doi.org/10.53975/cwwp-mrgc

Seeds of Science

Why Proposal Review Should Be More Like Meteorology

Stuart Buck[1]

The process of evaluating research proposals for funding is often based on subjective assessments of the "goodness" or "badness" of a proposal. However, this method of evaluation is not precise and does not provide a common language for reviewers to communicate with each other. In this paper, we propose that science funding agencies ask reviewers to assign quantitative probabilities to the likelihood of a proposal reaching a particular milestone or achieving technical goals. This approach would encourage reviewers to be more precise in their evaluations and could improve both agency-wide and individual reviewer calibration over time. Additionally, this method would allow funding agencies to identify skilled reviewers and allow reviewers to improve their own performance through consistent feedback. While this method may not be suitable for all types of research, it has the potential to enhance proposal review in a variety of fields. [abstract generated by ChatGPT]

We're all familiar with meteorologists' forecasts: a 5% chance of rain, a 95% chance of rain, etc. It would be far less useful if you turned on the Weather Channel only to hear on every occasion that there were two possible forecasts: "Rain will happen today," or "rain will not happen today," with the forecast being wrong half the time. It would be only slightly more useful if the forecasters used a scale of 1 to 9: "Chance of rain today is a 7" versus "chance of rain is a 5." What does any of that actually mean? Should you carry an umbrella or not? Turn off the yard sprinklers? Plan for extra time on the drive to a meeting?

This is not unlike the situation we're in with proposal review at the moment.

Proposal reviewers are constantly expected, even if implicitly, to make predictions about the success of a particular research proposal. But to my knowledge they are *not* asked to provide actual probabilities, let alone probability distributions of likely impact. Instead, they're asked to make judgments about whether a proposal is "good" or "bad", or to rate proposals on a scale from 1 to 9,

[1] Executive Director, Good Science Project; Senior Advisor, Social Science Research Council; email: stuartbuck@gmail.com

or something else that doesn't actually require them to estimate and articulate probabilities of success.

As an experiment in enhancing proposal review, science funding agencies ought to try having peer reviewers or other evaluators *assign quantitative percentages* to a proposal's likelihood in reaching a particular milestone or technical goals. This idea might work better for some research areas (goal-directed projects, clinical trials, etc.) than for others (e.g., basic exploratory science), but focusing on probabilities would likely benefit all proposal review.

How might this look in practice? Reviewers would likely show some independent and diverse judgment (Lee et al., 2013). For example, Reviewer 1 gives a proposal a 50% chance of success in reaching milestone 1, while Reviewer 2 gives it 85%, and Reviewer 3 says it only has a 25% chance. These kinds of judgments can then be scored over time. Better yet, reviewers could provide confidence intervals, e.g., "I'm 95% confident that the probability of success is between 40% and 60%."

It might turn out that Reviewer 1 is pretty good (or is "well-calibrated" in forecasting parlance) in that the proposals that she predicts to have a 50/50 chance generally turn out to succeed (or fail) half the time. On the other hand, Reviewer 2 might turn out to be biased towards positive judgments, assigning higher probabilities to proposals that don't succeed very often, indicating he should perhaps be more conservative in his estimates. Likewise, Reviewer 3 might be far too negative, underestimating how successful proposals turn out to be. The point is not to expect perfect accuracy from all reviewers (although that would be nice). The point, instead, is that by requiring reviewers to make probabilistic forecasts, funding agencies can identify skilled reviewers, and reviewers can use their feedback to improve their own performance.

Here are some of the potential benefits:

- Everyone would have to be more precise in their evaluations. It would not be sufficient to say "this proposal is good, bad, or ok," but would require a reviewer to specify the actual chances that the approach would work or not. This would help reviewers to speak a common language with each other during proposal evaluations, as well as allow an agency to see the spread of forecasts among reviewers for any given proposal.
- Forecasting could improve both agency-wide and individual reviewer calibration over time through consistent feedback and review. Feedback loops are essential, after all, for improving performance (e.g., Goetz, 2011). If you're constantly making implicitly predictive judgments, but never have

to provide explicit probabilities and thus can't receive feedback as to whether your forecasts were accurate, you will miss out on many opportunities to improve.

- An agency would be able to visualize its overall portfolio with the proportion of its funding on research or projects that are truly risky (perhaps with under 20% chance of success), somewhat risky (perhaps closer to 50%), and – where risk may be less appropriate -- perhaps verge closer to slam dunks. In an agency that promotes a "high-risk, high-payoff" mission, funding proposals that on average have an 80% chance of success might be overly conservative. Other funding agencies might want a different risk profile, but probabilities would help them both in establishing what that means and whether they're aligned with that profile.
- Forecasting could also help reduce (or at least reveal) hindsight bias (Roese & Vohs, 2012). When a project fails, people will commonly opine that they always suspected that it was never going to work. This can lead to an impression that the investment was somehow foolish, wasteful, or unnecessary. But if reviewers had earlier given that project a 90% chance of success, and it failed, reviewers and organizations examine their perhaps unwarranted optimism. Forecasting can turn failed projects into learning opportunities. At the other end of the spectrum, success will often invite the response, "well, we always knew that would work," again suggesting that the research was unnecessary or somehow self-evident. Probabilistic forecasts would help temper that reaction, however, had the project been assigned an average of 30% chance of success during proposal evaluation.
- We could also test the effect of peer review discussions, which can potentially create peer *pressure* to conform to the judgments of other reviewers (Lane et al., 2021; Pier et al., 2019). For example, we could get reviewers to make a probabilistic forecast while blinded to what any other reviewer thinks; then we could gather the reviewers for a discussion (such as an NIH study section); and finally we could ask them to make a second probabilistic forecast. This would allow for tests of which forecasts are more accurate—independent or peer-influenced? Organizations could then adopt the more accurate procedures.

Given these potential benefits, the next obvious question is why the approach hasn't been widely adopted or at least attempted. Searching the published literature and consultations with members of the forecasting community revealed nothing. This absence of evidence is puzzling, but – like the proverbial economist who turns up his nose at a $20 bill on the sidewalk because it must be fake (otherwise someone else would have picked it up) – we shouldn't assume that it's

also evidence of absence. If such a presumably easily implemented and seemingly useful idea hasn't been tried, maybe there's a good reason.

One reason might be that some reviewers feel that they can't be more precise than to put proposals into, say, three buckets of predicted performance: "probably not; 50/50; and most likely." To make them assign a probability of 37% or 68% might feel like an artificial level of precision. After all, why 37% and not 38% or 36%? Who could possibly explain that kind of fine distinction? This is a fair point, but if reviewers want to round up (or down) their probability assessments to the nearest 10, or 25, or any other percentage, the beauty of forecasting is that they are free to do so. They need only be somewhat rational (give higher probabilities to proposals they think are more likely to succeed) and be willing to assign probabilities that can be scored.

This seems worthwhile: if a reviewer claims to be "skeptical" about a proposal, we should try to elicit a more exact probability. Maybe the reviewer has very stringent internal standards, and anything short of a 90% chance of success makes him or her skeptical. Alternatively, maybe the reviewer's skepticism maps to a 20% chance of success.

Requiring the reviewer to articulate a probability – *any probability* – makes it easier for others to understand what the reviewer actually means (and potentially to help the reviewer consider what he or she means as well)
Another reason for resistance may be that reviewers are uncomfortable with their performance being judged, which may reflect the organizational or scientific culture in which they operate. For example, the CIA was infamously uncomfortable with Sherman Kent's efforts to make analysts provide probabilities with their analytic judgments. Instead of judging the likelihood of an event being "possible, with medium confidence," Kent wanted analysts to put a probability on their judgments. When someone accused him of subsequently trying to turn the CIA into a betting shop, Kent's response was, like himself, rather legendary: "I'd rather be a bookie than a goddamned poet" (Ciocca et al., 2021).

While that sentiment may be a bit harsh, it is not wrong. If people – reviewers, program officers, agency heads – are determining whether and how to spend millions or billions of federal dollars, they should care about improving their own calibration and judgment. Indeed, I would argue that NIH and NSF should systematically capture the judgment and calibration of *all* reviewers and program officers. I could even imagine an ongoing set of awards: "Top-scoring NIH peer reviewer," "top-scoring SRO," and the like. This would further incentivize reviewers to make more accurate predictions.

If the National Weather Service's Global Forecast System started proving to be inaccurate (e.g., rain occurred only 70% of the time when the model predicted 80%), meteorologists would be able to jump into action to recalibrate the models, check the data inputs, and draw from a long historical database of past forecasts and outcomes. But they can only do this because they have a system and a culture of keeping score. Right now, scientific funders can't do anything comparable. As Fang and Casadevall write:

> "At its extremes, the error and variability in the review process become almost laughable. One of our colleagues recently witnessed an application to receive a perfect score of 1.0 when part of a program project application, but the identical application was unscored as a stand-alone R01. Almost no scientific investigation has been performed to examine the predictive accuracy of study section peer review…Putting a stronger scientific foundation and accountability into peer review would enhance confidence in the system and facilitate evidence-driven improvements"

As the best way to get ahead is to get started, we could all benefit from being a little more like meteorologists in this respect.

Acknowledgements

Thanks to Cory Clark and Phil Tetlock for helpful comments.

References

Ciocca, J., Horowitz, M., Kahn, L., & Ruhl, C. (2021, May 9). *How the U.S. government can learn to see the future*. Lawfare. https://www.lawfareblog.com/how-us-government-can-learn-see-future

Goetz, T. (2011, June 19). *Harnessing the power of feedback loops*. WIRED. https://www.wired.com/2011/06/ff-feedbackloop/

Pier, E. L., Raclaw, J., Carnes, M., Ford, C. E., & Kaatz, A. (2019). Laughter and the chair: Social pressures influencing scoring during grant peer review meetings. *Journal of General Internal Medicine*, *34*(4), 513-514.

Lane, J. N., Teplitskiy, M., Gray, G., Ranu, H., Menietti, M., Guinan, E. C., & Lakhani, K. R. (2022). Conservatism gets funded? a field experiment on the role of negative information in novel project evaluation. *Management Science*, *68*(6), 4478-4495.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2-17.

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science*, *7*(5), 411-426.

Gardener Comments

Andrew Neff (Assistant Professor of Psychology & Neuroscience):
Fascinating concept, persuasively and clearly written, I support it. I think the most interesting part is the concept of providing feedback for reviewers. As is mentioned in the article, this could help reviewers update their own biases, or help agencies like NIH/NSF get rid of bad reviewers, promote good ones, and be more inclined to bring new people into the fold. New people please.

Reading this makes me want a little more context for how forecasting could be incorporated into a holistic review process. You're not suggesting this, but a yes/no funding decision can't rely solely on the likelihood of success on its own, since nobody wants to fund only low-risk projects. Agencies need to add a second criteria for evaluating projects, something like 1) likelihood of success X 2) the value of success. Do we also need to quantify the value of success, or is that a different question?

I'm also thinking about how reviewers should judge the likelihood of success. At least at the NIH, it's based on 1) the investigators history, 2) the technical-soundness of the method, and 3) the supportiveness of the environment. Should reviewers be constrained in their methods for assigning % likelihoods of success (as they currently are to an extent), or should forecasters be given more leeway?

My forecasted increase in the probability that the NIH/NSF change their funding policy in the next 2 years (if this article is published vs not published): 0.01% (sorry)

Antman (PhD in behavioural biology):
This is a nice, constructive suggestion. I completely agree that putting hard numbers of judgements, on a clear scale with meaningful units, is constructive and helpful, and will make reviewers think more carefully about their reviews.

However, the proposal has obvious weaknesses which are not directly addressed.

Firstly, and perhaps most banal, the title contains the term "peer review," and most academics might assume this refers to peer review of research articles. However, the work is more limited in scope, and only really directly relevant to proposal reviews.

Secondly, and more importantly, this proposal would be most powerful in systems where repeat reviewers are tracked and scored, to identify super-predictors. However, no data is provided on how often, on average, an individual reviewer is asked to review different proposals for the same agency. The measure of prediction accuracy will only start, itself, to be useful after at last 5 or 6 reviews, and then still with large confidence intervals. Is it really realistic to expect most reviewers to have already performed so many reviews for any specific grant agency?

Thirdly, and related to the above, the literature on prediction science shows that it is very hard to identify super predictors, and very easy to confuse them with people who just got lucky - especially at the low sample sizes we would be dealing with here.

Finally, it should at least be acknowledged that 'chance of success' is not the only criterion which proposals should be judged on. An equally important measure would be "importance of the research", and these should interact to give a final score.

In essence, then, while I agree that adding a scale with meaningful units to reviews is helpful, I feel like many of the additional benefits of the system proposed here are not realistic, and thus oversell what is, in essence, a valuable exercise in improving prediction readability and reviewer reflection.

DK (PhD in psychology):
I would support the publication of this article, but only if the serious concerns that I have are addressed. My main serious concern is - the author claims that the current peer review procedures are not optimal because manuscripts are rated using arbitrary categories such as good-bad or certain scales (e.g., 1-9). However, I do not see an argument that would convincingly show me that using percentage/chance would not result in reviewers consistently choosing few arbitrary categories - e.g., 25%, ,50%, 75%. I would like to see a convincing argument regarding whether and why this would not be the case (there is a large

literature on how people perceive and estimate percentages which should be consulted), and I would also like the author to describe the measures which would ensure this does not happen.

Étienne FD:
This is a good article. It provides an interesting idea to help improve peer review. I will not comment on its "chance of success," but there's probably value in making success probability estimates in science more common than they currently are.

I do want to offer some feedback on the meteorology analogy, as I'm somewhat familiar with that field. While weather forecast computer models do produce precise numerical probability estimates for precipitation, it's common for forecasters to round these to a few "bins" for public communication. A weather forecaster I know tells me, for instance, that they purposefully never state a 50% probability, and in practice only round to a select few numbers, something like 0, 30, 40, 60, 80, and 100%. This is not very different from the 1 to 9 scale that the author offers as an example of what meteorologists do not do.

This speaks to concerns about how to communicate probability effectively, and whether it's actually useful to provide probability estimates when you don't have a more precise idea than "somewhat likely" or some such. It seems possible that moving to probability estimates would not in fact improve peer review, while giving it an undeserved veneer of credibility. It would be nice to discuss these concerns somewhat more than they already are, but I'm happy to recommend this article for publication regardless.

Josh Randall:
This article provides an interesting solution to one of the issues of the model of publicly funded science. Specifically, many of their points are describing ways of improving the study proposal mechanism associated with funding research, likely in human focused fields. Their suggestions and description of possible responses from reviewers seem reasonable given the increasing workload associated with writing and proving the possible success of new studies. However, unlike in meteorology in which success can be determined by whether rain occurred or not; many disciplines of science that do not strictly follow falsification or hypothesis driven research might not have an obvious outcome of success and failure. In research programs focused on natural history, a proposal might be describing a plan to describe the physiology of an organism that hasn't received much attention previously. Would a failure be impossible or occur if their physiology is not interesting? Many of these fields also construct experiments

with a goal of a positive outcome in all situations. In a publishing environment that seems to disregard 'negative' results, this is necessary. Should failure be a logistic failure of not publishing in these situations? Requiring that reviewers are considering risk-reward of studies in a quantitative manner seems useful, but whether this institutionalizes pre-existing biases against minority and female researchers should be considered as well.

Patrick Wilson:
It's clearly written but the idea is flawed. Here are 3 objections top of mind:

1) I am uncomfortable with the current trend, possibly inspired by EA and the super-forecasting movement, to get everyone to give percentage probabilities for every prediction. I have recently been taking part in the Tetlock hybrid forecasting and persuasion tournament, and it made me realise how hard it is to assign percentage probabilities for X-risk and other future events.

2) Percentage probabilities are not inherently superior to 9-point scales or other fuzzier and multidimensional measure, and also they're just not how human minds work. The probabilities will tend to cluster on key deciles / extremes and eg 25%, 33%, 75% because it's easier and that's how we are used to thinking.

3) Proposal evaluation / peer review is not only judging how likely a research project is to achieve its stated goals and limited objectives, but also how worthwhile that project actually is. For example I could submit a proposal for something with high confidence the results will achieve what I set out to, and peer-reviewers might agree with me, but it could be something so trivial as to not be worth studying. There's enough science like this around!

Peer review might be broken in many ways, but this isn't one of them. I expect if implemented, this proposal will drive even more people away from peer-review, through pedantic and over-prescriptive insistence on percentage predictions - a fool's errand.

Thomas Gladwin:
The paper addresses a very important topic, is well-written, and presents an interesting focus on requiring probabilities as a building block to further improvements and innovations.

EMSKE:
1. Does the article contain novel ideas that have the potential to advance science?

An incremental improvement to proposal peer review like the author describes would increase the level of accountability for research projects with uncertain outcome, and a step toward making the review process more meritocratic. I find the piece also inspiring thinking on whether project peer review could be crowdsourced with historical success weights (i.e. as a quantitative reputation-tracking system) of reviewers, rather than reviewers having to be sourced strictly from within the four walls of a funding agency.

2. Does the Seed include adequate justification for its ideas and how they could advance science?

Broadly yes; Obviously debatable, but I'm not sure that the 'proverbial economist who turns up his nose at a $20 bill on the sidewalk' is the best metaphor for why this improvement to proposal peer review hasn't been done already. The role that organizational politics play shouldn't be understated IMO. Anyone would want to sit in a position of power to 'pick winners'. Favor-trading, empire-building, dare I simply say, 'nepotism' - these behaviors aren't unknown to the world of scientific peer review.

3. Does the Seed contain high-quality writing?

For ease of reading, and ease of the reader following the author's train of thought, I wouldn't mind section headers to visually break-up the text. Can a couple of relevant metaphorical graphics be sourced from Unsplash?

William Collen:
A splendid idea. I try to do this anyway when making truth claims during conversation. The point about being able to analyze and modify our assumptions and calibrate our predictions was well put.

Dan James:
This article is written in the style of an opinion piece or blog post and whilst lacking the rigour of a more traditionally formatted academic paper, is no less engaging and thought-provoking.

Metaresearch – the scientific study of how science is conducted - is an area where new ideas to solve perceived problems are urgently required. This article focuses on proposal review, a vital first step to secure funding for many research enquiries.

Unfortunately, as the article explains, proposal review has a high error rate in predicting research success, one far less reliable than the comparator the paper chooses - a meteorological forecast. This article suggests one possible solution - reviewers should assign probabilities to a proposal in much the same way as weather forecasting.

This idea has a distinct 'Bayesian' feel, and whilst a Bayesian approach to assigning probabilities is not mentioned or developed by the author/s, it is a measure of how well the paper succeeds in presenting a generalised discussion that invites further reflection from the reader.

My specific Bayesian reflection on reading the paper (that could add to the argument in favour of assigning probabilities) is to break proposals down into incremental stages and stage payments (already common practice), with each successive stage updating the priors for any project. Assigning probability scores on a stage-by-stage basis is arguably more realistic/practical than the difficulty for a reviewer of estimating a probability score for the whole project. After the successful completion of the first stage of a proposal, probabilities for following stages could be updated by reviewers with far more confidence.

This article is a helpful contribution to an area that needs original thinking to address problems that have proved intractable to the traditional ways in which science has been conducted, consequently I would definitely recommend this paper for publication.

Dr. Payal B. Joshi:
The premise of peer review as a forecast seems amusing. It is believed that most of them "free" their time to review grants & manuscripts, I think this may work well. Authors could have structured this article in a slightly better manner as the methodology of applying it is not quite clear. How do you propose forecast-based review? Provide a clear example in a lucid way to put across the idea. Overall, the idea may not be pathbreaking, yet has no potential flaw & certainly deserves to see the light of the day.

Jack Arcalon:
A good idea if it causes researchers to think of unknown probabilities extending beyond their specialties in new ways.

Fred Nix:
I'm not convinced a rating system like this would actually work like the author imagines. Who would risk their cushy job or current status with an objective

rating? Would it just skew predictions so everybody shoots for 50/50? How long would it take for a reviewer to build up enough reviews and outcomes to get a valid statistic? Would the rating just show as pending until there was a valid statistic, or just publish the preliminary number? But the current system is broken, and innovation and progress are suffering. It's an interesting idea and worth experimenting with to see how it works.