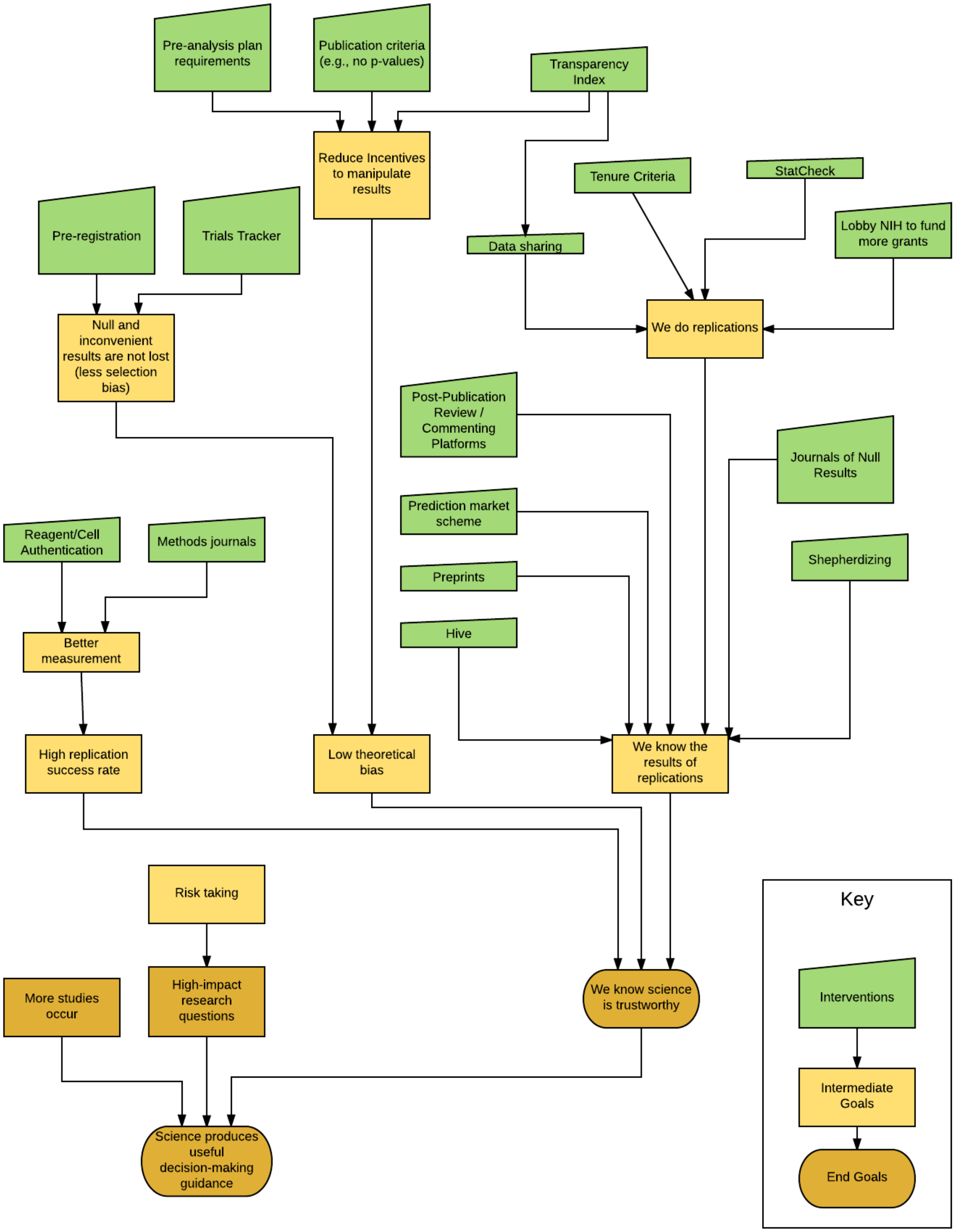
Report for LJAF on Scientific Reproducibility

Stephanie Guerra, Kathryn Rough, L. Vivian Dong, Katriel Friedman, and Eric Gastfriend

January 2016

Figure 1: Model of Factors Influencing Scientific Reproducibility


Pre-analysis plan requirements
Publication criteria (e.g., no p-values)
Transparency Index
Reduce Incentives to manipulate results
Tenure Criteria
StatCheck
Pre-registration
Trials Tracker
Data sharing
Lobby NIH to fund more grants
Null and inconvenient results are not lost (less selection bias)
We do replications
Post-Publication Review / Commenting Platforms
Journals of Null Results
Prediction market scheme
Reagent/Cell Authentication
Methods journals
Shepherdizing
Preprints
Hive
Better measurement
High replication success rate
Low theoretical bias
We know the results of replications
Risk taking
Key
Interventions
More studies occur
High-impact research questions
We know science is trustworthy
Intermediate Goals
End Goals
Science produces useful decision-making guidance

Introduction

Reproducibility is a proxy for truth. We can never know with certainty that a study/experiment has produced a true result, but if the result comes up the same every time when the study is performed by different, independent teams, then it probably has. Independent, partial reproduction of a study, such as reanalysis of the data, also strengthens the evidence.

The value of reproducibility can be characterized by a Value of Information calculation.[1] In essence, the Value of Information depends on the stakes of the decision (how bad it would be to make the wrong choice) and the probability that new information will change our decision. If we are highly uncertain and the information greatly reduces that uncertainty, then its value is greater. From a philanthropist's perspective, scientific reproducibility matters only in fields where there is high uncertainty/disagreement, many people (thousands or millions) are affected, and decision-making (by foundations or government actors) is evidence-based.

The terminology is not standardized. We use the term "verification" to describe the general category of any ex-post activity (reanalysis, validation of reagents, repeated experiments) that investigates the reliability of a result. We use the term "replication" more narrowly to refer to re-conducting an experiment and comparing the result to the original study. The term "reproduction" is often used to refer to a reanalysis of a paper's data.[2]

Roughly, science's usefulness is a function of the *number*, *reliability,* and *impact* of results. Most scientific results are less than completely reliable. Experimental errors, contaminants, unlikely statistical flukes, errors in calculations, fraud or distortion can never be totally ruled out when examining a scientific result. We use scientific results to develop new technologies and make policy choices (although not in all cases, unfortunately). Uncertainty about results leads to errors of both inclusion and exclusion: good policies that do not get adopted, useful technologies that do not get developed, bad policies that *do* get adopted, and ineffective technologies (e.g., medical or psychological treatments) that go into use.[3] In short, we think of science as *trustworthy* when results are highly reliable; *productive* when it generates many results; and *impactful* when it works on problems of great importance.

To identify the right strategy for improving reproducibility in science, ask:

1. Can we increase the reliability of results without reducing the number or average impact?
2. Are there ways of trading off a little productivity or impact for a lot of reliability (or vice versa) in a way that makes science more useful overall?

[1] Value of Information, Wikipedia, https://en.wikipedia.org/w/index.php?title=Value_of_information &oldid=736031788 (last visited Jan. 7, 2017).

[2] Replicability vs. Reproducibility, Tel Aviv University: Replicability Research Group, http://www.replicability.tau.ac.il/index.php/replicability-in-science/replicability-vs-reproducibility.html (last visited Jan. 7, 2017).

[3] In the physical sciences, developing a technology can often serve as a replication: if you build a prototype based on a novel principle and it does not work, the failure is obvious. In the life and social sciences, the "technologies" often can only be observed to "work" or "not work" through statistical inference.

Framing the replicability challenge

Figure 1 shows the determinants of reliability and how they relate to each other and to scientific productivity. It can describe science as a whole or individual disciplines. Trustworthiness and productivity, at the far right, are the main outcomes.

Moving to the left, three ingredients make science trustworthy:

1. Knowing replication results
2. Little bias in results of replications.
3. High replication success rates

(1) *Knowing replication results*

Consistent validation of results by replication is an indicator of reliability. In order to trust the results, consumers of scientific output must have access to that indicator: replications must be carried out *and* the results must be disseminated.

Interventions that drive *more* replications include:

- Norms or curiosity that intrinsically motivate scientists to conduct replications
- Funders' requirements to conduct replications as part of grants (greater access to resources as reward for conducting replications)
- Universities' promotion and tenure criteria to conduct replications (greater professional reward for conducting replications)
- Mechanisms for quantifying researchers' contributions through replications
- Greater transparency: availability of data from past experiments that enable replications
- Mass replication drives

Unfortunately, even when an adequate number of replications are being carried out, too often scientists are unaware of the results of those replications. This can have ripple effects where although a paper has been retracted, there will be no indication of the retraction in subsequent papers that cite the retracted paper, thus propagating misinformation in the literature.[4]

Furthermore, many valid replications may never be published due to the bias towards novel and positive results -- this is especially problematic for failed replications.

[4] Cat Ferguson, More evidence scientists continue to cite retracted papers, Retraction Watch (2015), http://retractionwatch.com/2015/02/18/evidence-scientists-continue-cite-retracted-papers/.

Interventions that improve *awareness* of replications include:

- Creation of publishing platforms for replication studies
- Improvements to information management systems (journal databases, web annotation platforms) that link studies to replication studies.
- Post-publication review
- Meta-analysis
- Tracking pre-registered replications.
- Better identification of retracted and contradicted papers

(2) *Unbiased results*

The replications must also be *valid* indicators of confidence: if there is a large bias towards confirming (or rejecting) earlier results, they contribute much less information.[5] We can't observe the bias in replications directly, and opponents of mass replication projects often argue that it is large. A common objection hinges on difficult-to-replicate "moderators" that teams who normally work on other issues lack the expertise to incorporate appropriately.[6]

Replication studies face the same sources of bias as original studies. Researcher decisions and selective publication of results lead to bias. Since most (almost all) published studies find a significant effect, the same practices (p-hacking, uncorrected multiple comparisons, interpretation of ambiguous cases) that lead to a bias toward significant results may also apply to the replication studies. These are amplified by potential confirmation bias if the replicators believe the original result to be true. However, they are potentially offset or reversed by incentives to overturn a well-established result.[7]

The main levers for influencing researchers' analysis and design decisions are:

- Improving researchers' knowledge of how bias gets introduced
- Promoting norms against introducing bias
- Greater transparency that increases the risk of being caught using bad methods
- Shifting incentives to favor methods that don't introduce bias (requiring or forbidding certain kinds of analysis, requiring pre-analysis plans, indexes/scorecards that reward researchers whose work is demonstrably reproducible)

[5] A canonical example of confirmation bias in follow-up studies is the gradual upward revision of estimates of the charge of an electron over the course of many decades.

[6] Lisa Feldman Barrett, Op-Ed: Psychology is not in Crisis, *New York Times* (2015), https://www.nytimes.com/2015/09/01/opinion/psychology-is-not-in-crisis.html.

[7] This is the allegation in the case of the 3ie re-analysis of a well-known experiment on the effect of deworming programs on schooling.

The main levers for reducing the suppression of null or inconvenient results are:

- Pre-registration of clinical trials
- Promoting norms of disseminating null results
- Results-free review
- Establishing and promoting platforms for sharing null results
- Developing processes that retrieve information about experiments that were not disseminated

The amount of noise *qua* noise in replication studies doesn't affect the level of confidence the body of replications implies for science or a scientific discipline as a whole. However, under realistic assumptions about the ratio of null to non-null results in the original literature, noise in individual replication studies generates bias towards "failure to replicate." Then, improvements to the accuracy of replication do lead to greater confidence about science as a whole.

(3) *High replication rates*

In addition to the channels described above as reducing bias, trustworthy science requires relatively low noise and consequently sound methods. Peer review is the traditional methodological gatekeeper across the sciences.

However, in some areas the professional consensus holds that certain methodological shortcomings are more or less tolerable. As a result, peer review doesn't challenge studies that would be expected to be relatively noisy. Prominent examples are, for instance, uncorrected multiple comparisons in the social sciences, underpowered experiments in psychology, and weak quality control of reagents in the life sciences.

Tradeoffs with productivity

Many of the interventions mentioned above aim to induce researchers to dedicate more time to various activities (preparing data and code for sharing, conducting additional review, conducting replications) that do not directly contribute to the volume of scientific output. More subtly, spending time documenting and preparing null results for publication could detract from time available to work on experiments of higher potential impact.[8] Changes to tenure criteria or a scholar-level index could plausibly reduce risk-taking by making high impact studies, which are less likely to replicate perhaps because they are groundbreaking, less rewarding from a career point of view.

The most prominent study to quantify the costs of irreproducibility found that the US spends $28B per year on unreproducible life sciences research; however, it did not attempt to estimate

[8] Which would be offset by publications' positive effect on productivity by improving selection of research topics.

an economically optimal level of irreproducibility.[9] We are not aware of quantitative estimates of the importance of these tradeoffs.

Interventions

In this report, we considered four broad interventions: post-publication peer review, quantitative metrics for transparency (e.g. “transparency index”), tenure criteria, and characterizing dependency relationships between pieces of research. All of these categories can be implemented in different ways, and can improve reproducibility through multiple channels. Post-publication peer review makes it easier to know the results of publications (e.g. in comments on PubMed Commons), and it also could reduce bias by creating a disincentive for scientists to manipulate results (for fear of facing more scrutiny). Quantitative metrics for transparency could change the culture in science towards more open practices, which would also open up more scrutiny of results as a disincentive to introduce bias, but also make it easier to produce replications (e.g. by having access to data and code). Tenure criteria similarly could create a culture shift, by requiring professors to perform replications and/or share data, but it seems difficult for philanthropy to have a direct effect in this area. And characterizing relationships between articles in scientific research can make research more efficient and accurate by allowing researchers to track unintuitive, tertiary effects of a particular finding in a paper.

Our recommendations are:

1. Fund and organize a conference to brainstorm and define a reproducibility metric, possibly a transparency index,
2. Approach [redacted], a stealth-mode startup seeking to quantify the reproducibility of observations in the biological sciences, with an offer of funding in exchange for more open availability of their product,
3. Fund the Global Biological Standard Institute to develop training modules for scientists and track adoption of better practices in the biomedical community that will improve reproducibility,
4. Speak with the team at Casetext, a legal innovation startup, on how their model of crowdsourcing the job of characterizing relationships between cases could be applied to science.

I. Intervention 1: Post-publication review

As outlined previously, one of the key causes of poor research integrity is bias: selective reporting, cherry picking of results, insufficient peer review, lack of internal validity, and lack of external validity (generalizability). These issues define the scope of the overall problem within the research community, namely a system that rewards flashy results rather than scientific rigor.

[9] Leonard P. Freedman, Iain M. Cockburn & Timothy S. Simcoe, *The Economics of Reproducibility in Preclinical Research*, 13 PLOS Biology (2015), http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002165.

Additional levels of scrutiny beyond traditional peer review may prove useful in combatting and incentivizing researchers to produce more robust reports.

Post-publication peer review, a process by which the work of the scientific community is opened to public review following its publication, is one mechanism for such additional scrutiny. This process can take on many forms and the reviewers of the work often utilize their own experiences, experiments, and expertise to complete the review.[10]

Models of post-publication peer review include formal review by an invited reviewer, formal review by a volunteer reviewer, and comments on affiliated blogs and journal databases independent of formal review. The first two models are publisher-driven and require the cooperation of a journal or organization to distribute published comments made by the invited or volunteer reviewer. Two concrete examples of this model include review articles written for traditional journal outlets as well as newer platforms such as the Faculty of 1000 (F1000) service, which publishes recommendations of biomedical articles via the opinions of their hand-picked faculty. The third model involves public commenting platforms, which come in a number of flavors depending on the source of comments, and can be anonymous or identified and can also come from a fellow scientist or a non-expert member of the public. A commenting platform is more open in nature, but with such transparency comes a number of challenges including active participation from the research community, the amount of expertise of the reviewers, and the possible fragmentation of the post-publication review discussion. These challenges can have societal costs, but may also be overcome with the right platform. More analysis of these challenges is presented in the following pages (see PubPeer, PubMed Commons) with a focus on commenting platforms due to their ubiquitous and open nature.

Beyond purely commenting on the potential validity of a report, an additional mechanism for post-publication peer review involves actually replicating these results in the laboratory. There are a number of organizations already funded by LJAF that make this strategy a priority. There are also additional organizations that aim to facilitate direct replication at a grassroots level within the community (such as Bio-Protocol).

Finally, another way in which scientists themselves can be reviewed following publication is to follow a name-and-shame model in which researchers are rewarded for high-quality research practices and punished for poor acts. Many organizations aim to enact a public incentivization program within the community (see GBSI). Additionally, creating a public platform where retracted reports and conflicting studies are more easily visible to the research community may prove beneficial (see part IV).

As of 2016, there has been no formal evaluation of the impact of post-publication peer review on research integrity but anecdotally, there have been instances where review by the community following publication of high profile papers have led to subsequent retraction of conflicted

[10] Eva Amsen, What is post-publication peer review?, F1000 Research Blog, F1000 Research (2014), http://blog.f1000research.com/2014/07/08/what-is-post-publication-peer-review/.

reports.[11] With respect to the "openness" component of post-publication review, a randomized trial of signed (and otherwise conventional) peer review by the British Medical Journal found no effect on review quality of sharing reviewers names with authors.[12] The EMBO Journal also reported no change in quality of reviews after posting review material online (but anonymously) [13]. Together these results suggest provisionally that openness in the review process will not greatly enhance the scrutiny given to results by invited reviewers.

In our analysis, we have identified four distinct existing organizations with different approaches and activities that enhance the post-publication peer review process. These include PubMed Central, PubPeer, and the Global Biological Standards Institute.

PubMed Commons

Two primary challenges to commenting platforms as a means of post-publication peer review involve maintaining the quality of the reviews and also guarding against the possible fragmentation and subsequent obscurity of the reviews themselves. PubMed Commons, a post-publication peer review system addresses both of these issues in its design. According to the The Scientist, the platform "allows users to comment directly on any of PubMed's 23 million indexed research articles, much in the way people review films on Rotten Tomatoes, evaluate restaurant service on Yelp, or grade purchases made on Amazon."[14]

The Commons was created in 2014 by several scientists in collaboration with the National Center for Biotechnology Information's (NCBI) biomedical research database, PubMed Central. PubMed houses 23 million research articles and is highly utilized by the research community with approximately 2.5 to 3 million people accessing the database each day according to David Lipman, the director of NCBI.[15]

Comments on the platform are non-anonymous and require users to be registered with the NCBI's online database (i.e. have authored PubMed-indexed papers), in an attempt to mitigate commentary and criticism from non-experts and to protect against possible ad hominem attacks which are more common on other anonymous platforms (blogs, see PubPeer). However, lack of anonymity also creates a culture where less established researchers would be more hesitant to post valid concerns and criticisms for fear of professional repercussions. Furthermore, this particular platform feature may discourage high usage of the commenting system. Since high

[11] The watchers on the Web, The Economist (2016), http://www.economist.com/news/science-and-technology/21709523-court-case-may-define-limits-anonymous-scientific-criticism-watchers.

[12] Susan van Royen *et al*., "Effect of open peer review on quality of reviews and on reviewers'recommendations: a randomised trial," *BMJ* 318 (1999), http://www.bmj.com/content/318/7175/23

[13] Bernd Pulverer "Comment: Transparency showcases strength of peer review," 468 *Nature* 7320, 29-31 (2010), https://depts.washington.edu/bhdept/Pulverer2010_Nature%7BEMBO_PR_Transparency%7D.pdf

[14] Aimee Swartz, Post-Publication Peer Review Mainstreamed, The Scientist (2013), http://www.the-scientist.com/?articles.view/articleNo/37969/title/Post-Publication-Peer-Review-Mainstreamed/.

[15] Aimee Swartz, Post-Publication Peer Review Mainstreamed, The Scientist (2013), http://www.the-scientist.com/?articles.view/articleNo/37969/title/Post-Publication-Peer-Review-Mainstreamed/.

participation from researchers is a key necessity for the success of these systems, PubMed Commons may face this additional hurdle.

“Comment boxes,” the submission format for PubMed Commons and PubPeer, may not be the most effective format for sharing information responding to specific points in a long article, according to a perspective we heard from a grantmaker working in scholarly publications.[16] Online publishing opens up the possibility of separating and recombining experiments within articles, and providing line-by-line annotation.

Another beneficial feature of this particular platform is its central nature. The literature database system is already in wide use and centralized. This makes the commenting system more attractive and seamlessly integrated into a researcher’s daily life. But do the statistics show that this system is working up to its potential?

The Commons has 10,632 members through October 31, 2016 who have posted 5,739 comments to 4,595 publications over the course of its two-year history. PubMed also reported that many of the conversations sparked in the Commons platform carried on in other venues such as Twitter and other social media platforms. In its pilot year of 2015, users posted approximately 4,000 comments to more than 3,300 publications. In 2016, the number of new comments dropped to approximately 1,700 on only 1,300 publications. Thus, there was a large drop-off in the number of comments and the number of publications just one year out of the platform’s pilot year. This does not signal great success of this program. Indeed, in its pilot year, the response rate to comments on the platform was under 10%.[17]

[Redacted]

At this time, we think the usage and growth of the platform is not robust enough to warrant funding. No funding or grants recommended at this time until more information can be gleaned from the NCBI about what additional funding would be used for on this platform.

Global Biological Standards Institute

One key hurdle in establishing best practices for research integrity within the scientific community is incentivizing researchers to adopt them. These practices such as providing detailed protocols and releasing raw data requires time, effort, and trust in fellow scientists. Currently, there are very few incentives in place to encourage adoption. The main levers of influence within academic science are funding agencies, journal publications, and university officials. In order to encourage research integrity, a top-down or bottom-up approach can be taken. Namely, grassroots adoption at the researcher level can spring up from the bottom, or

[16] Information from a phone conversation [redacted].

[17] PubMed Commons Team, PubMed comments & their continuing conversations, PubMed Commons Blog (2016), https://pubmedcommonsblog.ncbi.nlm.nih.gov/2016/11/21/pubmed-comments-their-continuing-conversations/. This information was corroborated in an email exchange with the PubMed Commons help desk on Nov. 16, 2016.

one or more of the main levers can require adoption from their constituents. The Global Biological Standards Institute (GBSI) takes both approaches and impressively maneuvers among the major levers and other key stakeholders including the press and researcher population. Established in 2013, GBSI is a non-profit organization with a mission to foster a scientific community in which best practices surrounding research integrity are not only accepted but also expected in order to succeed. They further their mission in three program areas: Science; Education and Training; and Policy and Awareness.[18]

GBSI has a strong track record of influencing key stakeholders. One example that highlights their ability to advocate for research integrity is their large collection of research articles and surveys that estimate the economic and cultural impact that poor research practices have on the community.[19] One of these high-profile papers estimated that $28B dollars of research funding is wasted each year on irreproducible research.[20] [Redacted] The very same morning that the GBSI article hit the press, NIH made a major notice on their new initiative on enhancing reproducibility through rigor and transparency.[21] This notice included wide-ranging changes to the funding process, reagent authentication guidelines, and a new website for reproducibility efforts with the NIH. [Redacted] In fact, one of the key tenets of these new guidelines, the authentication of reagents such as cell lines, has been a major cause of GBSI since its inception. This highlights the influence that GBSI practices have over this major funding agency. Much of this influence is attributed to their connections within the NIH as well as the scientific credibility of their leadership and board members.

One mechanism for insisting on best practices within the community is establishing accountability for guidelines provided by organizations such as the NIH. In that spirit, GBSI launched the Reproducibility2020 campaign as “a challenge and action plan for the biomedical research community to significantly improve the quality of pre-clinical biological research by year 2020.” The main mission of this project is to codify key practices for research integrity including setting community-wide standards for reagent validation, promoting the sharing of data and protocols via innovative platforms, and improving training of scientists via training modules.[22] A key part of Reproducibility2020 is to measure the success of this action plan from year to year to hold researchers accountable for their progress. In that spirit, GBSI plans to issue annual reports on the progress toward these priorities. The first report is due in early 2017 where it will be presented at the AAAS meeting in February framed as an update on the NIH Reproducibility and Rigor standards one year later. Elements of this report will include a status

[18] *See generally* http://www.gbsi.org. We also learned this from a phone call with Leonard Freedman (President), Rosann Wisman (Executive Director), and Michael Byrne (Director of Strategic Program Development) of GBSI, conducted on Dec. 5, 2016.

[19] Available at GBSI.org Current Publications, Global Biological Standards Institute, https://www.gbsi.org/publications/.

[20] Leonard P. Freedman, Iain M. Cockburn & Timothy S. Simcoe, *The Economics of Reproducibility in Preclinical Research*, 13 PLOS Biology (2015), http://journals.plos.org/plosbiology/article?id=10.1371 /journal.pbio.1002165.

[21] National Institutes of Health, Enhancing Reproducibility through Rigor and Transparency, https://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-103.html.

[22] Reproducibility2020, Global Biological Standards Institute, https://www.gbsi.org/work/reproducibility2020/.

report on all key stakeholders including journals, private funders, and pre-print servers. GBSI is looking for grants to cover the costs of measuring project success including two surveys per year, annual reports, and ongoing advocacy to ensure that research integrity and reproducibility remain a central focus of researchers, policy makers, funders, and journals. The estimated cost of Reproducibility2020 is approximately [redacted] until the year 2020.

Besides Reproducibility2020, GBSI is also in the process of creating training modules under the umbrella of its Good Research Practice (GRP) project. This project involves a "baseline assessment of the current state of knowledge and skill level across the fundamental components of research fidelity" via a number of surveys within the community. The second part of this project involves developing interactive training modules to address key areas of biomedical research. The first module on cell line authentication will be available in Spring 2017. GBSI is in the process of creating additional modules and aims to release six more the community in 2017. The cost of building each training module is approximately [redacted].

This non-profit organization currently has one R25 grant from the NIH for their cell line authentication project [redacted]. Other funding sources include [redacted].

Recommendation: Through multiple discussions with GBSI leadership, the team appears extremely capable of running successful programming. Additionally, their large publication record and history of interventions (namely the Cell Line Authentication Alliance) indicates a strong track record of influence. **We recommend investing in the Reproducibility 2020 project as the goals of this project are most aligned with LJAF's history of supporting large-scale projects with the potential to make paradigm-shifting changes.**

II. Intervention 2: Quantitative metrics for reproducibility

While many agree that increases in the transparency and reproduction of research findings would be beneficial to science, there is less agreement on concrete steps that can be taken to improve scientific behaviors. Norms for best practices, including the Center for Open Science's TOP guidelines,[23] have been proposed previously, yet there is a lack of incentive for researchers to adopt them, unless required by a specific publication or funding institute. Therefore, we explored the creation of a quantitative metric to measure best practices in reproducible and open research as a possibility for incentivizing better research behavior.

Currently, there is no widely-accepted metric for measuring best practices in reproducible and open research. As a result, it is difficult to objectively compare the performance of institutions, journals, researchers, or individual publications. There is evidence that quantitative metrics to measure "impact," such as the Thomson Reuters Impact Factor, h-index, and i-10 index, are

[23] The Transparency and Openness Promotion Guidelines, Center for Open Science, https://cos.io/top/#summary. The guidelines were initially published in *Science*. Brian Nosek et. al., *Promoting an open research culture*, 342 Science 1422 (2015), http://science.sciencemag.org/content/348/6242/1422.full.

influential on researcher behavior.[24] This is unsurprising, given that these metrics are frequently used by hiring, tenure, and grant review committees as a way to compare candidates. The goal of this intervention to harness the same competitiveness and social status ranking that drives organizations to care about impact factors and use it to encourage greater transparency and reproducibility in science. There are three channels by which the intervention effects this outcome. We refer to a transparency score/ranking below, but the general idea of a quantitative metric could incorporate practices that go beyond just transparency—we discuss this under **Specific Proposals**.

(1) *Encourages the relevant actors to adopt practices calculated towards their score/ranking*

The most obvious route through which this intervention can encourage better scientific practices is by making researchers, institutes, and/or journals adopt the specific practices that are calculated towards their transparency score. The primary incentive effect depends on the thing being indexed—indexing papers, research institutes, and researchers themselves would make researchers become more transparent, indexing journals' editorial policies and practices would make journals become more transparent. There are also secondary effects; indexed journals may reject certain articles because their score might take a hit, which makes researchers eventually share more data so that journals are likelier to accept them.

Reasons why entities would care for the ranking in the first place so as to adjust their behavior include prestige, pride, funding, and hiring/tenure committee decisions. Therefore, even if the ranked entity doesn't believe in the concept of the transparency index/ranking, as long as relevant community members do so, it may adjust its behavior for self-interested reasons. Achieving that community buy-in is a serious obstacle to this idea that a later section discusses more in-depth.

Assuming for now that such a transparency ranking could have some cachet within scientific communities, we can look to impact factors as a comparison. Despite widespread criticism,[25] a journal's impact factor is still considered essential, possibly determinative, to its reputation. The way that Thomson Reuters calculates journals' impact factors therefore incentivizes journals to act a certain way. Some responsive practices may be good, like preferencing papers that would attract a higher number of citations, an admittedly flawed proxy for interest to the scientific community. Some responsive practices may be bad. For example, to boost their "numerator," journals engage in soliciting coercive citations.[26] The aggressive pursuit of seemingly groundbreaking articles has also arguably exacerbated the replicability problem through selective publishing of positive results. Relatedly, a journal's impact factor is correlated with its retraction rate.[27]

[24] *See generally* part III of this report.

[25] *See, e.g.*, James Wilsdon & Ismael Rafols, Just say no to impact factors, *The Guardian* (2013), https://www.theguardian.com/science/political-science/2013/may/17/science-policy.

[26] Allen W. Wilhite & Eric A. Fong, *Coercive Citation in Academic Publishing*, 335 Science 6068 (2012), http://science.sciencemag.org/content/335/6068/542.

[27] Ferric C. Fang & Arturo Casavadell, *Retracted Science and the Retraction Index*, 79 Infect. Immun. 10 (2011), http://iai.asm.org/content/79/10/3855.full.

(2) *Enables resource-constrained readers to more efficiently discern between works*

Even if the ranking prompts no changes in behavior for the relevant actors, it can still promote greater scientific integrity by allowing researchers to better evaluate the trustworthiness of their sources. For example, the RetractionWatch “retraction leaderboard” likely does not provide any additional deterrence for researchers who routinely commit fraud or make serious errors.[28] However, the information is useful for researchers, who are aware that certain authors are not to be trusted.

(3) *Generates awareness, dialogue around transparency practices*

Beyond just incentivizing entities to adopt specific practices that are calculated towards their transparency score, we expect that the publication of a quantifiable metric and accompanying ranking for entities will generate more awareness around the importance of transparency in science. For better or for worse, people tend to debate and discuss rankings in a particular concept more than the concept itself.[29] Any score or ranking assessing something as qualitative as “transparency” will be deficient in some respect. Hopefully, the criticism this generates will spur debate over the qualities of various entities and also the relative importance of specific practices in protecting the integrity of the scientific process.

[28] Ivan Oransky, The Retraction Watch Leaderboard, Retraction Watch (2015), http://retractionwatch.com/the-retraction-watch-leaderboard/.

[29] *See, e.g.*, Phil Baty, Rankings' role in MENA up for debate at Going Global, Times Higher Education (THE) (2015), https://www.timeshighereducation.com/world-university-rankings/news/mena-up-for-debate-at-gg

Specific proposals

Transparency Index

In 2012, Ivan Oransky, who founded Retraction Watch and serves as the global editorial director at MedPage Today, wrote several articles outlining his idea for a metric.[30] Dubbed the "Transparency Index," the idea was to create a journal-level metric that "will signal to the scientific community how willing editors and publishers are to share how they make decisions."[31] The original proposals outlined ideas for the types of components a metric might contain (including journal review protocols, editorial transparency, handling of conflict of interest disclosures, requirements for data sharing, etc.), but did not list exact criteria or specify how they would be measured.

Since the two articles appeared in 2012, nothing on the topic has been subsequently published. We contacted Oransky and had a conversation with him over the phone about his proposal for a Transparency Index. While he has not pursued the development of the index further, he was still enthusiastic about the idea. [Redacted.]

Oransky emphasized that a metric should be created thoughtfully. His strong suggestion was to convene a workshop with multiple stakeholders (including scientists, journal editors/publishers, transparency activists, and funders) to move the idea forward. [Redacted.]

Two big organizational questions are key to the successful implementation of a transparency index. First, what criteria make up the index? Second, what entity administers the index? For the first question, Oransky emphasized the need to consult with researchers who currently review articles and/or sit in editorial positions. He strongly suggested that LJAF fund a conference where such researchers could meet up to discuss this intervention if LJAF was interested in pursuing this further. There are some obvious key criteria like pre-registration and availability of information on editorial responsibility, but consultation would be necessary to find some important, overlooked criteria, especially those that would create real variance among the major journals. The Center for Open Science's Transparency and Openness Promotion Guidelines could also form criteria for a transparency index.[32]

The second question is key to the success of the transparency index in effecting change, as the index's success is so dependent on community buy-in. An independent, reputable group ought to administer the index. The group would most likely non-profit, but could potentially be for-profit, as Thomson Reuters is. [Redacted.]

[30] Adam Marcus & Ivan Oransky, Bring on the Transparency Index, The Scientist (2012), http://www.the-scientist.com/?articles.view/articleNo/32427/title/Bring-On-the-Transparency-Index/; Ivan Oransky, The Retraction Watch Transparency Index, Retraction Watch (2012), http://retractionwatch.com/the-retraction-watch-faq/transparencyindex/.
[31] *See supra* note 33.
[32] *See supra* note 26.

Unfortunately, we do not know of any organization that is an obvious contender to administer the index, nor of any organizations that have seriously looked into this idea. We thought of the Committee on Publication Ethics (COPE) as a potential candidate—a COPE-administered index would certainly have a lot of cachet, as COPE has thousands of members and is made up of journal editors.[33] However, from browsing their website it seems COPE is concerned with a narrowly tailored definition of academic misconduct—they have not published anything in-depth about concerns with replicability. COPE is an interdisciplinary organization, while the replicability crisis is only really applicable to the social and natural sciences. But organizations adapt to different needs. COPE may be interested in pursuing this intervention.

Additional potential barriers include difficulties with implementation, validation, scalability, automation, and adoption, none of which were addressed in the original proposals for the Transparency Index.

Recommendations: Individuals we interviewed showed great enthusiasm for a transparency index, often with insightful suggestions about what might make the metric more successful. [Redacted.]

As a result, **we recommend and endorse the suggestion to convene a workshop with a group of relevant stakeholders to work towards the development of a Transparency Index, or similar metric**. This investment would be relatively low cost and low risk, but has the potential to make a substantial impact if it results in the creation and adoption of a metric to measure best practices in research reproducibility, transparency, and integrity.

Transparency Index variations: similar scores for articles, researchers, or research institutes

[Redacted.]

We think ranking research institutions by transparency would theoretically be promising. Given the variance in retraction rates and the underlying drivers for retraction observed between countries,[34] it seems reasonable to conclude that there are almost certainly environmental and cultural factors that contribute to the incidence of academic misconduct among researchers. These factors would vary between institutions, and even departments within institutions. Evaluating and ranking institutions could incentivize the changes necessary to improve cultures currently unamenable to research integrity.

The feasibility of the idea depends on which criteria are ultimately included in the ranking. The accessibility and objectivity of criteria included in these hypothetical indexes will strongly factor into the feasibility of creating a Transparency Index for journals, articles, researchers, or institutions.

[33] *See generally* About COPE, Committee on Publication Ethics, http://publicationethics.org/about.

[34] Ivan Oransky, Which countries have the most retractions, for which reasons?, Retraction Watch (2014),http://retractionwatch.com/2014/05/15/which-countries-have-the-most-retractions-for-which-reasons/.

OSF Badges: for papers

The Center for Open Science created badges for journals to use to denote publications that meet certain transparency criteria as part of its Open Science Framework.[35] Currently they have three badges: for open data, open materials, and pre-registration. Though a badge is not a score, it shares the same function—provides quickly accessible information on a publication's transparency. As the Open Science Framework is such a well-known organization in this sphere, some publications have already adopted the badges:[36]

- AIS Transactions on Replication Research
- American Journal of Political Science
- Clinical Psychological Science
- European Journal of Personality
- Internet Archaeology
- Journal of Social Psychology
- Journal of Research in Personality
- Language Learning
- Psi Chi Journal of Psychological Research
- Psychological Science
- Social Psychology

Psychological Science also has a list of its authors who have received badges.[37]

Though we thought these badges could start a great trend, money cannot add much value here anymore. The goal of OSF for these badges seems to be more widespread adoption by journals.

P-curves: for researchers, journals

We spoke with [redacted] at the Harvard Business School early in the fall about his thoughts concerning replicability. One concept he mentioned was the *p*-curve. The statistical concept was pioneered by Uri Simonsohn, Leif Nelson, and Joseph Simmons, who describe it as such:

> *P*-curve is the distribution of statistically significant p values for a set of studies (*p*s < .05). Because only true effects are expected to generate right-skewed *p*-curves—containing more low (.01s) than high (.04s) significant *p* values— only right-skewed *p*-curves are diagnostic of evidential value. By telling us whether we can rule out selective

[35] Badges to Acknowledge Open Practices: Openness is a core value of scientific practice, Open Science Framework, https://osf.io/tvyxz/wiki/home/.
[36] Badges to Acknowledge Open Practices: Adoptions and Endorsements, Open Science Framework, https://osf.io/tvyxz/wiki/5.%20Adoptions%20and%20Endorsements/.
[37] Authors Leading the Way in Open Science, Association for Psychological Science, http://www.psychologicalscience.org/publications/psychological_science/badge-earners.

> reporting as the sole explanation for a set of findings, *p*-curve offers a solution to the age-old inferential problems caused by file-drawers of failed studies and analyses.[38]

This simple statistical concept can be applied to any body of work large enough such that there are enough studies with significant results from which to derive a p-curve. This includes journals and some extremely prolific researchers. Some researchers have already applied the concept to make inferences concerning the prevalence of p-hacking in large bodies of research.[39]

Though Simonsohn et. al. held out the p-curve to be applicable to both experimental and observational research, Stephan B. Bruns and John Ioannidis have criticized the use of p-curving for the latter.[40] They showed that with observational research featuring minimal omitted variable bias, p-curves based on true effects and p-curves based on null effects with p-hacking cannot be reliably distinguished. They state that the same problem may persist for non-randomized experimental research. Nonetheless, Bruns and Ioannidis do not conclude that the p-curve ought to be wholesale abandoned for observational research, just that researchers generating a p-curve must empirically calibrate their results. As Simonsohn et. al.'s paper was only published in 2014, additional research will likely lead to refinements of the *p*-curve. But the central concept remains promising for its elegance. P-curves could be generated automatically, unlike a sophisticated transparency index, so the cost of labor of creating a p-curve based metric would be near-zero.

An additional weakness we thought of was that If p-curving is used to attribute a score to a particular researcher, it can deter researchers from pursuing some worthwhile, more ambiguous research questions because such research is less likely to turn up low p-values. Pursuing such research routinely may skew leftward the researcher's p-curve, opening herself to accusations of p-hacking. If researchers feel there is nothing they can do to improve their p-curves, publishing p-curves will have limited effects. If complemented by new channels for reporting null results, *p*-curves may be relatively more effective, because they will then provide researchers working on these ambiguous areas a way to improve their scores (offsetting the losses in high-impact publications due to reducing p-hacking behaviors).

[Redacted]

III. Intervention 3: Tenure Criteria

In higher education, being awarded tenure grants a professor permanent employment at his or her institution, protecting against unjust dismissal. Tenure policies were originally created to

[38] Uri Simonsohn et. al., *P-Curve: A Key to the File-Drawer*, 143 Journal of Experimental Psychology 2, 534 (2014), http://pages.ucsd.edu/~cmckenzie/Simonsohnetal2014JEPGeneral.pdf.

[39] *See, e.g.*, Megan L. Head et. al., The Extent and Consequences of P-Hacking in Science, 13 PLOS Biology (2015), http://journals.plos.org/plosbiology/article/file?id=10.1371/journal.pbio.1002106&type=printable.

[40] Stephan B. Bruns & John Ioannidis, *p-Curve and p-Hacking in Observational Research*, 11(2) PLOS One (2016), http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0149144.

ensure academic freedom of speech and thought, especially for professors holding views that were controversial or unpopular. In general, faculty are reviewed for tenure after working at the institution for a set period of time; those not awarded tenure are typically dismissed from the institution. Tenure is highly sought-after for academics due to the job security and prestige it endows, but the number of tenured faculty positions is small compared to the number of academics, making it a highly competitive process.

Because tenure considerations are an important driver of academic behavior, it is an intuitively appealing area for an intervention to improve research reproducibility; changes to tenure criteria could be one way to reward and incentivize practices of open research. Currently, evaluation of faculty for tenured positions is typically based on three broad "pillars" of academic contributions: research, teaching, and service. While nearly all universities consider the three pillars in their tenure criteria, how those three concepts are conceptualized, measured, and rewarded is highly heterogeneous. As a result, tenure criteria and policies vary dramatically between institutions.

Issues with tenure criteria and possibilities for improvement

At a number of institutions, both the number of publications and the prestige of the journals in which they appear is either implicitly or explicitly rewarded in tenure policies.[41] This reward system result is a widely-recognized "publish or perish" mentality among faculty.[42] Several prominent figures, including the director and principal director of the National Institutes of Health,[43] have described tenure criteria as being among the causes of the observed lack of reproducibility of many fields. Journal editors also seem to agree; [redacted] identified the need to publish in high-impact journals to satisfy tenure criteria as a strong incentive to selectively report or manipulate results.[44] [Redacted.]

There has been particularly pronounced backlash against the widespread use of Thomson Reuter's Impact Factor in the tenure decision-making process. The San Francisco Declaration on Research Assessment (SF-DORA),[45] initiated by the American Society for Cell Biology, is a statement denouncing the use of the Impact Factor to compare research output across individuals in funding, hiring, and tenure decisions. SF-DORA asserts that the Impact Factor is especially ill-suited for this purpose, explaining that it was "originally created as a tool to help librarians identify journals to purchase, not as a measure of the scientific quality of research in an article" and identifying numerous problems with the measure: citation distributions within journals are skewed, the measure can be easily manipulated through editorial practices, and the

[41] Chiara Franzoni, Giuseppe Scellato, & Paula Stephan, Changing Incentives to Publish, 333 *Science* 6043, 702 (2011), http://science.sciencemag.org/content/333/6043/702.
[42] Seema Rawat & Sanjay Meena, *Publish or perish: Where are we heading?*, 19(2) Journal of Research in Medical Sciences 87 (2014), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3999612/.
[43] Francis S. Collins & Lawrence A. Tabak, *NIH plans to enhance reproducibility*, 505 *Nature* 7485, 612 (2014), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4058759/.
[44] Information from a conversation with [redacted].
[45] San Francisco Declaration on Research Assessment, *available at* http://www.ascb.org/files/SFDeclarationFINAL.pdf?x30490. See more information at http://www.ascb.org/dora/.

calculation of the metric is not transparent. The SF-DORA document currently has over 12,000 individual & 900 organizational signatories.

There is evidence that researchers consider Impact Factor in their decision-making processes, which may arise from pressure to conform to tenure committee expectations. Researchers are more likely to cite articles when they appear in high Impact Factor journals,[46] and are more likely to preferentially submit manuscripts to journals with a high Impact Factor.[47] [Redacted.]

Therefore, one potential mechanism to improve tenure criteria is the creation of a quantifiable metric for rewarding reproducibility and open science.

Beyond the replacement of the Impact Factor with another metric, there are additional ways tenure criteria could be altered to enhance incentives for reproducible research. For instance, policies could reward the open sharing of materials, including detailed research protocols, datasets, manuscripts, code, and reagents. While this practice is not currently widespread, selected institutions, including the National Institutes of Health, already explicitly reward the timely and open electronic sharing of data in their internal tenure criteria.[48]

Enacting change in tenure policies

While there are clearly alterations to tenure criteria that could reward and incentivize better practices in reproducible and open research, it is less clear how to enact a meaningful change. The unique nature of institution-specific tenure policies and practices creates and environment of decentralized decision-making related to tenure. Therefore, to better understand what an intervention into tenure might look like, we examined groups that have previously issued guideline documents related to tenure with the potential to influence tenure criteria across institutions. Please note that none of these guidelines has directly addressed open science or reproducibility.

American Association of University Professors (AAUP)

The American Association of University Professors (AAUP) is a non-profit membership association of academics who "shape American higher education by developing the standards and procedures that maintain quality in education and academic freedom in this country's colleges and universities."[49] The organization functions largely to protect the interests of faculty working in academia, and it has an affiliate, the AAUP-CBC, that acts as a labor union.

[46] Vincent Lariviere & Yves Gingras, *The impact factor's Matthew effect: a natural experiment in bibliometrics*, submitted to arXiv on Aug. 21 2009, *available at* https://arxiv.org/abs/0908.3177.

[47] V. Calcagno et. al., Flows of Research Manuscripts Among Scientific Journals Reveal Hidden Submission Patterns, 338 *Science* 6110, 1065 (2012), http://science.sciencemag.org/content/338/6110/1065.

[48] Criteria for Tenure at the NIH, National Institutes of Health, https://oir.nih.gov/sourcebook/tenure-nih-intramural-research-program/criteria-tenure-nih.

[49] About the AAUP, American Association of University Professors, https://www.aaup.org/about-aaup.

In 1958, the group's Committee A published its first guideline statement on tenure and faculty dismissal proceedings.[50] Since then, the Committee has issued approximately a dozen periodic updates to its guidance in a document known as the 'Recommended Institutional Regulations on Academic Freedom and Tenure' (RIR); the latest RIR was published in 2014.[51] The RIR guidelines currently focus on ensuring academic freedom, outlining conditions for termination of tenured faculty, and the prevention of discrimination. The statement includes little to no guidance on tenure promotion criteria. The basic principles outlined in the RIR statement appear to be widely adopted and reflected in tenure criteria at institutions across the United States.

Along with the American Council on Education, the AAUP also published "Good Practice in Tenure Evaluation" in 2000.[52] The four chapters of the document focus on clarity of tenure standards, consistency in evaluation, candor, and treatment of unsuccessful candidates. There is a much greater focus on advising the implementation and communication of tenure standards than the design or components of the standards.

We reached out to AAUP's Committee A to ask to discuss changing tenure criteria to facilitate increased research reproducibility; however, a representative for the organization, [redacted], said that the group generally refrains from commenting on the actual content of tenure policies, so long as the criteria is faculty-controlled and communicated clearly by the institution.

American Council on Education (ACE)

The American Council on Education (ACE) is the major coordinating body for institutions of higher education in the United States and is frequently involved in policy-making activities at the national level. They provide guidance on a variety of topics, and have occasionally released documents relating to tenure.

In addition to the Good Practice in Tenure Evaluation document noted above, the group published "Internationalizing the Tenure Code: Policies to Promote a Globally Focused Faculty" in 2015.[53] The focus of this document was largely evaluating the extent to which university tenure criteria incentivize international teaching, service, and research. This was presented alongside case studies which illustrate how universities can implement more international-friendly criteria. However, the content of this document appears less influential, and we found little evidence that indicated widespread adoption of its suggestions. [Redacted.]

[50] Statement on Procedural Standards in Faculty Dismissal Proceedings, American Association of University Professors, https://www.aaup.org/report/statement-procedural-standards-faculty -dismissal-proceedings.

[51] American Association of University Professors, *Recommended Institutional Regulations on Academic Freedom and Tenure*, Academic Due Process (2014), *available at* https://www.aaup.org/file/RIR%202014.pdf.

[52] American Association of University Professors et. al., *Good Practice in Tenure Evaluation* (2000), *available at* https://www.aaup.org/sites/default/files/files/Good%20Practice%20in%20Tenure%20 Evaluation.pdf.

[53] American Council on Education, *Internationalizing the Tenure Code* (2015), *available at* http://www.acenet.edu/news-room/Documents/Internationalizing-the-Tenure-Code-Policies-to-Promote-a-Globally-Focused-Faculty.pdf.

Recommendations

After reviewing the available evidence, it appears there is room for positive change in tenure policies to incentivize open science; however the major obstacle in changing tenure policies is their institution-specific nature. We were unable to identify any individuals or organizations actively working in this area to recommend for grant consideration and conversations with experts led us to believe that other policies that may be more effective in creating change.

[Redacted.]

Therefore, at this time, we do not recommend funding any specific interventions related to changing tenure criteria. However, we acknowledge that other interventions we have recommended, including the creation of a reproducibility metric, may indirectly impact tenure decision making if it becomes widely adopted and recognized.

IV. Intervention 4: Characterizing dependency relationships (“Shepherdizing”) in scientific research

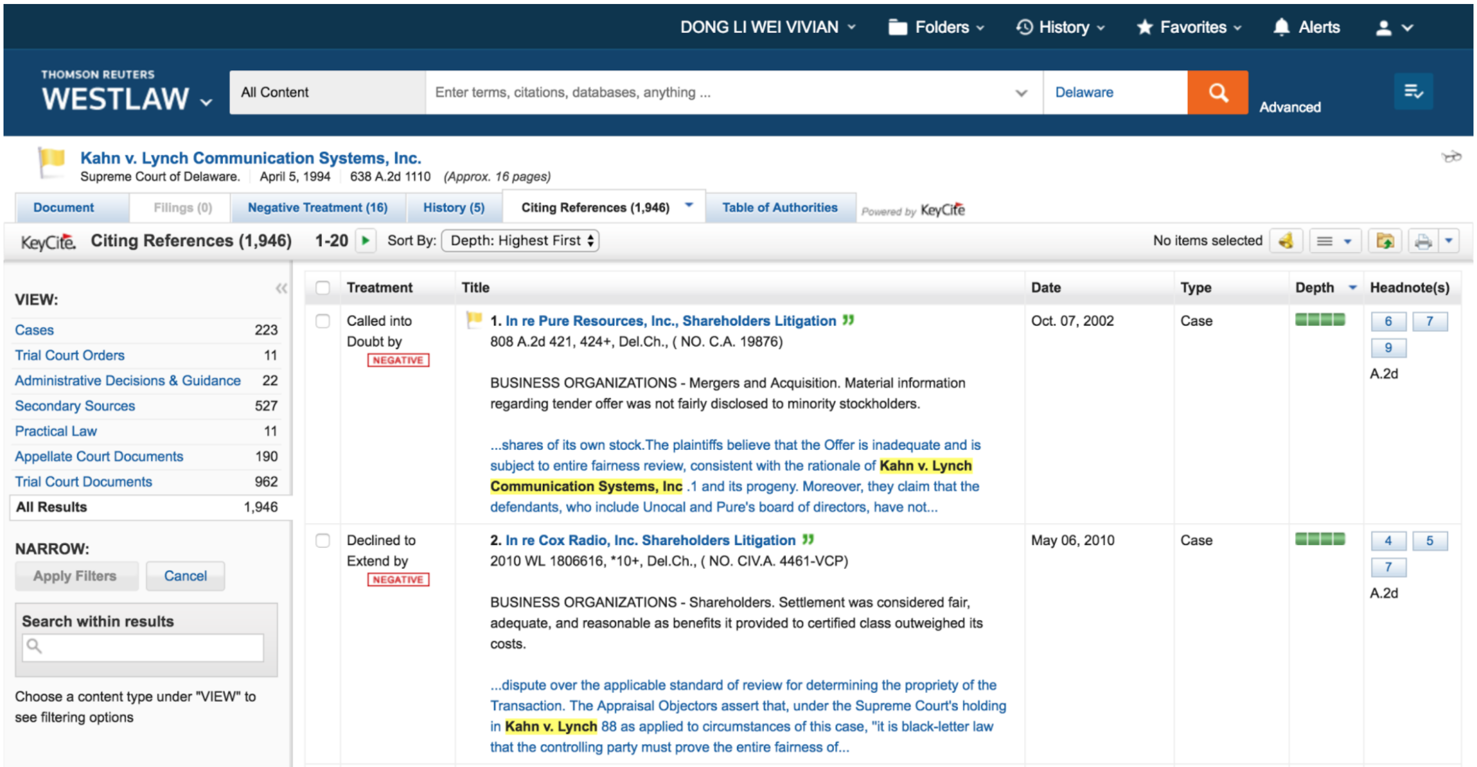
In an Edge.org post, Brian Christian argues for the necessity of “dependency graphs” in organizing scientific literature. He illustrates the problem with the current state of organization:

> It amazes me how poorly the academic and scientific literature is configured to handle even retraction, even at its most clear-cut—to say nothing of subtler species like revision. It is typical, for example, that even when the journal editors and the authors fully retract a paper, the paper continues to be available at the journal's website, amazingly, without any indication that a retraction exists elsewhere, let alone on the same site, penned by the same authors and vetted by the same editor. [...] The scientific literature, taken as content, is stronger than it's ever been—as, of course, it should be. As a form, the scientific literature has never been more inadequate or inept.[54]

More concretely: When a researcher goes on PubMed and looks up a paper, she can see a list of articles the paper cites, but she cannot figure out the nature of the relationship between the paper and the cited articles unless she actually reads the paper and/or articles. This makes for inefficient research. But more significantly, as Christian points out, this means that “academic literature makes no distinction between citations merely considered significant and ones additionally considered *true*.”

[54] Brian Christian, Scientific Knowledge Should Be Structured as Literature, Edge.org (2014), https://www.edge.org/response-detail/25514.

Westlaw and LexisNexis, the two major online legal research services, are more advanced than PubMed in this regard. The two companies hire legally trained researchers to read through cases and describe the logical relationship between a case and the case it cites (known as "Shepherdizing" among lawyers). So when a Westlaw user pulls up the tab that shows all the cases that cite to the case currently being read, he sees how the citing cases treated it:



Kahn v. Lynch Communication Systems, Inc.
Supreme Court of Delaware. April 5, 1994 638 A.2d 1110 *(Approx. 16 pages)*

KeyCite. Citing References (1,946) 1-20 Sort By: Depth: Highest First

No items selected

VIEW:

Cases	223
Trial Court Orders	11
Administrative Decisions & Guidance	22
Secondary Sources	527
Practical Law	11
Appellate Court Documents	190
Trial Court Documents	962
All Results	1,946

NARROW:

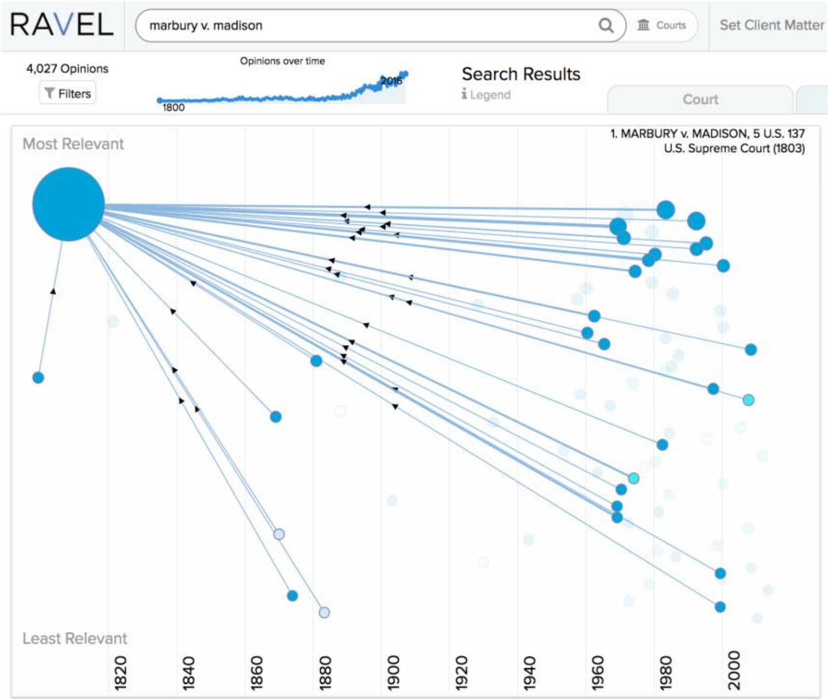
Apply Filters Cancel

Search within results

Choose a content type under "VIEW" to see filtering options

□	Treatment	Title	Date	Type	Depth	Headnote(s)
□	Called into Doubt by NEGATIVE	**1. In re Pure Resources, Inc., Shareholders Litigation** 808 A.2d 421, 424+, Del.Ch., (NO. C.A. 19876) BUSINESS ORGANIZATIONS - Mergers and Acquisition. Material information regarding tender offer was not fairly disclosed to minority stockholders. ...shares of its own stock.The plaintiffs believe that the Offer is inadequate and is subject to entire fairness review, consistent with the rationale of **Kahn v. Lynch Communication Systems, Inc** .1 and its progeny. Moreover, they claim that the defendants, who include Unocal and Pure's board of directors, have not...	Oct. 07, 2002	Case	🟩🟩🟩🟩	6 7 9 A.2d
□	Declined to Extend by NEGATIVE	**2. In re Cox Radio, Inc. Shareholders Litigation** 2010 WL 1806616, *10+, Del.Ch., (NO. CIV.A. 4461-VCP) BUSINESS ORGANIZATIONS - Shareholders. Settlement was considered fair, adequate, and reasonable as benefits it provided to certified class outweighed its costs. ...dispute over the applicable standard of review for determining the propriety of the Transaction. The Appraisal Objectors assert that, under the Supreme Court's holding in **Kahn v. Lynch** 88 as applied to circumstances of this case, "it is black-letter law that the controlling party must prove the entire fairness of...	May 06, 2010	Case	🟩🟩🟩🟩	4 5 7 A.2d

The descriptions of these relationships are Westlaw's proprietary information. Thomson Reuters, which owns Westlaw, affords to pay for the significant labor costs involved in this endeavor because it charges significant fees to its clientele. Ravel Law, a startup seeking to provide a similar service, does not currently offer descriptions of relationships between cases but does offer spatial representations of citing cases over time:



We talked to [redacted] at Harvard Law School for his thoughts on the feasibility of creating a “Westlaw for science.” [Redacted.] As a result, our research into this promising intervention is less developed than the others’.

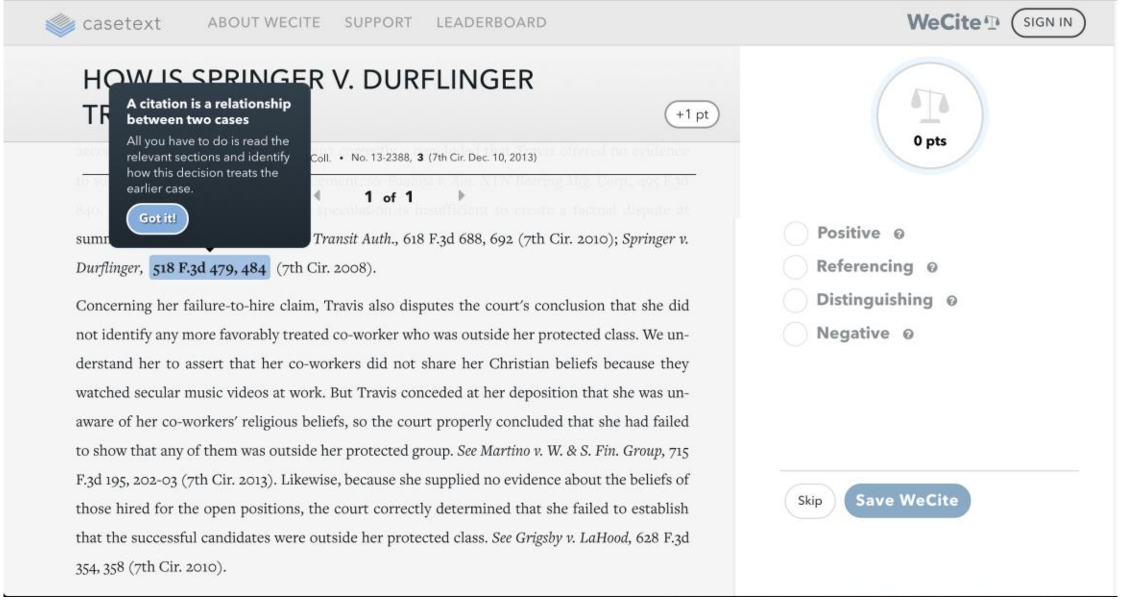
[Redacted] thought that the basic concept as we described to him was plausible, though he caveats that he does not have a scientific background. His main concern is that the scope and character of the relationships between articles in science are more sophisticated than between legal cases. This may undermine the usefulness of such a feature to science researchers—it will not make researchers much more efficient if researchers must read a long description to get an accurate sense of the relationship between two articles. More critically, to be feasible in the long-term as a non-profit, philanthropic endeavor—unlike Westlaw—this feature must avoid high labor costs. For this reason, we prefer that these relationship characterizations are eventually generated through AI or crowdsourcing, not paid human labor.

Current AI/Natural Language Processing technology is not sufficiently advanced to generate these characterizations. [Redacted] thinks this will be hard, even for law, where the relationships are simpler and the Bluebook citation form (for example, the use of signals *see*, *e.g.*, *cf.*, *but see*, etc. and explanatory parentheticals) makes figuring out the characterization of a particular citation much easier. Therefore, we believe that for science, crowdsourcing the characterization of citations would be a more feasible option than a machine learning approach.

As for what organization could actually be a “Westlaw for science,” [redacted] said that one major bottleneck for finding such an organization is that the organization must have access to the data from which to generate the characterizations. In other words, the organization must have rights to the underlying articles themselves, not just access to search engines or databases that host the articles. [Redacted.]

Given this bottleneck and the technological sophistication the intervention requires, we think at this time it makes more sense to first focus on finding and funding an organization to build the underlying technology instead of a full platform. [Redacted.]

Though we did not get the opportunity to speak to them, we found Casetext particularly impressive. Namely, Casetext has a feature, WeCite, that invites law students and lawyers to read through cases and do essentially what Westlaw pays researchers to do—characterize the relationships between cases. If a website user volunteers to “WeCite” for Casetext, they get taken to a page that looks like this, where they are invited to read and characterize the relationship (on next page):



HOW IS SPRINGER V. DURFLINGER TR

A citation is a relationship between two cases

All you have to do is read the relevant sections and identify how this decision treats the earlier case.

Coll. • No. 13-2388, 3 (7th Cir. Dec. 10, 2013)

sumr Transit Auth., 618 F.3d 688, 692 (7th Cir. 2010); *Springer v. Durflinger*, 518 F.3d 479, 484 (7th Cir. 2008).

Concerning her failure-to-hire claim, Travis also disputes the court’s conclusion that she did not identify any more favorably treated co-worker who was outside her protected class. We understand her to assert that her co-workers did not share her Christian beliefs because they watched secular music videos at work. But Travis conceded at her deposition that she was unaware of her co-workers’ religious beliefs, so the court properly concluded that she had failed to show that any of them was outside her protected group. *See Martino v. W. & S. Fin. Group*, 715 F.3d 195, 202-03 (7th Cir. 2013). Likewise, because she supplied no evidence about the beliefs of those hired for the open positions, the court correctly determined that she failed to establish that the successful candidates were outside her protected class. *See Grigsby v. LaHood*, 628 F.3d 354, 358 (7th Cir. 2010).

Law students rack up points for contributing WeCites, which then lead to prizes, using an innovative gamification system to incentivize the crowdsourced data-gathering.

Recommendation: Assuming enough people contribute WeCites and do so competently, Casetext’s WeCite feature could provide a model for science research. Pre-existing databases could crowdsource the characterizations of relationships between articles to researchers and/or students. **[Redacted]; we would recommend that LJAF continue this line of research by speaking to Casetext about this concept**.

Final Recommendations

We reviewed a range of ideas and interventions in preparation of this report. Many of these show significant promise. From our survey, we arrived at four recommendations for concrete actions LJAF can immediately pursue:

1. Fund and organize a conference to brainstorm and define a reproducibility metric, possibly a transparency index,
2. Approach [redacted], a stealth-mode startup seeking to quantify the reproducibility of observations in the biological sciences, with an offer of funding in exchange for more open availability of their product,
3. Fund the Global Biological Standard Institute to develop training modules for scientists and track adoption of better practices in the biomedical community that will improve reproducibility,
4. Speak with the team at Casetext, a legal innovation startup, on how their model of crowdsourcing the job of characterizing relationships between cases could be applied to science.

Appendix

[Redacted]