Beyond Statistical Ritual: Theory in Psychological Science

Travis Proulx

Cardiff University

Richard Morey

Cardiff University

Abstract

Over 40 years ago, Paul Meehl (1978) published a seminal critique of the state of theorizing in psychological science. According to Meehl, the quality of theories had diminished in the preceding decades, with statistical methods standing in for theoretical rigor. In this introduction to the special issue 'Theory in Psychological Science', we will apply Meehl's account to contemporary psychological science. We will go on to suggest that by the time of Meehl's writing, psychology found itself in the midst of a crisis that is typical of maturing sciences, where the theories that had been guiding research were gradually cast into doubt. Psychologists were faced with the same general choice when worldviews fail: face reality and pursue knowledge in the absence of certainty, or shift emphasis toward sources of synthetic certainty. We will suggest that psychologists have too often chosen the latter option, substituting synthetic certainties for theory-guided research, in much the same manner as Scholastic scholars centuries ago. Drawing from our contributors, we will go on to make recommendations for how psychological science may fully reengage with theory-based science.

With ‘Theory in Psychological Science’, we present a special issue that surveys the state of theorizing in the decades since Paul Meehl’s (1978) seminal critique. By Meehl’s reckoning, the quality of theorizing had diminished substantially in the preceding decades, with statistical methods standing in for theoretical rigor. In the estimation of our contributors, this trajectory has continued over the subsequent decades, with the predicted consequences and renewed calls for reform. Over the course of this introduction to the special issue, we will outline Meehl‘s account as it is relevant to contemporary psychological science. We will suggest that the trends identified by Meehl and our contributors began with a crisis of confidence faced by all developing sciences, where serious doubts emerge about the theories that direct scientific inquiry (Kuhn, 1962/2012; Lakatos, 1970).

As with any other science, psychological theories are worldviews, imparting the values that determine the *important* phenomena for scientific communities to explore and understand, along with methodological tools developed for this purpose (Kuhn, 1962/2012). Over time, novel or existing tools begin to turn up phenomena that cannot be accounted for by existing theory, or observations that directly contradict these worldviews. As theories run their course, the complexity of observed phenomena begins to outpace the capacity of tools to disentangle their causal pathways. By the time of Meehl’s writing, these trends were well underway, making scientific gains more difficult to determine and motivating efforts to identify progress elsewhere within statistical methods. Over the past decade, these trends may have accelerated as a purported replication crisis shook many subdisciplines of psychological science.

The cumulative effect of these worldview violations is *anxiety* among scientists. It is the same anxiety felt by anyone encountering an *existential* crisis, whereby the theories guiding our actions are gradually (or suddenly) undermined. Individuals, societies and sciences all face the same choice in response to this anxiety: acknowledge a shifting reality and pursue knowledge in the absence of certainty; or turn away from uncertain realities and toward sources of synthetic certainty. Centuries ago, Western academics chose the latter option. Combining religious dogma and dialectical logic, Scholasticism side-tracked Western scholarship for centuries following the upheavals of the Dark Ages, before the Enlightenment's re-emphasis on theory-guided empirical observation. Faced with its own crises and complexities, we will argue that significant strains of psychological science have recapitulated the deemphasizing of practical theory in favor of synthetic certainties. This retreat replaces messy realities with statistical determinations of what is "real". It replaces uncertainty-laden observations with rituals that can be registered with authorities for maximum certainty – or, rather, the *feeling* of certainty. It is in this context that we draw on our remarkable array of contributors for their assessments and recommendations on how psychological science may fully reengage with theory-based science, moving past its partial retreat into methodological dogmatism.

Meehl and the decline of psychological theorizing

At the outset, it is important to acknowledge the varieties of psychological sciences; it may be more useful to call psychology many sciences, not one. The perspective taken here will be applicable in different degrees to different research programmes. The psychophysicist may find Meehl's depiction of psychology unrecognisable; the cognitive psychologist may find it true to a point. To the extent that these issues are cultural, however, and to the extent that "psychologists" form a loose group in our scientific culture, we believe that psychologists of many stripes will find something familiar in it.

Moreover, it is not our intent to paint a totalisingly bleak picture of theory in psychological science, especially given that broad fields or sub-disciplines that can still be characterized as bastions of theory-directed research . For example, developmental psychology is historically grounded in broad theories of cognitive (Piaget, 1964), social (Vygotsky, 1967), moral (Kohlberg & Hersh, 1977) and identity (Erikson, 1994) development, with rich explanatory accounts of executive function (Zelazo & Müller, 2011) and theory of mind (Gopnik & Wellman, 1992) guiding research programs that range from language development (Sodian, Kristen-Antonow & Kloo, 2020; Gooch et al., 2016) to the causes of autism (Baron-Cohen, 2000; Tager-Flusberg, 2007). Within social psychology, foundational theories like cognitive dissonance continue to be applied to a range of phenomena (e.g., ideological adherence, Jost, Pelham, Sheldon & Sullivan, 2003; fluency and affect, Winkielman & Cacioppo, 2001), bridging into adjoining fields like social neuroscience that offer deeper explanations (e.g., Harmon-Jones, Amodio & Harmon-Jones, 2009). In applied

disciplines, neurobiological and pharmacological theories have revolutionized the diagnosis and treatment of depression and anxiety (Young & Craske, 2018). And we have no doubt that psychological societies reading this special issue can generate myriad examples of good theory practice which are often formally acknowledged – for example, the Wegner Theoretical Innovation Prize (Society for Personality and Social Psychology, n.d.). There are also numerous positive examples discussed by our contributors as exemplary guides for scientific action.

Nevertheless, the picture painted by Meehl over 40 years ago (Meehl, 1978), was a bleak one, and we believe this portrait depicts much of the modern psychological landscape. By his account, much of our science had entered a holding pattern, where the broad, generative perspectives of the previous century had given way to "theories" that were little more than tautological descriptions of effects (i.e., "pseudo-theories", Fiedler, 2004) . For example, a decades-long proliferation of experiments demonstrated that threats to the self evoke efforts to defend the self; these findings stood in support for ego-defense theories, from which we derive the hypothesis that when the self is threatened, people will defend it. While Gestalt Theory or Behaviorist principles could be applied to a multitude of human capacities, Meehl bemoans the "Rubber Band Theory" of intimacy (ones draws closer as another pulls away) and a multitude of other empty metaphorical descriptors applied only to the very effects they purported to 'explain'. In spite of — or because of — the growing complexities that made progress more difficult to assess, entire fields of psychological science began to seek statistical advances with a rapidly-expanding library of behavioral effects supported by null hypothesis significance tests (NHST). At best, this focus on NHST was orthogonal to scientific progress; at worst, it was accelerating the trends that appeared

to place much of psychology's best days behind it. If psychology had become the 'sick man' of the mature sciences, Meehl's proposed treatment began with a frank diagnosis of what made psychological science a unique challenge at the outset, and recommendations for statistical approaches that were better suited to a renewed focus on broad theoretical frameworks.

In the intervening decades, elements of Meehl's critique have been applied (Gigerenzer, 1998) and reapplied (Muthukrishna & Henrich, 2019) to a science that remains characterized by the identification of *effects* (van Rooij & Baggio, this issue). Often, the relative importance of these effects is determined by their purported 'effect size' (Broers, this issue) and their supposed robustness in the face of NHST conventions (Robinaugh et al., this issue), rather than from within the context of explanatory theoretical paradigms (Navarro, this issue). Nearly a decade ago, this approach manifested as broad failures of reproducibility (Szollosi & Donkin, this issue); as it turned out, a science that measures progress as the accumulation of significant effects incentivized scientists to report larger and more significant effects by the conventions of NHST (Simmons, Nelson & Simonsohn, 2011). The subsequent failure of other scientists to reproduce these effect sizes – or many of these effects – *might have* highlighted some broader and deeper implications for the manner in which we had been construing scientific progress. Ongoing objections to the effects-focused nature of our science *might have been* incorporated into the main thrust of the subsequent reform movement. Instead, these failures in scientific research methods have been diagnosed by a new generation of statistical methodologists who have judged it to be a 'crisis of replication' (Maxwell, Lau & Howard, 2015). Rather than turning to

developed theory to guide subsequent replication efforts, (Irvine, this issue), this 'crisis' has been addressed by ever-more formalized statistical conventions (Scheel et al., this issue), overseen and enforced by bureaucracies with no analogue in mature basic sciences. Our faith in statistical ritual has become, in the Popperian sense, an unfalsifiable theory (Gigerenzer, 2004).

Why not a theory of theory advancement? It seems to work well for other sciences, why not us – or rather, why not us to the same extent (Borsboom et al., this issue)? As noted by Meehl (and many others), psychology is a challenging endeavor at the outset. At the heart of any satisfying psychological science is our own experience of mental states, which remain unobservable. This left us with approaches that largely ignore them (e.g., Functionalism; James, 1890), infer them (e.g., Cognitive Dissonance Theory; Festinger, 1957) or deny them (e.g., Behaviorism; Skinner, 1963). Our mental traits are also unobservable, warping and shading into different contexts. While measuring unobservable constructs is challenge that even the 'hard' sciences must face (Kellen et al., this issue), the extent of the challenge is unique to our science. Assessing 'pride' across a variety of individuals and cultural contexts — or even just trying to understand whether such assessment is meaningful — is a challenge a particle physicist will never face. Nor will a particle physicist be faced with the causal complexity of predicting how traits and states with manifest within an array of social and motivational contexts.

This final challenge – the unfathomable *complexity* of psychological phenomena – means that by the 1970s, the common tools of our science had turned up the robust, pan-contextual processes that we have been specifying (and re-"discovering") for the past 50

years. Our early theories had guided us to the phenomena that human experience deemed *important* (cognitive conflict, approach/avoidance, associative learning…), but our initial feeling of rapid progress receded into an ocean of multi-causality (Eronen & Bringman, this issue); our scientific reach largely exceeded our methodological grasp. Even cognitive neuroscience, with its expanding array of measurement tools, has failed to accumulate knowledge at the pace of its own early days; mired in its own complexities, it is, at best, decades from the promise of reliably assessing mental states — let alone in real time, across contexts (Churchland, 1981).

So what is a science to do when growing complexity renders the easy gains harder to come by? If other sciences are any indication, the typical approach is to stick with theory and gut it out (Kuhn, 1964). In contrast, our own science has taken a different approach, one that may be unique amongst mature sciences: we often de-emphasize knowledge accumulation scaffolded by developed theory. By the time of Meehl's writing, this shift was observable in his evaluation of the fragmenting theoretical landscape. Over the past decade, the revaluation became explicit, with formalized norms eschewing research programs guided and judged in terms of theoretical paradigms. This "*Phenomena first*" psychological science (Gray & Wegner, 2013) is a collection of findings that constitute "*interesting* research" as judged by what is salient to the lay intuitions of the general public. To that end, findings should "*be surprising*", both novel and *counterintuitive* relative to common intuitions and established theory. Where once the most common basis for a rejected manuscript was "did not advance theory" (Greenwalt, 2012), we are now more likely to face the "The Pink Floyd Rejection" of findings that *too intuitively* build on existing perspectives, i.e., "just another brick on the wall of science" (Kail, 2012). In the current

climate, researchers are motivated to actively conceal the extent to which a given finding may follow predictably from an existing theory to avoid having their research judged as insufficiently "groundbreaking" (Coyne, 2015), resulting in the *theoretical amnesia that* characterizes much of our science (Borsboom, 2013).

For example, in part to avoid the admonition that "this is *just* cognitive dissonance theory" (Campbell, 2014), social psychologists develop and apply this paradigm across theoretical labels that often obscure the origins, be it "worldview verification theory" (Major, Kaiser, O'Brien, & McCoy, 2007), "system justification theory" (Jost, Banaji, & Nosek 2004), "uncertainty management model" (Van den Bos & Lind, 2002), "compensatory conviction" (McGregor, Zanna, Holmes & Spencer, 2001) or our own "meaning maintenance model" (Proulx & Inzlicht, 2012). (Though it might be reassuring that Eddie Harmon-Jones won the Society of Experimental Psychology Career Trajectory Award for developing the Action Based Model of Dissonance [2009], a direct articulation of cognitive dissonance theory). More generally, a theory in social psychology can often be characterized as a statement of purported motivation underlying an observed effect (e.g., achievement), followed by some combination of words associated with whether the motivation is sated or deprived (e.g., disruption) followed by the word "theory" or "model" (e.g., "Goal Disruption Theory: A Theoretical Startup," Siegel, 2013).

More broadly, research based on these norms should be primarily aimed at "*Grandmothers, not scientists,*" where common intuitions are perceived as a more stable grounding for research importance than "changeable paradigms" (Gray and Wegner, 2013). This final aspect of effects-based science pointedly turns traditionally Kuhnian notions of

scientific development on their head. Kuhnian accounts portray young sciences as polyphonies of 'theories' operating as descriptions of observed effects. Over time, these descriptions coalesce into broader paradigms that explain what had been observed, as well as guiding and focusing future work. From this perspective, the practical justification of theories lies not in whether we can determine which turns out to be *true* (Greenwald, 2012) but in their ability to stand as an authority that tells scientists what phenomena are important and how to assess them. Beyond methodological techniques, paradigmatic theories provide and impart *values*, which is what differentiates scientists from engineers. Broad agreement on important tenants of a theory encourages the development of increasingly powerful tools, bestowing a progress narrative that meets a deep psychological need among scientists: we may not be getting closer to t*he truth* —whatever that may be— but as we develop a paradigm, there is no question that we have accumulated more knowledge than we could in a pre-paradigm science.

Nevertheless, these generative worldviews all seem to meet the same fate. The expansive scope and focused tools of a scientific paradigm begins to turn up observations for which the core conceits cannot account. The paradigm buckles under the complexities of emerging phenomena, often at inconvenient times, even prior to the establishment of new paradigms toward which scientists can 'shift'. During these interscene periods, maturing sciences are faced with a complex reality imposing itself upon a failing authority. This "essential tension" between paradigmatic expectation and empirical experience is a very real and anxious state (Kuhn, 1977). How do scientists handle this existential crisis?

Existential Anxieties and Academic Absolutes

Over the course of the 20th century, theorists summarized this psychological process from a variety of academic domains as it plays out among individuals and societies. Within the existentialist literature, Albert Camus (1942/2004) described three pathways in the face of "undermined" worldviews. The first is the most common: retreat to a dogmatic faith that everything still makes sense. Alternatively, we can default to an alternative certainty: totalizing despair that all is lost. The third path comes with maturity: a middle way that accumulates knowledge and acquires values with the acceptance that neither is grounded in certainty; we're forever groping forward in fits and starts towards workable understandings of reality. Developmental psychologists outline a similar transition from adolescence to functioning adulthood. Over the course of a common adolescence, the complexities of life can overwhelm our draft theories of knowledge, values and personal identity. This crisis has been called *The Othello Effect* (Chandler, 1987), whereby "the well of certainty is poisoned" and the attendant anxieties motivate a rigid and solipsistic thinking typical of adolescence. With received wisdom in question, young people search for renewed absolutes either through dogmatism or nihilism (Gadamer, 1975/2002) – "secret sharers" that identify ultimate truths or deny truth altogether. However long these inclinations persist, they are not merely orthogonal to our psychological development; they undermine the processes that allows us to continue building better lay theories. Growing up means travelling the third path: facing up to the very complexities that make us anxious and gutting it out. We triangulate our perspective with others, building more powerful theories and accepting that there will always be experiences that we cannot fully predict, control or even understand.

We also gain the confidence to accept that there are worse things than being wrong (Boyes & Chandler, 1992).

In *Escape from Freedom*, Eric Fromm (1941/1994) argues that developing societies often respond in the same manner as developing humans in the face of failing worldviews: by turning towards a source of perceived certainty. Whether through "destructiveness" or "conformity", the nihilism/dogmatism arc for societies represents the same effort to quell uncertainty through commitment to an absolute worldview. At the time of Fromm's writing, global depression and geo-political realignments following the first World War created profound uncertainties. Totalitarian strongmen offered freedom from the complexities of life in exchange for complete control over the same. Centuries earlier, the fall of the Roman empire plunged Europe into a Dark Ages marked by rampages and plagues, and Western civilization found comfort within the increasingly rigid dogma of the Medieval church. Scholarship of this era fell within the tradition of *Scholasticism*, a mix of theology and classical philosophy that aimed towards coherent systems of dialectical logic (e.g., Thomas Aquinas' *Summa Theologica*, 1485/2012). Eschewing the direct examination of reality, it was believed that nature could be inferred from within the logic games and assumptions of abstract models. When brute empirical realties imposed themselves on these dogmas, paternal authorities shamed and threatened heretics away from their telescopes and back towards their sanctioned construals of so-called progress (e.g., *The Galileo Affair*, Finocchiaro, 2012). Scholasticism wasn't merely orthogonal to knowledge accumulation; for centuries, priestly abstractions and coercive punishments actively impeded the intuitive reality-testing that allow for productive frameworks.

Scientific Anxieties and Statistical Absolutes

By the time of Meehl's critique, the complexities of psychological phenomena were surpassing the sophistication of our paradigmatic methods. The 'essential tension' became growing anxiety, and psychological scientists were faced with the same choice as individuals and societies when worldviews begin to fail: face the complexities and gut out the anxieties, or retreat to the soothing certainties that impede scientific progress. As we make our way into the 2020s, we believe that much of our science has chosen the palliative path. Judged from above, psychology programs continue to put 'bums in seats', with psychology degrees remaining a prerequisite for many professional trajectories. Our graduate programs continue to attract talented students, many of whom go on to research careers. They fill our journals with effects that have been vetted by even more stringent methodological requirements, approved by newly-empowered science-registration bodies. Though if it is claimed that 'pre-registration is hard' (Nosek et al., 2019), it is a very simple game relative to good-faith research in the hardest science (Cesario, 2014, Guest & Martin, this issue).

A closer look reveals that our highest-impact journals are filled with studies that may report *a priori* power analyses and display pre-registration badges, while the complexities that bedeviled our first-wave theoreticians are seldom acknowledged or addressed. We do little to grapple with the vagaries of construct validity (Grahek, Schaller & Tackett, this issue) and even manipulation checks have become scarce (Fielder, McCaughey & Prager, this issue). Such matters are often taken for granted, as if the hardest science is now the easiest. In spite of these complexities, psychological scientists in the early to mid-20th century

channeled their uncertainty into the painstaking accumulation of effects, which coalesced into phenomena and eventually resulted in the kinds of broad explanatory paradigms that characterize a mature science. Within other mature sciences, researchers will agree on important hypotheses, pooling resources to build tools capable of discovering novel effects. While physicists build particle accelerators to grope for paradigmatically predicted particles, psychological scientists pool resources to determine whether a single, previously published effect 'replicates' (Gervais, this issue) – to see if it is *real* (or perhaps, to 'expose' that it isn't e.g., Wagenmakers, Wetzels, Boorsboom, van der Maas, & Kievit, 2012; Wagenmakers et al., 2016). While it could be argued that such replications are rhetorically useful – many researchers may even agree that the original work was questionable (Draper et al, 2015) – it is hard to see how such a 'whack-a-mole' approach can lead to a productive science. (Fortunately, more constructive approaches have arisen such as the Psychological Science Accelerator [e.g., Jones et al, 2021], modelled after high-energy physics.)

At the same time, we have turned to NHST rituals to circumvent replication reality-testing altogether in determining which previously-published effects (Francis, 2012), which research literatures and which researchers can be deemed uncontaminated by NHST manipulation (Simonsohn, Nelson & Simmons, 2014) – based upon NHST inferences resting on further NHST assumptions (Johnson, 2013; Kvarven, Strømland & Johannesson, 2020; Morey, 2013). Much like the Scholastics, the transgression of 'questionable research practices' is treated like an original sin that cannot not be redeemed of scientists' own accord; instead, the path to redemption is attained through extrinsic accountability to a higher authority. At present, we are a priestly science, testing effects-bounded hypotheses,

pre-registered by paternal authorities and investigated by inference from ever-more ephemeral statistical games.

Statistics is no substitute for theory

While Meehl despaired that statistical methods had already taken on an equivalence with scientific methods, there can be no question that statistical inferential methods have been a great boon to science over the last century. From one perspective, statistics injects healthy skepticism into the scientific process (Fisher, 1973): informally, what we see cannot be taken too seriously, because it might have arisen by "chance". Probability theory – developed for modelling dice and card games – has now been adopted to create models of human behavior. Statisticians have long been clear about the uneasy relationship between statistical models and science. Fisher, for instance, insisted that in scientific applications, "populations" are fictions, "products of the statistician's imagination…" (1955, p. 71); in a response, Pearson (1955) – one target of Fisher's comment – completely agreed, framing significance tests as a way of exercising that imagination. Box (1972) emphasized that "all models are wrong," and that one cannot arrive at a "correct" model (p. 792). Nevertheless, Box notes that statistical modelling is an iterative, creative process of development and critique. From the Bayesian perspective, Gelman and Shalizi (2013) describe the same iterative process; Morey, Romeijn and Rouder (2013) again emphasize the fictional nature of statistical models. More recently, Hennig (2020) draws attention to the gap between statistical models and reality.

Useful statistical reasoning, then, has a number of features: creative iteration, healthy self-skepticism and subordination to scientific concerns. Science unmoored from statistics is hobbled, but possible; in contrast, statistics unmoored from science is mere game playing (or, as Box put it, "mathematistry"). As early as 1919, Boring noted that "statistical ability, divorced from a scientific intimacy with the fundamental observations, leads nowhere" and suggested that skepticism of "significance" was good scientific practice. Given the centrality of self-doubt to statisticians, it is interesting that discussion of statistical behavior has been so prominent in the reaction to the replication crisis. Demanding pre-registration of analyses, on its face, devalues creative iteration by explicitly calling it suspect. Using pre-specified alphas for scientific decisions (Benjamin et al, 2018; Lakens et al, 2018) erases the chasm between the imaginary statistical universe and our own. Moreover, it reverses the relationship between statistics and science. In a kind of paradox, statistical *thinking* is about injecting uncertainty; but statistical *behavior* is supposed to increase credibility (Vazire, 2018). It is unclear how to bridge the gap.

Individual studies tell us very little about human behavior, and so it difficult to understand what the increased credibility would be *in*. What do we believe with respect to a potential demonstration of an effect? It cannot be that the effect is *real*; an effect is more than a few statistically-significant demonstrations. It is not a highly-precise measurement of an average effect size; any useful definition of a psychological effect is theoretical, a set of boundary conditions and manipulations that cause changes related in behavior or perhaps with a tentative mechanism. Although individual studies may form a web of support for theories, no single demonstration stands alone, and the way these studies relate to one

another is not primarily statistical in nature. Theories, not decisions based on arbitrary significance thresholds, provide the means of guiding future experimentation.

The replication movement has had to manage this paucity of theoretical understanding. Consider the practice of involving original researchers in the process of replication. This is done to ensure fidelity of a replication (e.g., Open Science Collaboration, 2015), obtain a prediction (e.g., Gronau, Ly, & Wagenmakers, 2017) or to define what a success would look like (e.g. Matzke et al, 2015). But none of these would be necessary if the effects were embedded in a developed theory. Having to go back to the original researchers for these purposes is a sign that the logic of the experimental design was questionable, and that the evidential import of a replication would not be clear. In these cases, the meaning of an experiment is heavily determined by the original authors, rather than by an author-independent relationship to a broader theoretical framework. The main function of replication attempts — particularly large-scale ones—has not been for refining specific methods or theoretical progress, but rather for rhetoric. Of what use besides rhetoric is an "average" replication rate across disparate subfields (e.g. "Overall, however, [a 39% replication rate in the 2015 Open Science Collaboration] points to widespread publication of work that does not stand up to scrutiny," Baker, 2015)? Effects-based thinking and paucity of theory has caused many researchers to be overconfident about what we know. Large pre-registered replication attempts have helped to undermine that overconfidence, which is (arguably) useful for those of us that agree that there are problems in the psychological sciences. We must be careful, however, not to confuse a practice that helps signal that there is something wrong with the solution to the problem.

The rise of meta-analytic forensic checks represents another inversion of statistical values (Morey, 2013). The scientific process across a literature is assumed to be governed by a simplistic statistical model. When these models don't fit, this is taken as evidence that the literature is problematic in some way: for example, that whatever 'effect' was under investigation is doubtful. Akin to Scholastic thinking, it has even been suggested that scientific behavior should be performed in such a way that these models can be assumed true: that is, science must conform to the models, rather than the models to science (Lakens, 2018; Frances, 2013b). Surely then we could trust scientific results. In order to buy this story, however, we must trust the models on which the meta-analysis is based, and we must believe any given grouping of studies submitted to a meta-analysis has some theoretical interpretability. The inference, after all – whatever it is – will be based on both. The irony is that forensic meta-analysis is often based on assumptions that are as strong as those made by the authors they are criticizing. Power-posing, for instance, was theoretically poorly defined. For their set of studies assessing power-posing, Simmons and Simonsohn (2017) depend on a list of studies constructed by the authors. Could they do otherwise? What determines whether an effect is a "power pose" effect is, in fact, nothing more than whether it was called a power pose effect.

Large-scale replications, forensic meta-analysis, pre-registration of statistical hypotheses, a focus on estimation of simplistic effects, reducing alpha for "discovery of new effects", Bayesian hypothesis testing to test whether effects are "real": all of these statistical reform activities reinforce the same ills they seek to solve. But if we agree that psychology has issues, and we reject these potential reforms as solutions, how can we move psychology forward? How do we return to testing reality with theory-driven hypotheses?

How does the preponderance of our science re-join the mature sciences, and reengage with theory?

Reengaging with theory

Returning to theory-based research will mean turning back to the complexities of our most difficult science, gutting out the anxieties, humbly acknowledging the limits of contemporary progress and broadly collaborating on the phenomena that our theories deem important. Within the critiques that comprise this special issue, there is much consensus on how we can broadly reengage with these principles.

Teach theory. As we grow up, we gradually apply internalized values to our own behaviors rather than awaiting judgment from an external authority. Similarly, mature sciences train mature scientists by ensuring that the succeeding generation has internalized the values of a generative theoretical paradigm. Rather than training another cohort of NHST-conversant research engineers, a new generation of psychological *scientists* will receive formal training in the application, assessment and construction of scientific theories (see Borsboom et al., this issue). If the next generation cannot grapple with the necessities of construct validity as graduate students (see Grahek, Schaller & Tackett, this issue), there is little hope that they can embody these values as subsequent researchers, reviewers or journal editors (see Fielder, McCaughey & Prager, this issue). Our science will continue to asymptote towards the same dead ends (see Eronen & Bringmann, this issue).

Use theory. The current incentive structure strongly rewards the discovery of 'novel' effects by individual labs, which must be verified as 'real' by multi-lab replication cohorts (see Irvine, this issue). In contrast, a generation conversant in theory-guided research will be less individualistic at the outset, adopting methodological reforms that aid in coordination of research efforts. They will begin with questions deemed important by developed theories (see van Rooij & Baggio, this issue), and will work together to explore the theoretically-relevant methodological space. Interpretations of results will be guided by how well these results fit together and new studies devised accordingly. To this end, pre-registration of research methods aids planning, communication and coordination. However, a developed theoretical paradigm prescribes these qualities at the inception of a research program (see Szollosi & Donkin, this issue). In a mature sciences, there is no need to pre-register hypotheses; published theories are pre-registered hypotheses (see Broers, this issue).

Reward Theory. As evaluators of scientific contributions, the next generation of psychological scientists will incentivize theory-grounded contributions over effects explained by 'theories'. For example, strong theory has been the ongoing hallmark of mathematical psychology as it is applied to cognitive processes (see Navarro, this issue). Mathematical and computational models (see Guest & Martin, this issue) enable a more transparent dialog through a common language, and hence easier communication about what matters in a theory. Although formal tools will look somewhat different in every subfield, looking to areas with thriving theory will provide sources of inspiration that weren't present at the time of Meehl's writing (see Robinaugh et al., this issue). Conversely, the demand for 'theoretical' explanations of every novel or unexpected finding creates an incentive for researchers to invent tautological theories and defend them, resulting in the

current landfill of disposable 'theory'. The next generation will allow authors to be humble and not to try to do everything at once; theoretical progress happens slowly across a field, in fits and starts of replication and extension. The units in which we divide our work should reflect that. More generally, genuine scientific progress is necessarily slow and is difficult to identify contemporaneously. Returns to extensive periods of exploration are often necessary to solidify theories worthy of renewed hypothesizing (see Scheel et al., this issue). We should not be discouraged by slow progress; looking for a way forward during periods of stagnation is common in science. We cannot fool ourselves into thinking progress is easier than it is by lowering the bar to novel simplistic effects (or replication of these effects).

Maintain a scepticism of statistics. Statistical analyses are not coextensive with 'results'; rather, statistics help reduce data to manageable forms and to, in some sense, separate interesting variance from uninteresting variance. Trivializing results by identifying them with 'what we can arrive at through the application of a statistical procedure' leads to science that is increasingly divorced from the reality these results are meant to represent. When applied poorly, statistical operationalizations become tautologically defined as the unobservable phenomena they are intended to assess (see Kellen et al., this issue). When applied well, application of statistics enforces humility (e.g., a wide uncertainty interval leading one to doubt the results) but statistical models should always be understood to be useful *metaphors*. More generally, broader coordination of research efforts will bring greater contact with existing bodies of accumulated scientific knowledge, and the broader realties they represent (see Lin et al., this issue). A greater focus on practical application of established phenomena is also essential is moving beyond narrow statistical rituals and into broader realities (see Berkman & Wilson, this issue).

Maintain a scepticism of reformist authorities. Whatever knowledge was derived from Scholastic scholarship, it was of no particular use in the prediction and control of observed reality. While the epistemic uselessness of dogma was apparent for centuries, scholastic practitioners maintained their authority by denigrating the value of empirical knowledge and redirecting attention to their models. When this failed, there were coercive threats to one's career and livelihood. In the current climate of psychological science, consider the source and nature of the reformist appeal. Is the purported authority a psychological scientist, or a statistical methods practitioner, drawing attention back towards the assumptions underlying their models (Devezer, Navarro, Vandekerckhove & Buzbas, 2020)? If they are primarily a psychological scientist, does their body of work exemplify the theory-based knowledge accumulation characterized by mature sciences, or do their strictures reinforce and incentivize their own effects-centered methods? If they advocate for pre-registration bodies, is it in the manner of persuasion, explaining why these authorities are uniquely necessary in our science? Or are these messages mainly expressed as coercive prerequisites for disseminating research and maintaining a career (see Gervais, this issue)?

Conclusion

At its inception, psychological scientists developed tests and measures and accumulated an impressive catalogue of phenomena. Over time, psychology matured, organizing itself into broad theories that offered explanations and directed research efforts. By and large, psychological scientists established constructs whose validity was determined with reference to explanatory paradigms. By the late 1970s, however, Meehl described a field where paradigmatic theories were less likely to direct research efforts. Instead, a

researcher was expected to determine an appropriate NHST given the effect they wished to demonstrate, design an experiment to meet the appropriate NHST assumptions and apply additional statistical techniques as auxiliary augmentations if the assumed conditions of the model were not met. In a Kuhnian sense, the turning away from paradigmatic science meant that all of the limitations of pre-paradigm science returned: 'theories' with no explanatory value, the re-hashing of previously identified effects and the absence of an agreed array of validated constructs. In an alternate trajectory, our society of scientists will re-incentivize the internalization of theory-based norms. Our journals will gradually be replenished with editors and reviewers who can determine whether these norms were present. Following from Meehl, we hope that this special issue is another nudge in the direction of theory-scaffolded empirical science. Scholasticism dominated Western academics for five centuries. We hope the current era of *statistica theologica* is somewhat briefer.

References

Aquinas, T. (2012). Summa theologica. Authentic Media Inc.

Baker, M. (2015). Over half of psychology studies fail reproducibility test. Nature News. https://doi.org/10.1038/nature.2015.18248

Baron-Cohen, S. (2000). Theory of mind and autism: A fifteen year review. Understanding other minds: Perspectives from developmental cognitive neuroscience, 2, 3-20.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., …

Johnson, V. E. (2018). Redefine statistical significance. Nature Human Behaviour, 2(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Boring, E. G. (1919). Mathematical vs. Scientific significance. Psychological Bulletin, 16(10), 335–338. https://doi.org/10.1037/h0074554

Borsboom, D. (2013). Theoretical amnesia. Open Science Collaboration. Retrieved from http://osc.centerforopenscience.org/2013/11/20/theoretical-amnesia/.

Boyes, M. C., & Chandler, M. (1992). Cognitive development, epistemic doubt, and identity formation in adolescence. Journal of youth and adolescence, 21(3), 277-304.

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Campbell, T. (2014) Never say isn't that just cognitive dissonance. [Web log post]. Retrieved from http://indecisionblog.com /2014/02/21/viewpoint-never-say-isnt-that-just-cognitive-dissonance.

Camus, A. (1942/2004). The myth of Sisyphus. In G. Marino (Ed.), *Basic writings of existentialism* (pp. 441–492). New York, NY: Random House.

Cesario, J. (2014). Priming, Replication, and the Hardest Science. *Perspectives on Psychological Science*, *9*(1), 40–48. https://doi.org/10.1177/1745691613513470

Coyne, J. (2015). Ten suggestions to the new associate editors of Psychological Science. [Web log post]. https://jcoynester.wordpress.com/2016/01/21/ten-suggestions-to-the-new-associate-editors-of-psychological-science/

Chandler, M. (1987). The Othello effect. Human Development, 30(3), 137-159.

Churchland, P., 1981. Eliminative Materialism and Propositional Attitudes, Journal of Philosophy, 78: 67–90.

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020). The case for formal methodology in scientific reform. BioRxiv, 2020.04.26.048306. https://doi.org/10.1101/2020.04.26.048306

Erikson, E. H. (1994). Identity and the life cycle. WW Norton & Company.

Festinger, L. (1957). A theory of cognitive dissonance (Vol. 2). Stanford university press.

Fiedler, K. (2004). Tools, Toys, Truisms, and Theories: Some Thoughts on the Creative Cycle of Theory Formation. *Personality and Social Psychology Review*, *8*(2), 123–131. https://doi.org/10.1371/journal.pone.0150205

Finocchiaro, M. A. (2012). Galileo Affair. The Blackwell Companion to Science and Christianity, 14.

Fisher, R. (1955). Statistical Methods and Scientific Induction. *Journal of the Royal Statistical Society: Series B (Methodological)*, *17*(1), 69–78. https://doi.org/10.1111/j.2517-6161.1955.tb00180.x

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*. Hafner Press.

Fromm, E. (1994). Escape from freedom. Macmillan.

Francis, G. (2012). The psychology of replication and replication in psychology. Perspectives on Psychological Science, 7(6), 585-594.

Francis, G. (2013b). We should focus on the biases that matter: A reply to commentaries. Journal of Mathematical Psychology, 57(5), 190-195.

Gadamer, H. G., Weinsheimer, J., & Marshall, D. G. (2004). EPZ truth and method. Bloomsbury Publishing USA.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology, 66(1), 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

Gigerenzer, G. (1998). Surrogates for Theories. *Theory & Psychology*, *8*(2), 195–204. https://doi.org/10.1177/0959354398082006

Gigerenzer, G. (2004). Mindless Statistics. *The Journal of Socio-Economics*, *33*, 587–606.

Gooch, D., Thompson, P., Nash, H. M., Snowling, M. J., & Hulme, C. (2016). The development of executive function and language skills in the early school years. Journal of Child Psychology and Psychiatry, 57(2), 180-187.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian t-Tests. The American Statistician, 74(2), 137–143. https://doi.org/10.1080/00031305.2018.1562983

Harmon-Jones, E., Amodio, D. M., & Harmon-Jones, C. (2009). Action-based model of dissonance: A review, integration, and expansion of conceptions of cognitive conflict. Advances in experimental social psychology, 41, 119-166.

Hennig, C. (2020). Frequentism-as-model. ArXiv:2007.05748 [Stat]. http://arxiv.org/abs/2007.05748

James, W. (1890/2007). The principles of psychology (Vol. 1). Cosimo, Inc.

Johnson, V. E. (2013). On biases in assessing replicability, statistical consistency and publication bias. *Journal of Mathematical Psychology*, *57*(5), 177–179. https://doi.org/10.1016/j.jmp.2013.04.003

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L. G., Bloxsom, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Saribay, S. A., Owsley, N. C., Jang, C., … Coles, N. A. (2021). To which world regions does the valence–dominance model of social

perception apply? Nature Human Behaviour, 5(1), 159–169. https://doi.org/10.1038/s41562-020-01007-2

Jost, J. T., Pelham, B. W., Sheldon, O., & Sullivan, B. N. (2003). Social inequality and the reduction of ideological dissonance on behalf of the system: Evidence of enhanced system justification among the disadvantaged. European Journal of Social Psychology, 33, 13-36.

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of framework justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. Political Psychology, 25, 881-920.

Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. Theory into practice, 16(2), 53-59.

Kuhn, T. S. (1977). The Essential Tension: Selected Studies in Scientific Tradition and Change. University Of Chicago Press.

Kuhn, T. S. (1962/2012). *The structure of scientific revolutions*. University of Chicago press.

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Lakatos, I. (1970, January). History of science and its rational reconstructions. In *PSA: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1970, pp. 91-136). D. Reidel Publishing.

Lakens, D. [@lakens] (2018, May 28). Exactly. If it looks bad, it's bad (under certain assumptions) and if it looks good, it might still be bad - just as with normal meta-analyses. Throw in a mixed bag of efects (e.g., JPSP) and there is always some true (trivial) effects in there. [Tweet]

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. Nature Human Behaviour, 2(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Major, B., Kaiser, C. R., O'Brien, L. T., & McCoy, S. K. (2007). Perceived discrimination as worldview threat or worldview confirmation: implications for self-esteem. Journal of Personality and Social Psychology, 92, 1068.

McGregor, I., Zanna, M. P., Holmes, J. G., & Spencer, S. J. (2001). Compensatory conviction in the face of personal uncertainty: going to extremes and being oneself. Journal of personality and social psychology, 80(3), 472.

Morey, R. D. (2013). The consistency test does not–and cannot–deliver what is advertised: A comment on Francis (2013). *Journal of Mathematical Psychology*, *57*(5), 180–183. https://doi.org/10.1016/j.jmp.2013.03.004

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. Proceedings of the National Academy of Sciences, 115(11), 2600-2606.

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., ... & Vazire, S. (2019). Preregistration is hard, and worthwhile. Trends in cognitive sciences, 23(10), 815-818.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? American Psychologist, 70(6), 487–498. https://doi.org/10.1037/a0039400

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A

preregistered adversarial collaboration. Journal of Experimental Psychology. General, 144(1), e1-15. https://doi.org/10.1037/xge0000038

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of consulting and clinical Psychology*, *46*(4), 806.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2013). The humble Bayesian: Model checking from a fully Bayesian perspective. British Journal of Mathematical and Statistical Psychology, 66(1), 68–75. https://doi.org/10.1111/j.2044-8317.2012.02067.x

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. Nature Human Behaviour. https://doi.org/10.1038/s41562-018-0522-1

Pearson, E. S. (1955). Statistical Concepts in Their Relation to Reality. Journal of the Royal Statistical Society: Series B (Methodological), 17(2), 204–207. https://doi.org/10.1111/j.2517-6161.1955.tb00194.x

Piaget, J. (1964). Cognitive development in children: Piaget. Journal of research in science teaching, 2(3), 176-186.Proulx, T., & Inzlicht, M. (2012). The five 'A's of meaning maintenance: Making sense of the theories of sense-making. Psychological Inquiry, 23, 317-335.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359-1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. Journal of experimental psychology: General, 143(2), 534.

Skinner, B. F. (1963). Behaviorism at fifty. Science, 140(3570), 951-958.

Sodian, B., Kristen-Antonow, S., & Kloo, D. (2020). How does children's Theory of Mind become explicit? A review of longitudinal findings. Child Development Perspectives, 14(3), 171-177.

Tager-Flusberg, H. (2007). Evaluating the theory-of-mind hypothesis of autism. *Current directions in psychological science*, *16*(6), 311-315.

Van den Bos, K., & Lind, E. A. (2002). Uncertainty management by means of fairness judgments.

Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress—Simine Vazire, 2018. Perspectives on Psychological Science. http://journals.sagepub.com/doi/10.1177/1745691617751884

Vygotsky, L. S. (1967). Play and its role in the mental development of the child. Soviet psychology, 5(3), 6-18.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. Perspectives on Psychological Science, 7, 627–633.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., … Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). Perspectives on Psychological Science, 11(6), 917–928. https://doi.org/10.1177/1745691616674458

Society for Persoanlity and Social Psychology. (n.d.). *Wegner Theoretical Innovation Prize*.

Retrieved March 1, 2021, from https://www.spsp.org/awards/annualawards/outstanding-contributions/wegner-theoretical-innovation-prize

Winkielman, P., & Cacioppo, J. T. (2001). Mind at ease puts a smile on the face: psychophysiological evidence that processing facilitation elicits positive affect. Journal of personality and social psychology, 81(6), 989.

Young, K. S., & Craske, M. G. (2018). The cognitive neuroscience of psychological treatment action in depression and anxiety. Current Behavioral Neuroscience Reports, 5(1), 13-25.

Zelazo, P. D., & Müller, U. (2011). *Executive function in typical and atypical development.* In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (p. 574–603). Wiley-Blackwell.