The Open Science Movement is For All of Us

Moin Syed
Department of Psychology, University of Minnesota
moin@umn.edu

Invited lecture for the Department of Psychology, Western Washington University

April 15, 2019

Slides and supplemental materials: https://osf.io/pm8bh/

****

I am troubled. I am deeply skeptical. I believe very little of what I read. That is not just because of high standards or expectations for what counts as evidence, but rather knowledge that we are doing bad science that is leading to individual studies of questionable utility, and thus a severe lack of cumulativeness. If we lack cumulativeness, we lack science.

I said "we" are doing bad science. That is a very intentional phrasing. It corresponds to the collective we—all of us—and is inclusive of me. I am not exempt, you are not exempt, we all engage in practices that compromise our work. And I continue to do so, even as I stand here and elsewhere to talk about these issues. If you do not think so, then I challenge you to allow me and a group of my trained RAs to conduct an audit of your lab workflow. I guarantee we will find some problems with how you do things. That is true now, and that will be true in five years even if you have dedicated every waking work hour to improve your process. It is not about reaching some specific standard; it is about recognition that you can always improve your work, recognition that perfection is impossible. Changing your ways, unlearning poor practices, is a process, one that takes time and effort and is ongoing.

Now many of you have probably heard of the "replicability crisis." If you have not, I would be tempted to say that you have been completely tuned out from the psychology world or, indeed the science world in general, as the issues that are being discussed go far beyond psychology. Nature conducted a study of 1500 scientists from a variety of fields and found that 90% agreed there is some sort of crisis whereas only 3% said there was not. There are certainly some variations across disciplines but none are exempt from the perception of problems.

If you are a student and have not heard of the replicability crisis, then I might say that your psychology program here has failed you. But I know that I cannot say that, as I know that knowledge of the replicability crisis is uneven across the field, and I know that many departments are reluctant to teach their students about it for fear of turning them off from psychology. A group of researches actually did a little study to see how students' learning about the replicability crisis was related to their attitudes towards psychology. They asked them some attitudinal questions one week prior to a lecture on the replicability crisis and then again one

week after the lecture. They found small to moderate effects on students perceiving psychology to be more like chemistry, physics and biology and for them to be more distrustful of the results from psychological studies—I see this as a good thing, we should be skeptical. There were no changes in how much they like psychology, whether they thought psychology was soft, whether they intended to continue in psychology, and so on. So, effects are small and this is just one study, but if anything expose to this content leads to students' perception that psychology is more rigorous than they thought.

Regardless of your prior knowledge, the "replicability crisis" is a thing that exists out there in the world—like most things, failing to talk about it does not mean it no longer exists. It is an idea that means different things to different people. It is a splashy name well-suited for headlines, but it is a rather limited characterization of what the issues are, as our current moment is not only defined by failures to replicate work, nor was it even fully fueled by high profile failed replications.

In fact, the three major precipitating events of the so-called replicability crisis were not really about replication at all, and all occurred around the same time yet independent of one another.

First was Daryl Bem's publication on extra-sensory perception in *Journal of Personality and Social Psychology*. Across nine studies Bem purported to demonstrate participants' ability to detect future events. All traditional (social) psychological methods and forms of inference were adhered to, which is why an obviously wrong paper was published. The publication date is 2011 but the paper was circulating early—I looked back in my archive and the first discussions were in September 2010.

Second was the Simmons, Nelson, and Simonsohn paper on "False-positive psychology," published in 2011. This article introduced the concept of "researcher degrees of freedom" (later p-hacking, garden of forking paths), demonstrating with an empirical example that it is possible to achieve an effect that is $p < .05$ that is literally impossible.

The third major event was Dierdrik Stapel, the Dutch social psychologist outed as a fraudster. Stapel was a highly regarded, productive researchers, and it turns out he was making up his data. Just completely creating data sets from nowhere. Some graduate student collaborators caught on and it turned out his fraud was much more widespread than anyone could have imagined; he is currently up to 57 retractions.

Bem, False-positive psych, and Stapel. Three important events in the history of psychology. All three spoke to the issues of impossible results that were impossible for different, yet overlapping reasons.

All of these events were in 2011 (sort of…Bem was preprinted in 2010), suggesting that this was a critical year for reform. Accordingly, some have suggested that 2012 should be considered Year 1, and that we more or less discard everything that came before it. But really these events served as a crescendo of some kind of crisis that had been brewing for some time. That all three

of these occurred independently at the same time is suggestive of a zeitgeist around research scrutiny, that there was a collective feeling across the field that was manifest in many different ways. And indeed, if you are inclined to do so you can find many papers published around the same time that are part of this zeitgeist. I will give a description of just a smattering of these papers.

John Ionnides and Daniel Fanelli had a series of papers about the problem of selecting for statistical significance, and how this practice is uneven across sciences.

Behavior geneticists eschewed the “candidate gene” approach due to under-powered studies that led to lack of replicability. First they required that all candidate genes studies include a replication, then they largely abandoned them all together. Importantly, this is not the case in other areas, such as social and developmental, who continue to use this approach, demonstrating the unevenness of reforms across areas.

Indeed, there was recently quite a lot of attention around a meta-analysis of the 5-HTTLPR serotonin transporter gene and links to depression, indicating no reliable effect. But there have been other meta-analyses showing this, going back to 2008, so problems with this work have been known for some time.

In 2009 Ed Vul and colleagues published a controversially-titled article on “voodoo correlations” in social neuroscience research—they ultimately had to change the title for the final version. This paper showed that the methods in this line of research were likely producing a huge number of false positives, for a variety of reasons. Tal Yarkoni wrote a commentary specifically calling out the under-powered nature of these studies. Four years later Katherine Button and colleagues published their “power failure” article that really formalized these previously-raised concerns. Yarkoni’s paper has been cited 300 times, Button et al.’s over 3000 times.

In 2010 Henrich et al. published the WEIRD paper, arguing that not only are the majority of studies in psychology based on populations in Western, Educated, Industrialized, Rich, and Democratic societies, but that these populations are “weird” in relation to the rest of the world, that they are the outliers vs. a general model for universality. Arnett had published a similar paper in 2008. Never under-estimate the power of a catchy acronym, as the WEIRD paper has been cited over 5000 times whereas Arnett’s just over 1000 times.

And so on. This is just a selection of greatest hits, there are many more examples of issues that were being raised around the same time, somewhat independently, most of which were not directly about failures to replicate specific studies. Much of that came later, after the general idea of “problems” came to light during this zeitgeist.

Moreover, all of this rested on a foundation of knowing that there were problems. It was that researchers just started to scrutinize methods at this particular time.

Sterling published a report in 1959 on the problem of publication bias, with very few published studies rejecting the null hypothesis.

Paul Meehl, the patron saint of Minnesota psychology (we light a candle for him prior to each workday), published throughout the 1960s and 1970s on the senseless routine of hypothesis testing and how most of the research in our field is essentially meaningless.

Around the same time Imre Lakatos articulated the distinction between progressive and degenerative lines of research, referring to how we related new information to our existing theories. Much work in psychology can be characterized as degenerative because we make modifications to our theoretical ideas in light of disconfirming evidence because we have to not because it improves the explanatory power of the theory.

Jacob Cohen first wrote about statistical power in 1962 and continued to do so throughout his career, arguing that psychology relies too heavily on under-powered studies and, paired with the ritual of null-hypothesis testing, means that we are producing a large volume of meaningless results.

Anthony Greenwald tried to institute reforms from the editorial side when he oversaw JPSP in the 1970s, highlighting problems with replication (even calling it a "crisis") and the need for thorough and transparent reporting. He had to resign before his term was over.

Rae Carlson wrote a few papers in 1972 and 1984 criticizing the methods of personality and social psychology, that our work had become too distant from the subject matter and diminished the meaningfulness of what we produce.

Norbert Kerr's article on HARKing – Hypothesizing after Results are Known – came in 1988. Everyone knew about this and how problematic it is, yet nothing really changed in practice. Probably all of us who have submitted at least a few papers for publication have received comments from editors and reviewers that we should modify our hypotheses to be more in line with the results. Perhaps some of us have even made such comments ourselves.

The lack of diversity of samples is an issue that has been raised for ages. This issue has been raised with regard to global diversity, for example by Arnett and others, as well as racial diversity, such as by Sandra Graham in 1992 and local celebrity Ana Marie Cauce in 1998 and really throughout her career. Notably, the WEIRD paper actually includes no meaningful discussion of racial/ethnic diversity.

All of this is of course relevant for replication but if you think of this only in terms of replicability then you are missing the bigger picture. For this reason, as well as an attempt to draw attention to the positive reform work being done, Simine Vazire has termed the current moment the "credibility revolution." Whereas the credibility revolution is a nice reframing of the issue that gets beyond some of the limitations of the term "replication crisis," it is still incomplete. Credibility is critically important, but it is really just one component of an even broader issue.

Indeed, at the current moment people are identifying several different crises—replicability crisis, theory crisis, measurement crisis—all claiming that they are highlighting the "real" problem. Of course, none of these is more important than the other, and they are all related.

Let's take a quick detour into my own substantive domain—research on identity development. We can propose and test a simple model, that cultural socialization leads to ethnic/racial identity development, which in turn leads to positive adjustment. And indeed, this model, or some variation of it, has been tested quite a lot. But if we look closer at this middle process of ethnic/racial identity, we can see quite clearly that there is a lot going on. Even just looking at the two dominant models, the Developmental Models of Ethnic Identity and the Multidimensional Model of Racial Identity, and the two corresponding measures, the Multigroup Ethnic Identity Measure and the Multidimensional Inventory of Black Identity, we see a profusion of subscales and constructs, some of which sound different but are actually identical (jangle fallacy) and others that sound the same but actually tap into different things (jingle fallacy). Any given study that you select will use or more of these measures, and there is rarely any good reason for one to be used over another. Moreover, rarely if ever do you see differential predictions for these dimensions. Rather, a vague hypothesis of the role of "ethnic/racial identity" will be advanced, and then one or more of these measures will be put forward as the operationalization. It should be clear that this is a major barrier to cumulativeness or even being able to make any kind of conclusions at all. Is this a problem of replication? Theory? Measurement? Yes, it is all of those things.

The problem is exponentially worse because you have the same issues with the other two elements of our rather simple model—cultural socialization and well-being. Each of those has a profusion of constructs and possible measures. When you combine all of these together, you get a collection of findings that are impossible to synthesize.

We can continue to identify barriers to synthesis by pointing out that ethnic/racial identity is only one domain of a broader identity that comprises many domains, and that identity is just one part of the much broader concept of personality. Is this a problem of replication? Theory? Measurement? Yes, it is all of those things.

The current moment is all about scrutiny, evaluation, and criticism—not just of other people's work, but your own as well. It is about working hard to do better science. If your response to all of this is, "my science is just fine, thank you." Then I am sorry, but you are not scienceing hard enough.

Enter "Open Science." Open science on its face is a broad term that may not have much meaning, or perhaps has limited meaning. Let's put some structure on it.

To be clear, there is no single, agreed upon definition of what open science is, what it pertains to, and so on. Some people have provided some definitions:

Katie Corker, the current president of the Society for the Improvement of Psychological Science wrote the following in 2018: “I want us to think hard about classifying Open Science as a **behavior**. Not as an identity. Not as a value. It is a set of practices that you **do** in order to make your work transparent to others, checkable and scrutinizable by others in the community.”

Munafò et al. (2017) defined open science as “the process of making the content and process of producing evidence and claims transparent and accessible to others” (p. 5).

Spellman et al., 2018 defined open science as “a term for some of the proposed reforms to make scientific practices more transparent and to increase the availability of information that allows others to evaluate and use the research.” Their chapter also included a glossary of terms, where they provided the following entry: “open science: a collection of actions designed to make scientific processes more transparent and their results more accessible.”

These definitions highlight that open science is about specific practices/behaviors that are mostly meant to increase transparency and accessibility. This is good but still a bit vague and imprecise.

Here is my working definition: *Open science consists of principles and behaviors that promote transparent, credible, reproducible, and accessible science.*

Corker really stressed that open science is about behaviors—actions you take—and she stressed the behavioral aspect in opposition to thinking about open science as a social identity—that there are open science people and not open science people, and you are either one or the other, all or nothing. That is clearly not a positive way to think about open science. At the same time, I think that stressing behaviors might also be limited.

For this reason, I define open science as a set of four principles, each of which has many associated behaviors. The reason for this framing is that it provides some higher order structure that helps guide, understand, and organize the different behaviors one might engage in. You could call these principles “values” as well, and that would not be wrong.

Four principles and associated behaviors:

Transparency pertains to researchers being honest about theoretical, methodological, and analytic decisions made throughout the research cycle. Transparency means rejecting incomplete or opaque explanations that can mislead others. It is about reducing, as much as possible, the knowledge asymmetry that is inherent to the process (see Vazire, 2017). Transparency is not only about fraud or otherwise intentional misreporting of the details of research. To paraphrase Feynman (1974), it is very easy to fool oneself without intending to. Thankfully, there are safeguards that can be put in place to reduce the frequency and severity of sub-optimal behaviors due to self-deception (e.g., preregistration).

A behavior that you can engage in right now to enhance transparency, for which there are no barriers, is to preregister your next study. You can learn more about that at the workshop tomorrow. Alternatively, if you are writing a paper right now, review your method and results sections and ask, with an extremely critical eye, whether you are reporting all of the important details of the study and how you treated the data.

Credibility is closely related to transparency: it pertains to how others will view your work and the work of a field in general. Credibility is the degree of trustworthiness and believability of the research reported in the literature. For a variety of inter-related reasons associated with suboptimal research practices, large swaths of the current body of research is not credible. The credibility of research is enhanced when it is transparently reported, well-powered, subject to replications, and relies on samples that reflect human diversity. Transparency is a precondition for credibility. Credibility was once determined by author status or by journal impact factor (correlation between retraction and impact factor). Preregistration can also enhance credibility as can making your data, materials and analysis code open. But even then, there can be some cherry picking of methods or the results may not be robust (e.g., many analysts, one dataset project). Accordingly, methods such as multiverse analyses or specification curves are strong approaches to addressing robustness and facilitating credible evaluations. Replication also facilitates credibility. If you can demonstrate a result in repeated direct replications that are preregistered and sufficiently powered, that will then enhance credibility (e.g., our redemption study).

Diversity of samples also falls under credibility. We need to appropriately generalize based on the nature of our samples, and be sure we are using samples that match our research questions and goals for generalization. We must ask ourselves, who is the population for my study? This is something we often do not do. Moreover, this should be done beforehand. One of the most irritating defenses of failures to replicate that I have heard from social psychologists is "contextual sensitivity." Contextual sensitivity on its own is an important concept that has been widely uses by all kinds of psychologists. But when it is used after the fact, when failures to replicate emerge, and the original study includes wild claims of broad generalization, when the findings are suggested to have implications for policies and practices without respect for context, then *ex post facto* claims of contextual sensitivity represent a degenerative approach to research and are not credible. For this reason, as well as others, issues of diversity are central to the open science movement.

A behavior that you can engage in right now to enhance credibility is to conduct robustness tests. Did you include controls in your model? Test with and without. Did you transform your data? Test on raw and transformed. Determine who your population actually is, then generalize to that population. If you are planning a study, build in a replication or plan to collect enough data to split your sample into two halves.

Reproducibility pertains to how well we keep records of what we do, at all phases of the research cycle, so that everything can be reproduced when needed (not *if* needed, because there will always be a need for reproducibility). Reproducibility is not only about enabling other

people to reproduce your work. Indeed, in a saying attributed to Mark Holder, "Your primary collaborator is yourself six months from now, and your past self doesn't answer e-mails." Using clear and consistent documentation in your workflow allows you to better remember what you did, why you did it, and importantly, allows you to do it again. Note that replication is not the same as reproducibility, although the terms are often used interchangeably. For me, replication is more closely aligned with the principle of credibility, whereas reproducibility is a broader principle that encompasses many different behaviors.

A behavior that you can engage in right now to enhance reproducibility is to examine your workflow. How much have you ever even considered your research workflow? When you are planning a new study, do you formally write down your research questions and hypotheses? Do you develop a corresponding analysis plan? Do you have a clear plan for how and when you clean and prepare newly collected data sets? Do you clean your data manually or using automatic scripts? Are clear and detailed codebooks available for all of your data? Do you have multiple versions of the same dataset? Do you have naming conventions for your variables? Are all of your files stored in a single location and clearly labeled? Do you have 17 versions of the same manuscript or do you practice version control? How do you handle collaborative writing and revisions of manuscripts? If you are anything like I was (and still am in many ways), you have not spent sufficient time thinking about these issues. My guess is that if you do so you will rather quickly identify some poor practices that could be changed relatively easily. Try to identify 1-2 practices that you can change right now. Do those, then identify 1-2 more practices to change, and so on.

Accessibility pertains to making all aspects of the research cycle open and available for those who are interested. This includes not only openly sharing data and materials, but sharing them in a format that is understandable, useable for outside researchers, and is easily accessible via functional weblinks or some other mechanism. Note that to be able to do this you have to have your data in a format that you yourself understand, so in this case the principle of reproducibility must really precede accessibility. Accessibility also includes making research products freely available to researchers and the public around the world. For example, posting non-copyrighted post-prints of articles and chapters on a server such as PsyArXiv increases accessibility. Similarly, posting psychological tests and measures online in an accessible format, free of cost to the user, can contribute to their broader use and thus broader range of samples from which we gain knowledge.

A behavior that you can engage in right now to enhance accessibility is to post one of your articles to PsyArXiv**.** PsyArXiv.com is a "pre-print server" for psychological research. A pre-print is a broad term used to refer to a version of a manuscript that has not yet been accepted for publication in a journal. This can include manuscripts that have not yet been submitted or those that are currently under review. Posting pre-preprints can facilitate receiving valuable feedback from colleagues and can speed up dissemination. PsyArXiv also accepts post-prints, which are final versions of manuscripts that have been accepted for publication. In most cases you cannot post the final journal-formatted version of the paper, but you can post your own Word doc (or equivalent). You can look up the journal in which you article appears on SHERPA/RoMEO to

determine if uploading post-prints is permitted. Uploading post-prints facilitates dissemination and increases access to scholars who may not have subscriptions to certain journals. This is especially the case for book chapters, which are notoriously difficult to locate. Uploading your chapters can greatly increase their reach.

Open science clearly covers a lot of ground, and those who are new to these issues can be quickly overwhelmed with the myriad tools and behaviors that can be used to advance open science in their own research. Honestly, even those who are very well-acquainted with the issues at hand and have been engaging in open practices for years can sometimes feel overwhelmed (e.g., me). Additionally, open science is not an "all or nothing" enterprise and as noted open science is not an identity group that people belong to or do not (Corker, 2018). Rather, open science is a varied array of principles and behaviors you can use in your research.

Importantly, and this is really critical, none of these principles are area or method specific. Some of the manifest behaviors may be, but the principles are not.

Moreover, and of course I am clearly biased, but I find it difficult to argue against the value of these principles. But there are arguments against.

A group of researchers published an opinion piece, "Open Science isn't Always Open to All Scientists." A key quote is "we need to reject the reactionary response of assuming that open science is without risks." There were many arguments advanced, including that there are financial limitations to open access publishing, and thus there is a risk of reinforcing existing hierarchies; that in general early career researchers may be at a disadvantage because more senior colleagues may not value such activities; that open peer review could lead to greater bias and retaliation; and that some data cannot be shared for privacy reasons or the risk of being scooped by other researchers who use your data.

These are all good points and issues that require great thought and scrutiny, but they are not sounds arguments *against* open science. The one that I personally think is a non-issue is the risk of being scooped. I am not alone in this feeling, as the "Open Scoop Challenge" was started in 2014, seeking to document any cases of actual scoopage. As far as I know it has yet to do so.

Rather, these are all arguments that we need to be careful how we proceed and that we cannot expect all researchers to do all things open science; there are limitations. Indeed, this has been a strong message in recent times, with several metaphors being advanced: Katie Corker's "not all or nothing"; Michele Nuijten's "good cherry picking"; Christina Bergman's "buffet model." This approach is certainly sensible and likely to be the dominant way of thinking in the years to come.

And now I will abruptly end and open it up for discussion.