I started my tenure track, the university told me what they expected in terms of publication output, grants, and PhD guidance, but they said nothing about peer review, and they never evaluated me on these terms. A decent peer review should almost count as much as publishing a paper?
- 'Factions' (bottom of first column on p3): what is that?
- Would the STRATOS initiative (www.stratos-initiative.org), coordinated by Willi Sauerbrei from Freiburg, be in line with the second issue? It brings together experts on different methodological topics in the context of observational studies, with the aim to provide guidance on how to address these topics. (I am a member of this initiative.)

Is the topic of the opinion article discussed accurately in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Yes

Are arguments sufficiently supported by evidence from the published literature?
Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Methodology, (medical) statistics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 25 Mar 2019

Jose Perezgonzalez, Massey University, New Zealand

Thank you very much for you review (and apologies for the long response time).

- *This is an opinion piece, so I reviewed it as such. The paper discusses a few possible directions for improving the scientific process. I largely agree with the expressed opinions. That said, I think that the text remains quite general and vague at times. For example, with respect to the second issue, what do you mean with 'tutorials which are independently*

created rather than centrally edited'? About the third issue: what is your specific suggestion? That researchers should better frame their work before starting it (as to what methodology and statistical approach they will use), and adhere to that when writing the report? This is unclear.

I agree that the text is "general and vague at times", albeit this is namely because I don't know in which form the recommendations will eventually be implemented.

Independently created tutorials calls for multiple and diverse works to be done by independent authors or groups of authors, as opposed to them being 'centrally edited' by a publisher (e.g., of textbooks on methods, or statistics). The noted sentence follows from an earlier assertion, that "methodological knowledge [is] too much textbook-based, thereby more aligned with editorial concerns than with the advancement of science (Gigerenzer, 2004)". At the time, I also found it difficult to write it better without repeating 'independently / independent authors', and 'centrally edited / textbook editors'. I have attempted it with synonyms here (albeit it may read a bit 'forced'). It now reads "…tutorials which are independently created by unfettered authors rather than centrally abridged by textbook editors…"

The third issue is a bit simpler than framing our work a priori (although it may help with this, as well). It is more like framing our work for publication. E.g., those who have followed a Fisherian approach, would say so but also write in a manner that is consistent with such approach, as the assumptions and inferential process are different to those following a Neyman-Pearson approach, or a Jeffreysian approach, etc. If no approach is clear or if they got mixed up in the process of doing the research, then no standard ought to be indicated (it would be misleading, otherwise).

The recommendation is more or less the following: In the same way we may decide to release a document under a particular creative commons license (or none), a license which we need to specify and which binds the document to it; we could also release a research report under a particular research standard (or none), and we shall specify such standard so that peer-reviewers can assess the manuscript as per compliance with those standards, and readers can understand the results in reference to such standards. This also means the standards need to be negotiated, approved and hosted as for facilitating quick referencing for the aforementioned peer-reviewers and readers (and, of course, authors).

- *Further comments:*
- *In the abstract, the statement 'warring among different perspective' may need a bit more clarity. The paragraph where the second issue is explained, could start with 'Secondly'. (The first issue is introduced with 'firstly', the third issue with 'finally'.)*

I have re- written the abstract, substituting 'warring' by "historical clashes among different perspectives, typically between frequentists and Bayesians".

The second issue actually starts with 'secondly', when I introduced the idea of 'minority movements'.

- *Table 1: too concise in my opinion. On what N are the p-valus based? Are the Cohen's d values observed ones? A bit more information on severity tests might be useful. Perhaps columns that belong together (p and Decision; BF and Evidence; if I'm correct) may be put together more clearly.*

I have added a new column to Table 1, now describing the test, degrees of freedom, and test statistic. I also noted explicitly that Cohen's *d* describes observed effects. The columns in the Table follow the suggested grouping:

- p-values and 'frequentist decision'
- Severity statistics (which is also a frequentist approach)
- Bayes factors and ‘Bayesian inference’

I also clarified a bit more severity testing.

- *The paragraph starting with ‘such enemy will be difficult to defeat’ is quite vague. E.g. what do you mean with terms like ‘tragedy of the commons’, ‘asymmetric information sharing’, ‘rationality of the enterprise’, and others?*

All those constructs are found in the book by Taleb (2018). The paragraph is meant to quickly brush over them as a quick introduction, as the really relevant constructs (also Taleb’s) follow: ‘scientific structures’, ‘minority movements’, and ‘soul in the game’.

- *Pre-print servers (line 2 of p3): do you refer to repositories like arXiv?*

Both, actually, as they often have such a dual role: either as a final repository or as a repository of manuscripts prior to them being sent for publication elsewhere (nowadays, many journals accept the latter, as long as they are in repositories / preprint servers). I have nonetheless added the concept ‘online repositories’ to the text.

- *Regarding scientific structures and peer-review: my issue is that peer review is very important (I am afraid that reviewers are sometimes the only people who check the contents of a study report), but it remains a largely uncredited – almost secret – endeavor. Don’t you agree that it still needs to be made more official, e.g. that universities do not only demand their researchers to publish papers, but also to deliver decent peer review reports? For example, when I started my tenure track, the university told me what they expected in terms of publication output, grants, and PhD guidance, but they said nothing about peer review, and they never evaluated me on these terms. A decent peer review should almost count as much as publishing a paper?*

I fully agree. In fact, I think that a decent peer-review should count as much as publishing a paper and academics could, ideally, gain tenure and the like on peer-reviewing alone, not just on teaching or researching, as this allows for good, committed peer-reviewers and increased quality standards (the problem is how to define ‘decent’!). That’s why I added an entry on ‘quality control’ under recommendation 1 for more resilient scientific structures. However, it will be very difficult to know who has peer-reviewed what and with what quality. This means that all comes down to trust: the academic on the tenure track would claim to have done ‘x’ reviews for ‘y’ journals…but it will be difficult to prove.

Thus far, I know of platforms such as Publons, PubPeer, FrontiersIn, and F1000Research, which would somewhat “reward” reviewers and/or publish peer-reviews. Of these, Publons actually rewards peer-reviewing and allows for generating some statistics in the form of percentiles and graphs. But reviews are only displayed depending on particular journal permissions. On the other hand, F1000Research does actually publish reviews and give them a doi. But there is no statistical summary of any form for reviewers. Thus, independently acknowledging / rewarding peer-review will be difficult unless the action of peer-reviewing have been somewhat independently ‘vetted’ (Publons have such control…but I am not too sure how well it works) or peer reviews are openly displayed (F1000Research). I don’t’ see Universities caring for a career in peer-reviewing any time soon, though.

- *'Factions' (bottom of first column on p3): what is that?*

Factions follow the “war” theme of Mayo’s latest book, consistent with similar concepts in the paragraph (enemies, polarization, warring, etc.). I have repeated Mayo’s reference in this paragraph.

- *Would the STRATOS initiative (www.stratos-initiative.org), coordinated by Willi Sauerbrei from Freiburg, be in line with the second issue? It brings together experts on different*

methodological topics in the context of observational studies, with the aim to provide guidance on how to address these topics. (I am a member of this initiative.)

Yes… I think that you have almost 'pulled the rug from under my feet' here. The STRATOS initiative is pretty much in line with that point. I have added the following reference to the manuscript ("*—see also the STRATOS Initiative, already working toward a similar goal, www.stratos-initiative.org*").

Competing Interests: No competing interests.