Paul Meehl's Vision for Psychology as an MSRP

Stanford Howdyshell

Asian University for Women, Bangladesh.

Presented at International Conference on Engineering, Economics, Management and Applied Sciences (ICE2MAS-24), Bangkok, 21-24 December 2024, organized by Academy of Art, Science and Technology (AAST).

https://doi.org/10.37082/IJIRMPS.ICE2MAS-24.7

Published in IJIRMPS (E-ISSN: 2349-7300), ICE2MAS-24

Abstract

In his 2021 Paper, "The Role of Replication in Psychological Science", Samuel C. Fletcher argues against Paul Meehl's view that psychology should be understood through the lens of a Methodology of Science Research Programme (MSRP), in the Lakatosian sense. An MSRP is a scientific theory with a hard core of widely accepted and nigh-unchangeable beliefs at the center of a science, which is then "surrounded" by many auxiliary hypotheses. When falsification occurs, then, the hard core of the theory remains while the auxiliary hypotheses are changed to accommodate the failure to accurately predict the results of a given experiment. Fletcher criticizes MSRPs of being unable to handle largely statistical sciences, such as psychology (which relies mainly on null-hypothesis significance testing), because they do not involve individual instances of falsification or correct prediction as MSRPs require, but instead on correlations within large samples (i.e. a significance test with a correlation of 0.6 and a sample of 100 people would count as 60 correct hypotheses and 40 falsifications). This paper will argue that Meehl's view agrees with Fletcher insofar as he rejects the compatibility between significance testing and MSRPs. It will develop Meehl's critique of significance testing and show that simple significance testing is not testing a valid hypothesis within the MSRP of psychology. I will then turn to Meehl's, unfortunately ignored, plan to develop psychoanalysis into an MSRP to show what a successful hypothesis looks like for Meehl. A successful hypothesis will not be the accurate prediction of a correlation but rather a precise prediction of a value on various scales that represent the phenomena under investigation.

Keywords: Philosophy of Science, Philosophy of Psychology, MSRP, Replication Crisis

Introduction

In this paper I will investigate Samuel C. Fletcher's criticism that psychology cannot be a Lakatosian MSRP and respond in line with Paul Meehl's conceptualization of psychology as a healthy psychological science and how it can remain an MSRP. To do so, I will start with a brief explanation of MSRPs, then launch into Fletcher's criticism, which is that psychology, as a statistical science, does not offer strictly falsifiable results, which are necessary to the functioning of MSRPs. I will then form a Meehlian response by first investigating Meehl's criticism of Significance testing, then how one can formulate a science that does not rely on it. Doing so will undermine Fletcher's criticism and allow this new understanding of psychology to maintain itself as an MSRP.

Psychology as an MSRP

A MSRP is characterized first by having a hard core of principles and beliefs that are central to the scientific program which are then 'protected' by a soft outer layer of auxiliary hypotheses (Lakatos 1978). The hard core is the foundational principles and beliefs of the programme which are essentially set in stone and whose rejection constitutes a rejection of the programme itself. For example, in cognitive psychology, the correlation of experience to internal mental states and viewing "learning and memory as forms of information processing" (Smith 1997, 832) would be part of its hard core. To reject internal mental states or the information processing model is to reject the paradigm itself. The hard core then requires auxiliary hypotheses in order to be instantiated or shown in experimentation (or other forms of data collection in less experimental MSRPs). To show this, Lakatos uses the example of Newton's laws of planetary motion, where in order to observe planetary motion to corroborate or falsify Newton's laws, we need to look at the planets through a telescope, which comes with a set of beliefs about how light works such that the telescope can function, that there are not mitigating factors that would displace the planet from its expected trajectory, and so forth (Lakatos 1978, 16). If we checked our hypothesis about the planets location and it was not there, we would not throw out Newton's laws (assuming we are Newtonians) but rather reject one of the auxiliary hypotheses (maybe our telescope is not functioning properly or there was a mitigating factor that displaced the planet).

The purpose of a MSRP is then to increase the programme's empirical content (Lakatos 1978, 48) through "ever more complicated *models* simulating reality" (Lakatos 1978, 50). In other words, the MSRP is to grow to explain more and more phenomena through the continual corroboration of a growing set of self-consistent auxiliary hypotheses. To continue with the example of cognitive psychology, this takes the form of the expanding field of experimental work (i.e. where one hypothesis is corroborated, so it is in turn used to build new experiments that grow the programme's empirical content and the development of psychological therapies (Meehl 2006c, 251). For example, the use of priming research to develop aspects of cognitive behavioral therapies (see Yang et al. 2020 and Borgeat et al. 2013 for examples).

Falsifications and corroborations then lead to problemshifts, where auxiliary hypotheses are rejected or elevated with respect to the empirical content. These problemshifts can be either progressive or degenerative, where a progressive problemshift is when a "series of theories leads us to the discovery of new facts" (Lakatos 1978, 34) and a degenerative one merely acts to save the programme as it stands without leading to an expansion of its empirical content. Psychoanalysis, for example, became a favorite target for claims of degeneration because of its failure to produce consistent experimental corroboration (Grant and Harari 2005), while always having some way of explaining away its failures through the manipulation of auxiliary hypotheses, or as Simone de Beauvoir cleverly put it twenty years earlier as displaying "an embarrassing flexibility on the basis of rigid concepts" (1956, 65).

Meehl casts psychology as an MSRP (or, more accurately, as a group of MSRPs) and through doing so shed light on how one would determine if a psychological programme was healthy or degenerative. He argued that well established theories, that could rightfully call for consistent adjustments to their auxiliary hypotheses without becoming degenerative, were ones that lived up to what he called the "money in the bank" and "damn strange coincidence" principles (Meehl 1990, 115).[1] The first, he claims, "gives the conditions under which it is rational to conduct a Lakatosian defense" (Meehl 1990,

[1] This seems to respond to the criticism that MSRPs cannot offer any normative claim on when a MSRP should be abandoned as degenerative (Fletcher 2021, 13). Though it is not a perfect response because much of it is still determined by the judgments of particular scientists, often based on their intuitions (Meehl 1990, 115), it at least gives guiding principles by which the scientist can make those judgments.

115) which is that a "theory having accumulated credit by strong successes, having lots of money in the bank" (Meehl 1990, 115). In other words, the "money in the bank" principle is that MSRPs become more worth defending and adjusting the more empirical content they explain and the more predictions they make that are corroborated. Furthermore, for Meehl, these corroborated predictions put more money in the bank the riskier they are. A more precise prediction, with a higher range of possible outcomes that could falsify the conjunction at the antecedent, the more impressive it is that it is corroborated, and the more likely it is that the MSRP that led to it has something going for it.[2] To use Meehl's example to clarify this: suppose a meteorological theory were to predict rain in April. This would be corroborated but unimpressive because it did not take any risks and simply predicted what every other theory (and common knowledge) was able to predict. On the other hand, if a meteorological MSRP predicted that it would rain 10+/-2 days in April and was corroborated, we would be more impressed because it made a riskier prediction. We place more confidence in it because the theory's hard core and auxiliaries were able to make a precise claim about the world and back it up. Finally, if the MSRP were able to predict which ten days it would rain, it would be all the more impressive and put even more money in the bank. If the last MSRP were to then make a falsified prediction, we would want to save it because it had previously delivered in such an impressive fashion, whereas if the first, that just predicted that it would rain, were to fail later, there would not be much motivation to save it.

The second principle, that of the "damn strange coincidence", states that an MSRP is more worth defending if its successes would amount to a staggering coincidence, such that the best explanation is the verisimilitude of the MSRP (Meehl 1990, 115). Meehl looks to the historical debate about the existence of molecules (whether they existed in reality or were merely theoretical constructs) as an example. In it, he argues that the sheer number of different ways that one can accurately calculate Avogadro's number (he counts 13) makes it more likely that something is actually being counted rather than the possibility that a mere theoretical construct being used to solve one specific problem also accurately solves 12 other specific, not closely related, problems. In other words, while it is possible that molecules are merely a theoretical construct, it would be a damn strange coincidence that they could be calculated in so many ways (Meehl 1990, 117-120).

It is worth reemphasizing here that neither of these confirm the theory (that would still be a logical fallacy), but rather that these reasonably raise confidence in the future ability of the programme to make predictions that are corroborated and to expand their empirical content in the future. This makes the adjustments made to the protective shell more likely to be progressive problemshifts and thus the programme to be more worth defending.

Statistical Sciences Cannot Be MSRPs

With the framework of psychology as an MSRP established, we can turn to Fletcher's argument against them. He establishes first that psychology, as a science, is built around statistical models that focus on correlations between phenomena. This can be seen clearly in psychology's reliance on significance testing. For example, a 2012 study found that analytic thinking was negatively correlated with religiosity by priming subjects with analytic puzzles before evaluating them through three different religiosity scales (finding the correlation to be -0.22, -0.15, and -0.18, where a 0 is no correlation and +/- 1 is a one-to-one positive or negative correlation) (Gervais and Norenzayan 2012).[3]

[2] Meehl specifies, though, that these MSRPs are almost certainly not true, but are more or less reliable models of reality that can take on more or less empirical content and lead to better or worse predictions about the world (Meehl 1990, 114).

[3] This study failed to replicate in 2017 (Sanchez et al., 2017).

As a statistical science, Flecther argues, they cannot be subject to falsification, which is essential for MSRPs. This is because falsification relies on the logical syllogism of modus tollens where we have a conjunction of premises consisting of the hard core and protective shell of the MSRP as the antecedent of the conditional and the hypothesized experimental result as the consequence. Then if the hypothesized result does not obtain, then, following the syllogism, we know that the conjunction of the hard core and protective belt cannot be true (i.e. (A&B&C&D…)->H, ~H, => ~(A&B&C&D…) (Fletcher 2021, 13). Statistical sciences are not liable to this form of falsification, though, because "if the conjunction yields only (nonzero/one) probabilistic predictions, then no observation can conflict with them" (Fletcher 2021, 13). Thus, there is no instance of a falsification because "Data may be quite (if not perfectly) unlikely, yet this warrants no inference via modus tollens" (Fletcher 2021, 13) and "logic thus demands no rejection of statistical auxiliary hypotheses" (Fletcher 2021, 13). In other words, without predicting a precise result (a zero/one correlation) that can obtain or fail to obtain, there is always enough ambiguity in the result such that the modus tollens syllogism is not necessitated. One could, of course, decide that the data is so unlikely that one ought to reject the conjunction, but this is an act of personal (or group) judgment rather than logical necessity.

Fletcher has thus set us up with a dilemma. On the one hand, we can reject the framework of MSRPs when evaluating psychology (and any other statistical science) as Fletcher does, or we can preserve the MSRP model of evaluating the sciences by rejecting the role that psychology gives to significance testing, i.e. making it so that psychology does not rely on predictions that hinge on simply whether or not there is a nonzero correlation. Each of these options comes at a cost, though. If we follow Fletcher, we are not simply rejecting the MSRP model with respect to psychology, but since it seeks to demarcate science as such from pseudo-science (Lakatos 1970), and healthy science from degenerative science (Lakatos 1974), rejecting the MSRPs from applying to a large subset of the sciences would be tantamount to rejecting MSRP as an evaluative paradigm as a whole.[4]

Meehl, on the other hand, maintains that psychology is an MSRP through a critique of the role of significance testing in the psychological sciences, as we will see below. The cost that Meehl incurs, then, is a strong negative diagnosis on the health of psychology as a science.[5]

Meehl's Rejection of Significance Testing

While Meehl does not specifically address Fletcher's argument that MSRPs cannot be statistical, he does agree that they should not be built around primarily predictions of correlations, which he calls directional predictions (i.e. where the content of the prediction is only the "direction" of the correlation, being positive or negative). In this section I will explain an (unexhaustive) series of key reasons why Meehl rejects significance testing of directional hypotheses.

First, and most simply, Meehl does not believe that these directional corroborations generate much "money in the bank." In other words, they are low-risk hypotheses that do not do much to establish the credibility of one programme over another, akin to predicting that it will rain in April. Thus, if everything is going well, and falsifications are not occurring, the programme will appear fine and healthy, but when a falsification occurs, there is not enough accumulated trust in the programme to motivate a Lakatosian defense. This is because all that significance tests "corroborate is the 'theory' that something nonchance must be at work in one direction" (Meehl 1990, 123) and that "there is a pretty big

[4] Fletcher is happy to do so, arguing that each science should be evaluated on its own terms from the ground up rather than from a "top-down" standard such as an MSRP.

[5] Though the current replicability crisis in Psychology may make this cost easier to bear (Open Science Collaboration 2012; Open Science Collaboration 2015).

class of actual and possible [theories] easily capable of generating a directional expectation along these lines" (Meehl 1990, 123).

Second, and more damning, is what Meehl calls the crud factor (1990, 123), which is "when the sample size is sufficiently large to produce accurate estimates of the population values, almost any pair of variables in psychology will be correlated to some extent" (Meehl 1990, 124). In other words, in any sufficiently large data-set, correlations will arise that are not related to the mechanisms being investigated and upon which the hypotheses are based, resulting in a scenario where one could "come up with a sizable number of apparent 'substantiations' of the theories even if they had negligible verisimilitude and there were no intrinsic logical connections between the theory and the pair of variables employed for 'testing' purposes" (Meehl 1990, 124). The paradigmatic example of this is seen in correlating economic, education, and health outcomes with population based on ethnicity or geography, which in turn correlates with socio-economic status, which in turn correlate with "personality or status characteristics that are not part of the definition of social class" (Meehl 1990, 125).

It is worth noting here that the correlations that occur are not false positives (Meehl 2006d, 454). These are real correlations that occur in the data-set. It is just that they could result from so many different plausible mechanisms that they do not lend much credence to the hypothesis they are meant to be corroborating. Since any well-educated and reasonably intelligent person with 15 minutes to kill could spin off a half-dozen equally plausible explanations for the correlation (that align with research being done in other programmes and as such are similarly well motivated), a correlation through significance testing hardly puts any money in the bank and is not a damn strange coincidence because a half-dozen competing theories predicted or could explain the same result.

This leads to the problem with the *ceteris paribus* clause in psychology. The *ceteris paribus* clause is part of the conjunction in the antecedent that states that "all else is equal" and is often the first to go if a theory is falsified because a complicating factor can be found that would throw off the experiment (Meehl 1990, 109). It can then be used, if it can be plausibly shown that there was a complicating factor, to not only defend the hard core of the theory but also much of the outer band of auxiliary hypotheses.

The problem here is, given the crud factor and the difficulties that arise when trying to isolate mental phenomena,[6] there can almost always be a plausible explanation for why the *ceteris paribus* clause failed to hold. There can always be a problem with the data-set where a crud correlation drowned out the one the experiment was supposed to corroborate, or a problem with experimental implementation (not design) where the experimenter accidently primed the subject.[7] In turn, then, the problemshifts can be limited because the response to a falsification can become "do more to limit mitigating factors" rather than adjust one's (other) auxiliary hypotheses in (ideally) progressive ways.

Worse yet, the weak *ceteris paribus* clause can render corroborations trivial. If it can be plausibly shown that the clause is not true, then the conjunction in the antecedent would also not be true. Thus, we have a case of A->B, ~A. This, of course, is valid regardless of the truth of B. Thus, if the *ceteris paribus* clause is weak, then it is not just that a corroboration is not deductively valid and merely puts money in the bank, but a corroboration and a falsification are just as good as one another. The money in the bank could have just as plausibly come from another source.

[6] See Barsalou 2016 for an example of this with priming studies.

[7] For Meehl, this explains why Kuhnian crises tend not to occur in psychology. There is always enough plausible deniability about any given anomaly that it cannot provoke a breakdown. Instead, old theories slowly phase out as they degenerate and fail to expand their empirical content.

Finally, Meehl points out that significance testing benefits from publication bias (Meehl 1990, 125). Much has been said about publication bias and its role in the ballooning of false positives in psychology.[8] A major example of this is what is known as the file drawer problem, where when an experiment fails to yield a significant result or disprove the null hypothesis, it does not get published such that the community would need to reckon with the falsification. Rather, it gets filed away (Romero, 2019, 4). Worse yet, falsifications being forgotten is often the best-case scenario. Oftentimes the data is reanalyzed (sometimes in questionable ways)[9] such that the null hypothesis can be, on later analysis, disproven. Thus, with a little massaging, a falsification can become a corroboration. While this problem is not necessarily exclusive to significance testing, it is exacerbated by the relatively low bar needed to clear the threshold of "results that are significant enough to publish" because it is easier to massage one's data to a low correlation with the right p-value[10] than it is do so to a more precise prediction.

This then compounds the above-mentioned problems with significance testing, where there are a significant number of seemingly random correlations in any given data-set. Even without QRPs, if all the data-sets where the random correlations don't align with the prediction are filed away, the one data-set with the random correlation that aligns with the prediction will appear much more significant. It will not appear as a case of three labs falsifying and one lab corroborating, for example, but of one lab corroborating. This makes it difficult for significance testing to live up to Meehl's "damn strange coincidence" principle because we do not have access to the information through which we can discern if it is such a strange coincidence that so much data is accumulating to corroborate a theory or if it is simply that only the occasions that it does get published.

Meehl's Framework for the Future

Through is work as a therapist, Meehl developed a respect for psychoanalysis, while maintaining that though it was empirical, it was not systematic or experimental (Meehl 2006b, 71) and thus was not a scientific research programme. He hoped, though, that some later researchers would seek to raise it to the status of a science and laid out the framework for how they might do so. This framework, importantly, largely did not involve directional significance testing. We can thus use it as a model for what psychology would look like that does not, in turn, rely on directional significance testing and that would maintain its status as an MSRP despite Fletcher's criticism.

One initial option, that Meehl would later distance himself from, was to replace directional significance testing with predictions of a band of correlations. To return to a previous example, if, hypothetically, instead of Gervais and Norenzayan simply predicting a negative correlation between analytic thinking and religiosity, they would have predicted that the correlation was somewhere between -0.15 and -0.25 then their theory would put much more money in the bank. And if they were able to do three or four experiments where their theory nailed the correlation value, it would become a stranger and stranger coincidence if they were wrong. Furthermore, it would seem to lower the chances of the crud factor leading to false positives, because it is less likely for a random correlation to fall within a predicted band than simply having the same direction. This would compound as more experiments are carried out.

[8] See Ferguson and Heene 2012 and Friese and Frankenbach 2020 for two examples. Also see Stevens 2017 and Ioannidis 2005 for the role of questionable research practices that this often encourages.

[9] This does not have to be malicious (and is often encouraged (Bem 2003)) nor does it have to be completely unjustified. Rather, an accumulation of somewhat plausible adjustments to the data-set can flip an experimental result from a falsification to a corroboration.

[10] See Head et al. 2015, for an overview on P-hacking.

Doing so, of course, would call for a much more robust theory (Meehl 2006a, 174). As the experiment stands, a -0.10 correlation would also have confirmed their hypothesis, as well as a -0.3, a -0.5, or a -0.8. Any correlation would do as long as it disproved the null hypothesis. Had they predicted a band from -0.15 to -0.25, a -0.5 would have been a falsification and as such they would have needed some theory that is sophisticated enough to explain why one would not expect such a high negative correlation (i.e. primed analytic reasoning is negatively correlated to religiosity, but the effect is not so strong as to overcome the effects of personal history, differences in religious tradition, social pressure, and so forth). This thus raises the bar for theorizing and sets the stakes for prediction higher.

While putting more money in the bank and yielding more damn strange coincidences, this tactic would still not solve Fletcher's problem because while making the predictions more precise, psychology would still remain a statistical science. In other words, no direct falsification would occur because the prediction is still of a nonzero/one correlation which, though maybe unlikely, does not guarantee that the result will fall within the band. The counterpoint would be that an accumulation of failures would make it increasingly unlikely that the MSRP is reliable and eventually the scientific community would rightfully lose respect for the current MSRP and make changes, but even still, there would be no direct logical link to the failure and as such it would not necessarily stand as a falsification.

Meehl's solution to this problem was to model the form of psychological studies on the harder sciences: to make a point prediction within an error band (Meehl 1990, 117). How one would do this, according to Meehl, is to develop scales that represent the phenomena under investigation. The IQ scale is the paradigmatic example of this, but they can be developed for all sorts of phenomena. To return to Gervais and Norenzayan, the prediction would change from predicting the direction of the correlation to predicting what the value on a numerical scale the analytically primed subjects would receive or the difference in value between primed and unprimed subjects. So, on a hypothetical 100-point religiosity scale where the average unprimed subject scores a 60, they would have to predict what number will come up after the subject is primed, i.e. 45+/-5 or 15+/-5 lower on the scale.

This then mimics how non-statistical sciences make predictions. For example, in physics or chemistry someone would not predict a correlation, but that a certain wave pattern will occur or a certain mass of a compound will result from the experiment. There is still error involved that requires statistical analysis, but this is generally subsumed into the error band of the prediction and the sample size rather than being the prediction itself. As such, a psychological experiment making predictions along these lines could be directly falsifiable in the same way that experiments in the harder, non-statistical sciences are falsifiable.

This requires theoretical robustness in two ways that are not necessarily present in directional significance testing. First, one needs a theory to assign numerical values to mental phenomena to develop the point scale that is being used. Second, one needs the theory to be robust enough to make precise point predictions. The consequence of this robustness would be the capacity to make riskier predictions that would put more money in the bank and if done consistently, make it much stranger if the corroborations turned out to be mere coincidence. The flip side is that falsifications would also happen, resulting in changes to auxiliary hypotheses and forcing problemshifts. This forcing of MSRPs to shift would result in some becoming progressive and thriving and others to degenerate and making it clearer which psychological programmes are more worth investing in and which ought to be abandoned.

References

1. Barsalou, Lawrence W. 2016. "Situated Conceptualization Offers a Theoretical Account of Social Priming." *Current Opinion in Psychology* 12:6–11.

2. Beauvoir, Simone de. 2010. *The Second Sex*. New York: Random House.
3. Bem, Daryle J. 2003. "Writing the Empirical Journal Article." In *The Compleat Academic: A Practical Guide for the Beginning Social Scientis*, edited by J.M. Darley, M.P. Zanna, and H.L. Roediger III, 2nd ed. Washington D.C.: American Psychological Association.
4. Borgeat, Fracois, Kieran O'Connor, Danielle Amado, and Marie-Eve St-Pierre-Delorme. 2013. "Psychotherapy Augmentation through Preconscious Priming." *Frontiers in Psychiatry* 4:1–8.
5. Ferguson, Christopher J., and Moritz Heene. 2012. "A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science's Aversion to the Null." *Perspectives on Psychological Science* 7 (6): 555–61.
6. Fletcher, Samuel C. 2021. "The Role of Replication in Psychological Science." *European Journal of Philosophy of Science* 11 (1): 23.
7. Friese, Malte, and Julius Frankenbach. 2020. "P-Hacking and Publication Bias Interact to Distort Meta-Analytic Effect Size Estimates." *Psychological Methods* 25:456–71.
8. Gervais, Will M. and Ara Norenzayan. 2012. "Analytic Thinking Promotes Religious Disbelief." *Science* 336:493–96.
9. Grant, Don C. and Edwin Harari. 2005. "Psychoanalysis, Science and the Seductive Theory of Karl Popper." *Australia and New Zealand Journal of Psychiatry* 39:446–52.
10. Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of P-Hacking in Science." *PLOS Biology* 13 (3): e1002106.
11. Ioannidis, John P.A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 19 (8): e1004085.
12. Lakatos, Imre. 1970. "History of Science and Its Rational Reconstructions." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1970:91–136.
13. ———. 1974. "The Role of Crucial Experiments in Science." *Studies in History and Philosophy of Science* 4 (4): 344–55.
14. ———. 1978. "Falsification and the Methodology of Scientific Research Programmes." In *Philosophical Papers*, edited by John Worrall and Gregory Currie, 1:8–52. Cambridge: Cambridge University Press.
15. Meehl, Paul. 1990. "Appraising and Amending Theories: The Strategyof Lakatosian Defense and Two Principles That Warrant It." *Psychological Inquiry* 1 (2): 108–41.
16. ———. 2006a. "Specific Etiology and Other Forms of Strong Influence: Some Quantitative Meanings." In *A Paul Meehl Reader: Essays on the Practice of Scientific Psychology*, edited by Neils G. Waller, Leslie J. Yonce, William M. Grove, David Faust, and Mark F. Lenzenweger, 175–91. London: Lawrence Erlbaum Associates.
17. ———. 2006b. "Theoretical Risks and Tabular Astericks: Sirl Karl, Sir Ronald, and the Slow Progress of Soft Psychology." In *A Paul Meehl Reader: Essays on the Practice of Scientific Psychology*, edited by Neils G. Waller, Leslie J. Yonce, William M. Grove, David Faust, and Mark F. Lenzenweger, 57–90. London: Lawrence Erlbaum Associates.
18. ———. 2006c. "Wanted-A Good Cookbook." In *A Paul Meehl Reader: Essays on the Practice of Scientific Psychology*, edited by Neils G. Waller, Leslie J. Yonce, William M. Grove, David Faust, and Mark F. Lenzenweger, 249–62. London: Lawrence Erlbaum Associates.
19. ———. 2006d. "Why Summaries of Research on Psychological Theories Are Often Uninterpretable." In *A Paul Meehl Reader: Essays on the Practice of Scientific Psychology*, edited by Neils G. Waller, Leslie J. Yonce, William M. Grove, David Faust, and Mark F. Lenzenweger, 445–86. London: Lawrence Erlbaum Associates.
20. Open Science Collaboration. 2012. "An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science." *Perspectives on Psychological Science* 7 (6): 657–60.
21. ———. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): 943.

22. Romero, Felipe. 2019. “Philosophy of Science and the Replicability Crisis.” *Philosophy Compass* 14 (11): e12633.
23. Sanchez, Clinton, Brian Sundermeier, Kenneth Gray, and Robert J. Calin-Jageman. 2017. “Direct Replication of Gervais & Norenzayan (2012): No Evidence That Analytic Thinking Decreases Religious Belief.” *PLoS One* 12 (2): : e0172636.
24. Smith, Roger. 1997. *The Norton History of the Human Sciences*. New York: W. W. Norton and Co.
25. Stevens, Jeffrey R. 2017. “Replicability and Reproducibility in Comparative Psychology.” *Frontiers in Psychology* 8:1–6.
26. Yang, Unbo, Ulrike Lueken, Jan Richter, Alfons Hamm, André Wittmann, Carsten Konrad, Andreas Ströhle, et al. 2020. “Effect of CBT on Biased Semantic Network in Panic Disorder: A Multicenter fMRI Study Using Semantic Priming.” *American Journal of Psychiatry* 177 (3): 254–64.