



AMERICAN
PSYCHOLOGICAL
ASSOCIATION

ESSENTIAL SCIENCE CONVERSATION:

SCORE AND THE STATE OF SCIENTIFIC CREDIBILITY: IMPLICATIONS FOR RESEARCHERS (JUNE 23, 2026)

SESSION DESCRIPTION

The SCORE program offers large-scale examination of reproducibility, replicability, and robustness across social and behavioral sciences. By coordinating replications, reproductions, and reanalyses, SCORE moves the field toward systematic assessment of repeatability and credibility. Join us for an Essential Science Conversation that explores implications for researchers and next steps for strengthening scientific credibility. Unable to attend live? Register anyway and about two to three weeks after the live session, we will send you the recording to view at your convenience.

Any opinions, findings, and conclusions or recommendations expressed by the presenters in this webinar do not necessarily reflect the views, policies or positions of the American Psychological Association.

UNOFFICIAL TRANSCRIPT

Shandol Hoover: Hello, everyone, and welcome. Thank you so much for joining today's Essential Science Conversation, *SCORE and the State of Scientific Credibility, Implications for Researchers*. I'm Shandol Hoover, Senior Director for APA Science Special Projects and Implementation. Before we get started, I'd like to share a few quick announcements. Today's session is being recorded, and within two to three weeks, everyone who registered will receive a link to the recording, along with a transcript and additional resources.

Please note that any opinions, findings, and conclusions or recommendations expressed by the presenters for today's webinar do not necessarily reflect the views, policies, or positions of the American Psychological Association. We encourage you to visit apa.org/science to learn more about how to engage with APA science. You can subscribe to our free science spotlight newsletter to stay current on news and opportunities for psychological scientists, and to our Editor's Choice newsletter for curated articles from across APA journals.

We also invite you to explore the Response Center for updates on policy developments and access to helpful resources and support. We have an exciting lineup of webinars coming this fall, and you can explore more than 70 on-demand sessions available for free on our website. We're continuing to add webinars for this fall and encourage you to check out these sites to register and access all of the sessions for free.

Today's program is part of APA's Essential Science Conversation Series, which creates space for open dialogue between experts in the broader community on emerging topics in psychological science. Thank you to everyone who submitted questions during registration. You can also submit questions throughout today's session using the Q&A box. If you experience any technical issues, please let us know there as well. With that, I'm pleased to turn things over to Dr. Adrienne Stith, APA Executive Lead Psychologist for Science, who will introduce and moderate today's session.

Adrienne Stith, PhD: Thank you so much, Shandol, and good afternoon, everyone. Welcome to today's APA Essential Science Conversation, *SCORE and the State of Scientific Credibility Implications for Researchers*. SCORE stands for Systematizing Confidence in Open Research and Evidence. We're really excited today to bring together experts in open science and methodology to talk about what we're learning from the SCORE program and about research credibility.

We'll spend the next hour exploring what these insights mean for researchers and how we can all contribute to more rigorous and collaborative science. With that, I'm going to start with introductions of today's panelists, and I'll go in alphabetical order and start with Flavio Azevedo, who is Assistant Professor of Interdisciplinary Social Science at Utrecht University and Director of the Framework for Open and Reproducible Research Training, or FORT, where he focuses on advancing open science practices and improving research transparency and education across disciplines.

Next, we have Jessica Flake, who is an Associate Professor of Quantitative Psychology at the University of British Columbia, and she's known for her work to integrate psychometrics, metascience, and open science practices to improve methodological rigor. Now, we have Brian Nosek, who is Co-founder and Executive Director of the Center for Open Science and a professor at the University of Virginia, and is widely recognized for leading global efforts to increase openness, integrity, and reproducibility in scientific research.

Our next panelist is Fred Oswald, who is Professor of Industrial Organizational Psychology in the School of Education at the University of California, Irvine. His research focuses on assessing workforce skills and improving the quality of psychological measures in educational and organizational settings. Welcome to all of you. I am so excited for the conversation today. I thought that we would just start with a little bit of background about what the SCORE program is, a bit about the background, and what it accomplished. I'm going to ask Brian if you could just say a few words about that.

Brian Nosek, PhD: Yes, happy to provide a brief summary, and thanks for organizing this event and everyone for attending. SCORE was a DARPA-funded program that had as its main goal to see if machines could tell us what findings are likely to replicate or not. This was conceived and begun in 2019 and was a distributed program with several different teams contributing different parts. One team was responsible for extracting claims from the social and behavioral sciences, a sample of about 60 journals, 3,900 claims from individual papers from those journals.

That would be the basis of trying to predict replicability. Then there were other teams that had humans who would make predictions through different types of methods for what they thought was likely to replicate. Then there were other teams that generated machine approaches to try to predict replicability. Notably, since this started in 2019, this is not any involvement of large language models or what the current AI craze is. It was what was the modern elements all those years ago, 2019.

All of those teams were trying to assess claims across a large sample of social behavioral science literature. Then, behind the scenes, there were many, many, many people that were trying to provide evidence about the replicability or the robustness, or the reproducibility of a sample of those claims. What ultimately came out of this project was a lot of data artifacts about the social behavioral sciences. A large sample of replication attempts of a sample of the literature, a large sample of reproducibility attempts, just trying to take the same data and apply the same analysis as the original study to see if you get the same result as reported in the paper.

A large sample of robustness tests where you would take the same question and apply alternative ways of analyzing the data in order to test that question and see how reliable the results were. Collectively, that body of research provided a groundswell of engagement and interest in how it is we can establish credible trustworthy research findings, and the role of these different types of ways of thinking about repeatability might contribute to that. In total, it involved about 865 researchers. I am just one of those. Thanks. That's it.

Adrienne: Wonderful. Thank you so much for that. Maybe I'll start with a question about what do you think the SCORE program revealed about the current state of reproducibility, replicability, and robustness in the social and behavioral sciences? Maybe I'll start with you and ask you that question and then ask others to chime in as well.

Brian: Sure. I'll give a short answer because there's lots of different perspectives, I think that could be revealed with this. To me, the basic lesson is that establishing repeatability is hard. That may be an obvious thing, but this is concrete evidence with a lot of experience, with a lot of different approaches across the social behavioral literature, that this is challenging. The basic reproducibility, I'm just trying to see if the findings that you report in the paper are the things that I can observe when I reanalyze the same data. The fact that we did not succeed a lot of the time in doing that is surprising.

Wait a second, are people messing up all of their findings in their papers? Probably not as often as it was observed in the SCORE paper, but the fact that it was challenging to establish that suggests that there's opportunities to improve how we share, how we document, and how we facilitate the ability to verify claims. Robustness revealed a lot of uncertainty that isn't often revealed in individual papers. When we read a paper, we say, "Oh, this is the findings from that data." Really, it's the findings contingent on the decisions that the authors made and how to analyze the data to come to those conclusions.

We don't observe the variability, the heterogeneity in those findings based on alternative analyses. It showed that there's a lot of variability. Then, on replicability, this is one that's just continuing on something that field psychologists have been working on for the last decade, which is, "Oh, it is challenging to establish the replicability of findings." It isn't that published is a synonym for true. Published is the start of a conversation. "Oh, here's some initial evidence. This is intriguing." Under what conditions is it repeatable? That's part of the work of doing science.

Adrienne: Thank you. I'd love to hear what others think.

Fred Oswald, PhD: Okay, I'll chime in here. I think what replication and reproducibility has brought into important conversations in the social sciences is the nature of process and the pieces that go into research that we're not only needing to be transparent about for the sake of reproducibility or replicability, if you want to build on it. I guess that's the latter point I want to make is that open science allows us to share in ways that advance our science.

That we're working together as a community on our research. We're not just siloing our own work within our labs. Of course, we have this thing called the internet where we can share all kinds of stuff, but really the open science framework and other frameworks, other information that'll be shared here, I think, are ways to let a thousand, let a million flowers bloom in terms of gathering information, insights, concrete materials for running studies in your own work to further our science. I think that materials point of view around the efforts toward reproducibility and replication is really important and, again, to do it as a community as Brian and his teams have in their multi-lab studies.

Adrienne: Thank you, Fred. Any other thoughts about that?

Jessica Flake, PhD: I'm just going to unmute. I think we said we might stay unmuted so that it's easier to talk amongst each other. I guess I would add to that I think something, the SCORE projects in particular, because the scale of the robustness analysis is really big. I think Brian Krugman, that's a first of its kind for that kind of analysis. It really builds off of previous reproducibility and replicability attempts, in addition to the reproducibility and replicability projects also being large.

I think it's really exciting to start to go from a really high-level view and think really differently about how we accumulate evidence. It suggests, I think, an epistemological or paradigm shift that I feel like we keep teetering on the edge of. I do a lot of meta-scientific thinking and meta-methodological thinking. How are we going to, instead of publish single studies or single claims, start to synthesize evidence across claims, across the literature, and add in AI, add in language models, think about how evidence synthesis is going to change.

Though it might seem a bit discouraging, some of the results of these projects, I think it's actually really exciting. It suggests to me the beginning of an era where we might be doing more of this meta-synthesis or meta-evaluation that's going to, I think, fundamentally change how we publish research, how we think about evidence, how we think about claims. Just this idea that we all know what sampling variability is or statistical uncertainty is.

The SCORE Robustness Project, I think, is big enough, but I hope it gets more people thinking about, well, there's this other kind of uncertainty in the results. We have the central limit theorem. We have a standard error to think about our sampling error. How can we think about this uncertainty, this distributional thinking about results that just comes from different analysis methods, say, for example? I think there's a lot of things that are exciting because they might be the beginning of a paradigm shift in how we think about research. That was one of my reactions looking through the projects and reading over the papers recently.

Flavio Azevedo, PhD: I'll just chime in on Jess's thoughts. I really think that what this project brings is a paradigm shift for cumulative research. It's in good time because meta-science is evolving at a very fast pace that allows us to look more at patterns or accumulation of knowledge rather than how science was construed 10 years ago. We only have, what, 10 years from the Open Science Collaboration? These conversations are rather new when you think about history of science, especially in social sciences.

I think what SCORE brings up is this potentiality for us to maybe come together as a community that is interested on, say, a scientifically relevant effect or a societally relevant effect and work together in several different ways to find what are the contingencies of this finding. Is it about different cultures? Is it about different timelines? Is there a political component to it or institutional component? I think, similar to Jess's point, it makes science way more interesting. I feel that it provides a appeasement to a little bit of an infight within science because we often say, hey, these people or that group is not paying attention on something that I really care about, whether that's measurement, whether that is epistemology, or whether that is something else.

Each one of us has this tendency about, this matters to me, why are you not taking this extremely serious? I think that the outcome for SCORE is that, actually, we need everybody. We need big team science to come together and work towards decreasing this uncertainty. Investigating more systematically, what are the boundary conditions of each finding for the ultimate goal of science, which is accumulation of knowledge?

Brian: Yes, I think that's right. If we're to take a meta-lesson from these types of projects, it's that we have overconfidence in the current literature. It's not that the findings shouldn't be taken seriously from the literature. That's part of what science is. Oh, that's interesting. I don't know. What about those? What about that? It's that we're overconfident in the simplicity of the extent to which we can draw conclusions on a paper-by-paper basis, rather than recognizing that real progress is made through that accumulative evidence that Jess and Flavio were talking about, that really it's condition-seeking and looking for how it is that we observe findings in order to gain explanatory power and ultimately be able to predict effectively how it is that these are going to play out.

That's hard to remember when we're rewarded and thinking about our work on a paper-by-paper basis. Well, this is what we found. This is what we now know when we know that that's not actually true.

Fred: I just wanted to build on that. I guess I'm doing a double comment here with the second round of comments. To me, an underappreciated paper by Brian Nosek and Errington. It's in PLOS Biology. It addresses this notion of generalizability across studies. If you think of-- and I'm not going to do justice to your thinking, Brian, but the way I might put it is, if you think about the generalizability aspects of studies taken individually, that's a different perspective than the meta perspective of how is every study speaking to generalizability. Whether they say so themselves, or whether you glean additional insights through the accumulation of studies, or whether you also have your own lens because you want to focus on your aspects of generalizability that you care about.

I just think that's-- The idea here is not that we systematize all our studies. The wrong message to get, I think, from Open Science is that here are all the rules for the game, necessarily. There are indeed some rules that are important, like the central limit theorem, et cetera. Oh, thanks for the link in the chat there. Everyone has their own approach as well. You almost can't avoid forking paths. In other words, you can't explore all options as a researcher. That would be impossible. Which options are you going to pick and explain and justify, and how does that fit into the larger puzzle in terms of the process of research and the outcomes in terms of our effects? Sounds like a sermon, but--

Flavio: Famously, Einstein said, God does not play with dice. I think physicists and chemists, they started taking this garden of forking paths way more seriously, given the uncertainty of their results, how they would do the same kind of result, and it would yield different results. I think social sciences need a little bit that flexibility in that there's a lot more investigation to be done after a paper is published, after a claim is made. We just need to a little bit learn from anyone, but embed that uncertainty into the way we do our science, into our processes, into how transparent we are.

Adrienne: Thank you. I'm going to pose a question to Jessica. How have open science practices like preregistration and data sharing, and registered reports inspire new ways of developing methodology?

Jessica: Yes, I added that question because I've been thinking about it a lot. I was interested, I saw this connection too in the discussion of the SCORE paper that was about robustness, that if we're going to think about this analytical uncertainty or analytical flexibility, we need to further develop specification curve analysis or multiverse analysis. These are approaches for trying to estimate or observe the heterogeneity in results across different data processing or analysis steps. Then it's really got me thinking, working with registered reports and working with reproducibility, that when

we develop the quantitative methodologies or the analysis tools themselves, we're not thinking about these problems.

Fred is the editor of Psychological Methods. It's a journal that publishes a lot of methodological research. I've been an associate editor there and at another methods journal. I don't get papers, methodological papers where people are developing methodologies where they say, we know that these are a problem. We know that reproduction is a problem. Actually, there's more work on reproducibility, I would say. Computational reproducibility is starting to have actual methodological development behind it.

I haven't seen something of like, well, we're developing a new method. We want to consider how it relates to analytical flexibility. We want to consider what kind of different results one would get if they use these 38 other methods that are in the literature. I don't think there's an answer to that question in the sense that a lot of people are thinking about it, but I'm really struck when engaging with the SCORE project and similar things about how we-- I think we need to think about how we develop methodology at the start with these issues.

When we're thinking about developing a new way of analyzing the data, how does that relate to the existing ways of analyzing the data? How does that increase or shrink analytical uncertainty? If we're developing methodologies, why isn't it that there is a research synthesis that describes all available methodologies? If we're going to do a multiverse analysis, why isn't there a methodological paper that determines the size of the multiverse or identifies all of the cells in a specification curve analysis?

I think there's a couple steps back with the development of methodologies where methodologists can engage in research synthesis, systematic review. These kinds of approaches don't really exist in methodological research. I've learned, trying to learn more about this over the last few years. I think we can think a lot about not telling researchers that they didn't pick the right method or saying that the method is wrong, but just that how many methods are there for doing something? What's the natural amount of analytical variability that we will observe because there are so many different methods?

I thought it was really interesting that the robustness paper considered design features. If an experimental design versus an observational design would show more heterogeneity in results or less robustness to analytical flexibility, well, are methodologists, are they developing simpler models? Are they developing more complex ones? I think model complexity is probably a more accurate way of thinking about that.

I don't think methodologists are doing research like this. They're not engaging with some of these challenges. I hope that given the scope of these projects, it can spark those kinds of conversations, not just in applied journals because the researchers aren't doing the right thing, but methods journals to say, well, how are we actually developing the ways of knowing via the development of quantitative methodology or research synthesis? I'm curious, Fred.

Adrienne: I see nodding. I was like, "I think, yes."

Jessica: Yes, I'm curious, Fred, if you were thinking at all about what we do over at Psych Methods in relation to these projects.

Fred: Yes, thanks for turning this into a conversation. Yes, of course, I think about these issues. I think the papers at Psych Methods tend to plug into that larger issue that you're speaking of, which

is, what are the analytic, not just analytic choices that Psych Methods might focus on, but also the research choices and the multiverse that is being explored? When it comes to published research, we don't have a section that talks about all the options that people considered.

We might infer the options from the collection of studies that we see. Like you alluded to, it might lean toward some biases that are picked up on the SCORE project, but it's also just pointing out the universe of what people did when they published. That's useful. What about what's under the hood? These are the things we talk about, maybe more in graduate school, where natural mistakes come up in the academic environment. We say, "Hey, don't touch the hot stove. Do this thing instead."

I don't know if it appears in our journals but being more open about those options. I think the hackathons that come up in the open science space speaks to that somewhat. I think you are speaking about being more open about these methodological choices. I'll put a plug in for you, Jessica, about the special issue you're leading in Psych Methods around simulations and really pushing the transparency that we should have in publishing those. Maybe some of you are thinking here is going to figure into that. Stay tuned, everybody, for that special issue. It's brewing.

Adrienne: Thank you. Any other thoughts?

Brian: Well, I agree with everything Jess and Fred are saying. To me, the broader message that it raises is the maturation of appreciating uncertainty, surfacing and recognizing that there are so many choices being made, sometimes by us in our individual project, but then also by those toolmakers and design decision-makers and statisticians that provided us with resources in order to do our work. All of those levels of decision-making create choices which create constraints which have potential implications for the confidence that we can have in our conclusions.

That isn't a fatalist comment. Uncertainty is just part of the work. The fatalist part would be saying, well, yes, it's all uncertain, but whatever, and just ignoring it and proceeding on our merry way of recognizing that our literature is almost always showing positive results. The challenge we face, I think, that's revealed by these projects is that in the way in which we've structured research, the way in which we reward research, we expect positive results to appear in the papers.

As a consequence, our papers are neat and tidy and seem to support claims very clearly. When you do census analyses of how often are people showing support for their hypotheses, it's at 90%, 95% of the time, the papers are supporting the hypotheses of the research. Then these replication projects and otherwise are saying, well, we can find consistent evidence with those original claims only 50%, 60% of the time. Then you asked about registered reports, a publishing model where you make commitments to publish prior to observing the outcomes of the research.

The success rates and observational studies of those is 50% to 60% of the time. We have this big gap between what the literature looks like, we're always right, versus when we have more controlled conditions, oh, we're not showing that evidence a lot of the time. That uncertainty now being recognized is a positive, I think, for us to be able to mature how it is we accumulate evidence and start to make claims and calibrate our confidence in those claims.

Jessica: I think something that comes out of thinking about some of these robustness issues is that that whole process of publication bias that Brian is describing also applies to how methods are developed. The publication bias version of a new method is that you publish a new method. It might not be that it's a unique claim, but there's some new method. It's newer and more complicated in some way than the other methods. They can have the same publication bias because

it's desired by the field to publish a new method as opposed to, I don't know, evaluate an existing one.

There's probably also some uncertainty in how well that method works, but that existence of that new method expands this multiverse. I think thinking a bit about, what's the cost-benefit analysis of introducing new methods when we don't know if they appropriately expand the multiverse? This is, I think, a controversial notion, but should we be thinking about shrinking? Should we say, "Oh, no, actually, these 10 methods were published." We should say, "No, they shouldn't be used."

There's this meta-problem with methodology where our version of publication bias is publishing new ways to analyze the data. Meanwhile, applied researchers are struggling with how to analyze their data, given there's the existence of so many ways to analyze it. I don't think anybody knows what the cost of that is. Are we finding the truth because we have more ways to analyze the data? Are we getting further from finding the truth because there are more ways to analyze the data?

This whole grappling with this aspect of the uncertainty, I think, is going to be really important moving forward. I don't think a new methodologist ever sits down and thinks, "Oh, I'm not sure I should publish this new method." It might just make things harder or more uncertain, or it works well in these conditions, but I really should check if it works on all these other conditions before I publish it, because if I publish it, somebody might use it.

There's some very recent research synthesis that shows a lot of methods are published. The percentages, we're still trying to know them, but they don't ever have a follow-up study about how well they work in another context. It's just like new, new, new, new, and this expands this element. It's like a dimension of uncertainty that expands.

Fred: I think you have Occam's razor outside of your Zoom window there. I totally appreciate this notion of what about what we have? How far have we explored with what we have? The message I take from your point is, how do we tell the difference between making innovations and advancements? How do we tell the difference between sophistication from confusion as we get into nuance? Not that we always need to march forward confidently. In fact, we need to march forward more cautiously, the more sophisticated we would get as a general rule. Anyway, totally resonate with your point.

People at some level need to know what they need to do to get the work done, as Brian was saying. We have software where we don't necessarily know what's under the hood, but we're using it anyway. Not everybody is a methodologist or wants to be or has time to be to figure that out. We have to have some kinds of trust, for better or for worse, to proceed.

Flavio: I'll just complement Fred's and Jess's point to say that, obviously, because they are brilliant methodologists and editors of *Psychological Methods*, they think in terms of methodology, but there is also the human part. We are all researchers. We are people. We have different backgrounds. We have different philosophical commitments. There's also that variance.

When somebody says, "Hey, the replication is 60% to 50%, the replication rate, and when we publish in journals, it's 90%, 95%," to me, discovering these sources of variability and heterogeneity is a great step for meta-science in the coming years, not only on measurement but on methods, on epistemologies. There's a wide range of sources that publication bias, I think, has led us astray.

Now that we are trying to have more open diamond journals, we are trying to implement open science into nearly every part of the way we do our science, possibly these sources of heterogeneity will become available for people to study their origins, the factors that predict this heterogeneity, and et cetera. Like Brian said, embracing that uncertainty, I think, is the future for both applied science and meta-science.

Fred: It's interesting to think that the more we embrace uncertainty, as you put it, Flavio, that we should be more confident in our science by expressing a lack of confidence in our findings, because that's how science works. I wanted to interject this idea as well of what do people do with those findings when we convey uncertainty?

If a doctor tells you there's a 40% chance this surgery will be successful, do you engage in the surgery or not? Do you follow up with the question, what would happen if I don't get the surgery, et cetera? Trying to tie the uncertain findings to dichotomous actions that people have to take in the real world is a challenge for everybody. I guess that gets to the question of science communication, which maybe others can take on here.

Brian: I love that you brought that up, because conveying uncertainty, it's not a, this will weaken science. It's, this will provide the best evidence we can for those that have to make decisions today, because the scientist has the luxury of not having to make any decisions at all about how to use the information. It's just studying whether it's true, under what conditions it's true, what are the claims about, but as a decision maker for having to do things, we have to ideally know the basis of uncertainty in order to make informed decisions.

That doesn't mean we make no decision when it's uncertain. The policy has to be made today. The decision about how the school is going to be run today, how the lesson is going to be delivered, has to be done today. Providing the uncertainty helps the decision maker recognize, I have to make a call, and I can't do it if given information that's overconfident, then I'm going to be misled. It may end up being the right call, but I will do so without the complete information. Embracing uncertainty is really just a matter of embracing reality and arming those that are having to make decisions in applied contexts with the best information possible.

Adrienne: I have a question about-- I'm going to turn a little bit to issues of training. Thinking about undergraduates, graduates, postdocs, how might you suggest changing training to better prepare scientists to do reproducible and collaborative research and create those environments? Maybe we'll start with Flavio.

Flavio: All right. Thank you for this question. I think that training is a very important part of the way we do our science. How we are teaching our students is essentially how we are training our future scientists. More than that, we're also training the future consumers of science. The way that we train folks in light of the replication crisis, in light of open science practices, and et cetera, becomes a fundamental issue for how science is perceived by the public.

I really think that reproducibility and replication should not be confined to one optional open science lecture. They should be embedded throughout methods, substantive training, epistemologies, and et cetera. I think open science has become a really big umbrella under scientific thinking.

Thinking what Brian just said, scientists don't have to make decisions. A part of open science is what we call participatory research, in which we use our scientific findings, and then we participate not only the people that are affected by our research, but also the stakeholders that are empowered

to take important decisions. Once we train individuals to be aware of all the multifaceted areas in which open science methods can be taught, and we speak about this great epistemological umbrella of research, how it's supposed to be taught when you embrace uncertainty and all of that, I think you are educating people in a much better way for the future.

If I may take this a little further, I wanted to speak a little bit about FORRT, the institution that I represent. There, what we try to do is align meta-scientific efforts such as replications, many, many replications and many reproductions, with training. We collected the world's largest database of replications and reproductions. Now we're going to provide a portal that for someone who is teaching, say, for example, age priming or any scientific effect, they can go to this portal and see how uncertain this finding is. Has it been replicated by how many people and et cetera?

We're trying to build on SCORE's work to provide everyone an ability to judge a given effect. Of course, this is endless work. Hopefully, this will be taken up by maybe governments or maybe Brian himself and actually do good work on that. We are trying to initiate that initiative. It's big enough. To mainstream replications and reproductions into the way that people are trained, in the way that you conceptualize scientific training, and the links are on the doobly-doo.

Adrienne: Thank you so much, Flavio. Other thoughts about training?

Brian: I love that response. The other thing I would add is that it does not take a lot of convincing for early-career researchers to embrace the idea of transparency in research or advancement of rigor or trying to achieve reproducibility. These are things that feel very natural for how science works when we are entering it. The barriers to it are practical and reward-based barriers. It's only after being in science for a while where I would get jaded of why would I bother preparing my dataset for sharing, because there's no value in it for me to have a career or advance my career?

Now, these things are changing, but the point being, in the context of training, is that we can leverage what people already expect to be doing in science to be transparent, to be open with this work, to look for areas of uncertainty. Those are things that are easy for people to get on board with.

The complementary item to me that's an important one is a mindset for researchers. There is a premium felt, especially when the literature is filled with all successes, and I feel as an early career researcher, "My God, everybody is so brilliant. They're producing all of these findings, going these new directions, and when I'm in the lab, I'm getting negative results. I'm getting a mess. My work doesn't look like what I see in the literature."

I think that cultivates a mindset of fearing error, of needing to be right, because that's what everybody else is. It seems like they're all right all the time. That's a dysfunctional mindset in the context of doing something that's very hard. Science is very hard. Discovering new knowledge that's truth, that things is very hard.

If we can shift the mindset from being right to getting it right, then I think it can reduce the worry about error. Error is going to happen in doing hard work. I expect that every project I've ever been involved with has errors that I did not detect in doing the project. The opportunity for transparency is to help detect those errors, to hold me accountable to those errors in a positive way, and to give the community an opportunity to scrutinize my work, to help discover those errors so that we can surface them and do better to improve.

If we can really embrace that mindset, then someone identifying an error in my analytic code is not to be met with shame or embarrassment or frustration or anger, but rather gratitude. "Wow, I wanted to get that right. That's my goal." They found something that got in the way of me getting it right, and now we've corrected it, and it's better as a consequence. That would be a very functional community mindset that would be, I think, much more beneficial to a lot of the issues that we're talking about.

Jessica: Just to make you feel validated, Brian, I teach introduction to statistics. A couple hundred students sit in a lecture and learn about the t-test and variance and stuff. When you're teaching the standard error of the mean, which helps you get a p-value, you always say, "If you sample again, you could get a different result. If you sample again, you could get a different result. If you sampled again and again and again, you'd get a distribution of results." The true mean is the mean of all those sampled means.

There's usually a student who's like, "Why don't we just sample again and again?" They're basically saying, "Why don't we do replications?" Because we know that there's uncertainty. These are usually first-year students, and I get to beam and say, "Oh, you just discovered something. It took psychology a really long time to value this idea that why don't we just run studies more so that we can observe the sampling distribution, basically."

I think Brian is really right that when people are learning about the way science works, it's like you have to infect them with the disease of the-- I tell students it's the enterprise of science or the business of science that comes along when you're learning about it. You have these ideas about how it should work, the enterprise comes along, and it changes how you would think about it naturally, but students are also very uncomfortable with uncertainty. Curricular change is hard.

Teaching introduction to statistics, it's like we have to teach this certain content, it seems like, and to add all of this other stuff to it requires this curriculum change. Are we still going to teach these certain kinds of statistics? Are we going to teach these other things that are basically for students to grapple with uncertainty? Then students really want to know the right answer. I feel the nuance you can have in the classroom is getting to be less and less because of pressures with teaching and enrollment. There's this challenge to really engage with students who have this capacity in implementing the curriculum. It's like we need to take a test. Students are less engaged, I think, in recent years since coming back from the pandemic.

I wonder if you guys have thought about bringing some of this stuff into the classroom and how you navigate that, because it seems like the classroom context is getting harder. At the same time, students do have the capacity to learn and engage, and then it seems like there's two things happening at once. One seems positive, but one also seems negative.

Fred: You sound like an excellent teacher, Jessica, to link concepts in an accessible way to undergrads. I hold this optimism that you and Brian have for the future. You heard it here that early-career researchers become late-career researchers. Pretty profound, right? They take what they learn under this new way of approaching open science and carry that forward as they become mentors.

I see this in my own grad students is why I am positive as well, that as comfortable as grad students might be in learning something new, they're comfortable preregistering, sharing materials and data and code, and trying to get things right. It does sound a little pie in the sky when we think about our current structure of the research enterprise.

I think in the future, it's a continuum, it's a messy multivariate space, but moving forward, we are moving into a more positive space, I think, in that you shouldn't feel like you did something wrong if you failed to reject the null hypothesis. You actually did something right. You did something right. If you're having problems getting those results out, of course, there are a multitude of issues in contending with research, but it's not the fact that you failed to reject the null hypothesis. Journals need to remove themselves from this notion that it's to be avoided or a sin.

As we know from the SCORE results that there's a lot under the hood that we haven't been revealing, and we need to push forward in being more, not just transparent in our practices, but turning journals into being more reflective of these processes that are more human, that are more complex. It's almost qualitative in a sense in terms of the individual study reporting these processes, but it becomes less qualitative the more you think about this multiverse in generalizability terms, like the constraints on generality, that kind of statement, around, "Hey, my research is constrained," and that's good context to provide.

If you think of it across research enterprises, what are their constraints? It becomes less constrained as you look across studies. This is more than meta-analysis, by the way. This is more, I guess, Jessica, as you referred to early on, is meta-analysis may be done differently because we can look across a sea of findings, for example, and look at which findings are more related, almost like a cluster analysis, more like a close replication versus conceptual replication. Meta-analysis could approach that in a more sophisticated way these days, I think. I'm just sitting in an armchair talking, but somebody might do the hard work of pursuing that.

Adrienne: Believe it or not, we just have a few minutes left. An hour went so fast, and I'm wondering if I could bring up one other topic for you today, and that is, what do you think about ways to better align incentives like funding, publishing, tenure promotion to reward rigorous and transparent science? I know that's very meaty but just wanted to get some thoughts about that. Who would like to start? Jessica, go ahead.

Jessica: I'd love to hear how other people see this as progressing, but I was involved with this group where we were trying to develop transparent, open practices that could guide tenure and promotion and hiring for research-promoting institutions. I should search their acronym. They're related to DORA, but not exactly DORA. I'll look at it before the end.

Brian: HELIOS is what it's called?

Jessica: What? Say it.

Brian: HELIOS is the US-based one.

Jessica: There's one out of Ottawa in Canada that I think has a different acronym. We sat in this meeting, and we talked all day, and it just seems like a huge challenge. It seems like research-producing organizations, which are universities and other such organizations, they don't really want to change hiring, tenure, and promotion to acknowledge that maybe we shouldn't just be counting impact factors and publications.

Now, some universities are moving away from impact factors of journals, but we talked a lot about, what about if you produce a dataset? What if you publish all of your materials? You really embrace some of these things. It's just everybody has a story about how their university, they tried to change some of these standards, and it just got completely shot down.

I just sat in a room with these people for the day, but I know some people here have been working on these initiatives at broader scale. I'm wondering if there's any optimism that anyone can share, because it seemed a bit like a dead end, like research offices saying, "No, we're not going to change that because this department said that they wouldn't," or it seems difficult to change that at the university level. I'm talking specifically about how universities hire, promote, and tenure people and organizations that are like universities in the sense that their primary outputs are research.

Brian: I would start with the positives, which are the change that's happened the most is at journals. The growth in requirements for transparency, data sharing, code sharing have blossomed across journals, and not just psychology but across several fields. Likewise, the adoption of registered reports as a model that removes outcome-based decision-making for what gets published. It addresses a lot of these challenges of, I need certain kinds of results in order to get a publication.

Instead, registered reports says you need to ask important questions and design really effective tests of those questions, and then the decision to accept is based prior to knowing the results. There are more than 300 journals, many of them in the social behavioral sciences, that have adopted that, and that continues to grow.

Funders is the second area where things are shifting. I think there's more change in funders outside of the US, where perhaps most of our attendees are, but still across the world, there have been shifts in greater requirements for more transparency and movement on rigor with funders. Where the least progress has been made is in institutions, as Jessica describes, and I think, in part, that's because it's the hardest problem.

When you talk to a provost or president of a university and say you need to change your hiring and promotion criteria, they say, "What are you talking about? I can't tell the poets and the political scientists and the chemists how they need to do their work." Those are things that are grounded in those disciplines. They decide what is good work, and we are just a coordinator of policies rather than a generator. We're not going to impose this.

That is a true statement in the sense that there is so much disciplinary variability and norms that it's very hard to ask institutional leadership to impose a change from the top. Really, the change has to happen from the department level. That's where the norms are set for how it is we think about what is good research in our kind of work. That's a long game. It's department by department, and you can get a groundswell of university change over time with several departments in different fields. Also, you can get department transfer within fields.

A great example of this is University of Maryland's Department of Psychology, Michael Dougherty. They went through a process of deciding what are our values and how do we make sure our values are instantiated in our policies and for how we evaluate ourselves and people for tenure, people for hiring, et cetera. They have a number of materials if you're interested in how they went through that process. We'd had a webinar with them to walk through it.

That change has been surfaced, for example, in COGDOP, which is the Society for Department Chairs in Departments of Psychology in the US. That is creating these occasions for saying, "Oh, we can make decisions about this. We can shift how we reward because it's up to us at the department level to do so."

Adrienne: Oh, I'm sorry. Just one minute, Flavio, and then I will have to close this.

Flavio: Just to say that the Netherlands have changed their policies for hiring to include practices on open science, open education. There are certain universities, for example, Utrecht University, who removed itself from these international rankings that lead to this system of incentives. There are some hopes out there that we are going towards a less infected system of incentives.

Adrienne: Thank you. Also, this is actually a nice segue, as you mentioned Maryland, Brian. We are hosting an Essential Science conversation on October 1st about modernizing tenure and promotion, and Mike will be a part of that conversation.

Brian: Perfect.

Adrienne: I know. You can all join us for that conversation. I cannot believe an hour has flown by, but I just want to thank you all so much for this really engaging conversation today about open science. I've just really enjoyed it. Thank you for joining us and lending your expertise and your thoughts and perspectives. Also, thank you to the audience for joining us today.

For the audience, you'll be receiving a one-minute survey after today's broadcast, and we'd really appreciate your feedback. We also want to hear from you about topics that you'd like for us to cover in these forums. You can always email us at science@apa.org with your recommendations. Lastly, you will get a link to this recording and a transcript in a couple of weeks. We really hope to see you at future events. Just a final thank you to the panelists for a fantastic conversation today, and hope that you all have a good rest of your day.