Running Head: 20 Questions

Moving Beyond 20 Questions: We (Still) Need Stronger Psychological Theory

Randall K. Jamieson

University of Manitoba

Penny M. Pexman

University of Calgary

Send correspondence to:

Dr. Randall K. Jamieson
Department of Psychology
University of Manitoba
Winnipeg, MB, Canada, R3T 2N2

Email: randy.jamieson@umanitoba.ca
Phone: 1-204-474-9360
Fax: 1-204-474-2599

Abstract

There has been growing awareness that many empirical demonstrations in psychology are difficult to reproduce: a problem called the replication crisis. To address the current *replication crisis,* Psychology has responded by re-examining its professional incentive systems, publication models, and research practices. Several reforms are now underway to correct for the problems, however skepticism is growing that psychology will escape the replication crisis by improvements in research practice alone. We address the *theory crisis*, the problems it poses for editors and reviewers, and we propose ways that reviewers and editors can contribute to addressing the replication crisis.

Keywords: Replication crisis, Theory crisis, Peer review, Editorial practice

Public Significance Statement

Many experimental reports in psychology fail to replicate. The situation has caused a great deal of disappointment and skepticism about psychological science. Much of the blame has been placed on how psychologists conduct experiments, the prevailing publication model, and how psychologists analyze their data. We join a growing debate that re-places the blame on a need for stronger mathematical approaches to theory building. We also point to ways that journal editors, scientific reviewers, and disciplinary incentives might be re-focused. Ultimately, we are optimistic that the replication crisis presents an opportunity for disciplinary self-improvement and growth.

Moving Beyond 20 Questions: We (Still) Need Stronger Psychological Theory

It is now accepted that many empirical demonstrations in psychology—even cornerstones of the discipline—are difficult to reproduce: a problem called *the replication crisis* (Open Science Collaboration, 2015).

Psychology has responded by re-examining its professional incentive system that has tended to punish careful research practice (Grimes, Bauch, & Ioannidis, 2017). It has also responded by interrogating and revising its research practices and policies: the formal frameworks we use to draw conclusions from data (e.g., Benjamin et al., 2018; Cohen, 1994, 1994; Nickerson, 2000; Wagenmakers, 2007), the questionable practices that have wormed their way into research (e.g., John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simohnson, 2011), and the misguided publication practices that, in retrospect, have promoted a deep file-drawer problem (e.g., Lindsay, 2017, 2018; Nosek et al., 2018).

Several reforms are now underway to correct for the problems. Researchers have taken up proposals to use alternative statistical methods (e.g., Cumming, 2014; Dienes, 2011; Kruschke & Liddell, 2018; Wagenmakers, 2007), Benjamin et al.'s (2018) proposal to draw more conservative conclusions from data, Simmons et al.'s (2011) recommendations to excise questionable research practices, and Nosek et al.'s (2018) recommendations to engage in Open Science.

Not to be outdone, journal editors have revised editorial policies and procedures. For example, the *Canadian Journal of Experimental Psychology* has introduced Open Science Badges to incentivize authors to provide complete reporting of data and methods (Pexman, 2017), added *Registered Reports* as a submission category (Jamieson, Bodner, Saint-Aubin, & Titone, 2019), and integrated manuscript submission with PsyArXiv—the discipline's repository for article preprints and Open Science practice.

From our perspective, the field has assumed a dignified stance on the replication crisis by acknowledging the problem, owning the blame, and developing remedies. Based on the response, we are optimistic that psychology will emerge stronger in the end (Lilienfeld, 2017;

Rodgers & Shrout, 2018). However, we are skeptical that psychology will escape the replication crisis by improvements in research practice alone.

The theory crisis

In Newell's 1973 commentary, "You Cannot Play 20 Questions with Nature and Win," he pointed out that, "Psychology, in its current style of operation, deals with phenomena," and that researchers too often frame their research questions in grand binary oppositions that are difficult to settle in a single experimental report. He went on to argue that continuing to conduct science by cataloguing behaviours and framing research questions in imprecise and oversimplified false dichotomies would leave the discipline disorganized and unwieldy in 30 years' time (i.e., in 2003).

In place of that strategy, he recommended that psychology re-focus on developing a strong, precise, and shared disciplinary theory that ties the empirical record together in a list of coherent psychological principles (see Surprenant & Neath, 2009). The shift, he argued, would transform psychology's contribution from a curio cabinet of loosely related behavioural facts into a productive and testable explanation of behaviour. Although Newell's (1973) warning is well known, the discipline has been slow to adopt his recommendations. We imagine that if Newell were alive to comment on the replication crisis, he might say something like, "I told you so."

Newell's (1973) thesis has been revived in relation to the replication crisis. For example, Oberauer and Lewandowsky (2019) argue that whereas discussions and solutions to the replication crisis have focused on changing research practice, the problem stems from an acceptance of rhetorical theories and the discipline's uncomfortable tolerance for weak and indirect logical links between theory and experiment: a problem they named *the theory crisis* (see also Muthukrishna & Henrich, 2018; Szollosi & Donkin, 2019). Based on their assessment, they recommended that psychology re-commit to developing and applying rigorous formal theories that promote strong and direct logical links with experiments.

In many ways, Oberauer and Lewandowsky's (2019) proposal echoes the current effort to reform research practices. However, their proposal extends the reach of that initiative to the reform of theory as well as data. As journal editors, we agree with Oberauer and Lewandowsky.

At the risk of spoiling the ending, we offer no magic bullet or wildly divergent perspective here; our thoughts echo what others have said on related issues (e.g., Gigerenzer, 2010; Oberauer & Lewandowsky, 2019; Szollosi et al., 2020; Yarkoni & Westfall, 2017). Rather, our contribution is in unpacking of the current crisis to support suggestions for next steps, especially from the perspective of editors and reviewers.

A tale of two traditions

Most manuscripts that we receive are rooted in *rhetorical theory*; what Oberauer and Lewandowsky (2019) call *discovery-oriented research*. In this tradition, researchers present a *rhetorical premise* and then demonstrate the feasibility of that rhetorical premise in a *narrative experiment*. For example, a researcher might propose that suggesting a concept (e.g., suspicion) will influence people's behaviour. Based on that premise, they might propose that introducing the smell of fish to a room will cause people to behave suspiciously—because "something smells fishy" (Lee & Schwarz, 2012). If people behave as anticipated, the data from the narrative experiment are interpreted as evidence for the motivating premise. However, if the premise does not *force* the experiment, the experiment cannot force the conclusion. Moreover, a narrative experiment is a creative, free invention and, therefore, can be devised to jury rig an outcome. Thus, although rhetorical research is important for experimental discovery, the use of inventive, open-ended narrative experiments invites false positives (Oberauer & Lewandowsky, 2019).

A minority of manuscripts that we receive are grounded in *formal theory*; what Oberauer and Lewandowsky (2019) call *theory-driven research*. In this tradition, a researcher presents a formal theory and then tests specific predictions computed from the theory. Indeed, as Yarkoni and Westfall (2017) have articulated, prediction can be distinguished from explanation and has several statistical and pragmatic advantages. For example, if one assumes, as is the case in the SIMPLE model of memory (Brown, Neath, & Chater, 2007), that episodic memories in multidimensional psychological space are located along a dimension representing temporal distance from the point of retrieval, the retrievability of an item is inversely proportional to its summed confusability with other items in memory, and the confusability of items along a temporal dimension is given by the ratio of the temporal distances of those items

at the time of recall, then people's recall should follow certain predicted patterns and, critically, not others. If people's behaviour is consistent with the predictions (measured by quantitative fit), the data are interpreted in support of the theory; more critically, if people's behaviour contradicts the predictions, the data are interpreted as evidence against the theory (e.g., Neath, VanWormer, Bireta, & Surprenant, 2014). Experiments conducted in this tradition are typically uninventive and introspectively unexciting. However, they have strong and direct links to their motivating theories, speak directly to the validity and precision of their motivating theories, and tend to replicate (Oberauer & Lewandowsky, 2019).

Finding our way out

At this point, one might expect us to argue that psychology should minimize its exposure to the replication crisis by abandoning rhetorical discovery-driven research in favour of formal theory-driven research (e.g., Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). However, both traditions are important to a vital and rigorous science of behaviour. Thus, the question is not how do we excise rhetorical research from our disciplinary strategy but, rather, how do we balance our commitments to rhetorical and formal research in a way that mitigates our exposure to the replication crisis but preserves the balance of experimental discovery and scientific rigour?

Theory-driven research

If psychology was a theory-driven science, the replication crisis would be more like a replication concern. In that world, theories would be formally specified, reviewers could objectively evaluate the theories, experiments would test *necessary predictions* of those theories, theory evaluation would proceed by comparing quantitative fits to experimental data, and the experimental record would form a coherent web of interrelated, mutually reinforcing incisive tests. In such a well-defined, transparent, and precise space, it is hard to imagine how we could arrive at a replication crisis. So, why don't we abandon risky discovery-oriented approaches and commit to a formal research program?

Theory-driven research is conservative by design and, consequently, short on party tricks. There are examples where well-specified theories have forced surprising insights and experimental discoveries. For example, the Rescorla and Wagner (1972) model of associative

learning was invented to explain cue-competition in blocking (Kamin, 1969), but went on to drive the discovery of new experimental phenomena thereafter (Miller, Barnet, & Grahame, 1995). However, theory-driven research's focus on precision and coherence rather than invention and discovery tends to constrain experiment breadth and, consequently, hobbles experimental discovery. Nevertheless, there are important and meaningful advantages of adopting a theory-driven approach to psychological science.

A cumulative experimental record. Psychology should commit to developing and testing shared formal theories. Mature branches of science are identifiable by their shared formal theories that record collective wisdom and organize a cumulative and interactive experimental record, typically in a system of equations. Physics, for example, has identifiable dominant theories that organize experimental work to derive and test increasingly specific and precise predictions. Because experimentalists proceed from common theories, the experimental data produced in different labs comingle in a shared, interactive, and cumulative experimental record that supports the kind of scientific progress to which psychology should aspire. In contrast, psychology's ballistic approach to investigating behaviour has produced an impressive but disorganized experimental record, where data produced in one lab too rarely intersect with or force consequences on other labs.

In the context of this special issue, and in relation to our thesis, psychology's loose and theoretically-promiscuous structure makes it difficult for reviewers and editors to judge the conclusions of experimental reports. Rather, editors and reviewers are left to assess each experimental report by its methodological and statistical rigour. The situation has severe and strange consequences. For example, we continue to publish "scientific evidence" that people possess latent supernatural abilities (*precognition*; Bem, 2011). If psychology is to escape the replication crisis, it is critical that we commit to requiring clear and formal scientific theories that not only imagine possibilities, but are held to the standard of providing a formal and precise account of the mechanisms that produce the behaviour in question.

Active correction. Shared theories act as a natural prophylactic to the replication crisis. A collective research program that proceeds from a shared dominant theory also organizes a quick feedback system of checks and balances to detect and correct false discoveries. When an

experimental discrepancy to collective wisdom arises, a correspondingly collective effort is launched to confirm or correct those mysteries (researchers in other labs are invested in the work arriving from other labs). Given a collective confirmation of the new fact, it becomes part of collective wisdom and the shared theory is revised or abandoned. If, however, the new fact is disconfirmed, it is removed from collective wisdom and business continues.

In contrast, psychology's theoretically-promiscuous state makes it difficult for reviewers and editors to recognize when a new fact contradicts collective wisdom. Consequently, false positives are difficult to spot in review and, once published, can remain undetected for long periods of time. Thus, although reforms to research practice are helping to solve the replication crisis, the replication crisis will persist without a commitment to strong theory.

In summary, no matter how transparent our data and research practices might become, our replication crisis will persist until psychology develops formal shared theories that codify collective wisdom (see also Szollosi et al., 2020). In our opinion, psychology should commit to developing a coordinated and theory-driven research program that promotes a coordinated and cumulative research record, one that naturally institutes checks and balances for the assessment and verification of new discoveries. Theory integration is essential; as Watkins (1984) first noted, psychologists tend to be averse to using one another's theories, likening it to "someone else's toothbrush – it is fine for that individual's use…" but not "…for the rest of us" (p. 86). Similarly, Mischel (2008) wrote about the incentives in the tenure and career systems that encourage researchers to distinguish themselves, a dictate that can run counter to developing a conservative and cumulative science. Overcoming the toothbrush problem and achieving integration will likely involve an emphasis on theory building in research training, cooperation rather than distinction in the theoretical field, interaction beyond our subdisciplines (e.g., Gigerenzer, 2010), a shift in career incentives, and encouragement from Editors and reviewers to develop and test shared formal theories.

Discovery-oriented research

Experimental discovery is critical for a healthy and vital science. However, discoveries come packaged with risks because they, by nature, challenge introspection, contradict established wisdom, and have the potential to do harm (e.g., the unbelievable and now-

discredited claim that changing your personality can prevent cancer; see Eysenck, 1991; Grossarth-Maticek & Eysenck, 1991; Pelosi, 2019).

Ongoing reforms to research and publication practices are already helping reviewers and editors to do a better job of detecting and correcting false positives in discovery-oriented research (e.g., Benjamin et al., 2018; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simohnson, 2011). However, the difficulties associated with deciding if an experimental discovery is merely an illusion will always remain difficult—especially in the absence of strong theories to evaluate the feasibility of exciting but mysterious experimental observations. So, how can we as editors and reviewers do our jobs better?

Converging evidence. Editors and reviewers should demand converging evidence for rhetorical conclusions. Lee and Schwarz (2012) speculated that priming a concept would influence people's behaviour. To confirm the premise, they reported evidence that introducing a "fishy" odour to a testing room can cause people to become suspicious. However, their rhetorical premise does not force their narrative experiment. Consequently, their evidence can serve as a sign of feasibility but cannot confirm their premise (in the same sense as Wason's, 1960, characterization of confirmation bias). Nevertheless, their premise is consequential and deserves investigation. So, how should researchers test a rhetorical premise that cannot be tested by counterfactual?[1]

As editors, we would be more comfortably convinced if researchers reported a series of narrative experiments to follow up on a demonstration of feasibility, all invented from the same rhetorical premise but differing in procedure and context. The larger the family of demonstrations, the harder it would become to doubt that any specific rhetorical demonstration reflects a methodological peculiarity rather than a confirmative demonstration. Ideally, each new test would test a variation on the original premise to rule out a possible alternate explanation. We would be even more comfortable if the grand conclusion was assessed by the aggregate weight of the different demonstrations against a demonstration-wise

[1] We appreciate that a failure to observe an experimental effect can serve as a sort of counterfactual. But, the failure to observe an experimental effect is an unconvincing counterfactual: the absence of evidence is not evidence of absence.

error rate (e.g., α = .05, divided by the number of demonstrations tried; see Benjamin et al., 2018).

In summary, rhetorical premises are difficult to test directly, but that should not discourage their investigation. Rather, in the absence of a direct test, reviewers and editors should compensate by demanding strong, conservatively assessed, and converging evidence over a coordinated series of indirect tests.

Deductive confirmation of abductive conclusions. Editors and reviewers should demand deductive confirmation of abductive conclusions. Conclusions in rhetorical investigations often rest on abductive inference; the process of moving from data to theory by seeking a coherent and reasonable explanation for observations in hand (see Kerr's, 1998, discussion of HARKing). Although abduction plays an important role in science (see Tukey's, 1977, discussion of exploratory data analysis), it comes packaged with risks if applied independently of deduction—reasoning forward from articulate and verifiable premises to precise and certain conclusions (i.e., see Tukey's, 1980, discussion on the coordination of exploratory and confirmatory data analysis).

As editors, we have observed worrisome conflations of abductive and deductive reasoning. When we request that authors re-conduct an experiment to convert their abductive inference (i.e., their theory consequent to the data) into a deductive conclusion (i.e., the data consequent to the theory), we have encountered resistance on the grounds that, "If the data support the abductive theory, then the corresponding deductive theory predicts the data."

As logicians know, abductive and deductive reasoning are different and license very different classes of conclusions. In relation to the problem stated above, the probability of deriving theory *T* conditional on data *D* (i.e., $p(T|D)$) is not necessarily equal to the probability of observing data *D* conditional on theory *T* (i.e., $p(D|T)$). Thus, converting an abductive inference into a deductive conclusion, even if only by direct replication, should be encouraged if not required.

In summary, changes in research practice are making it easier for reviewers and editors to do a better job of detecting false positives in discovery-oriented work. However, researchers reporting experimental discoveries from rhetorical theory can help reviewers to do a better job.

In the absence of a strong theory that demands specific experimental tests (not just confirmation of feasibility), authors can present converging evidence from a series of narrative demonstrations, none of which might be forced by but all of which are motivated by the same rhetorical premise. Secondly, reviewers and editors should encourage authors to follow through on discoveries by converting their post hoc abductive explanations into a priori deductive conclusions.

Publication problems

We have argued that adopting collective theory-driven research programs and refining assessments of discovery-driven research can help editors and reviewers to do a better job of regulating the replication crisis. However, editors are under additional pressures that have little to do with science but, nevertheless, contribute to the problem.

As was the case in Newell's (1973) day, editors are rewarded for publishing discoveries that attract downloads, citations, and press. They are also punished for publishing safe scientific reports that do not capture readers' imagination. In difference to Newell's day, those incentives are now monitored and published as *impact factors* and *altmetrics* (i.e., scientific value assessed by social media attention). If an editor does not attend to those numbers (e.g., devising editorial policies to inflate the numbers), they receive fewer submissions and those submissions are of lesser consequence. Although we have not calculated the lopsided ratio between journal pages dedicated to boring-but-safe versus exciting-but-risky research reports, editors know that devoting their journal pages to exciting discoveries rather than safe incremental science is "good business."

Fortunately, the replication crisis has reorganized the incentive system. The history at *Psychological Science* offers a case study of those changes.

Psychological Science is one of the discipline's most prestigious journals. Early on, its reputation grew based on publishing exciting discoveries that inspired enthusiasm and drew corresponding downloads, citations, and press. However, the flagship announcement of the replication crisis spoiled the party (Open Science Collaboration, 2015).

In response, the journal's editors launched a dignified and brave sea change in the journal's editorial policies. Eich (2014) required authors to report adequate statistical power,

effect sizes to supplement *p*-values, and meta-analyses. Eich also introduced Open Science Badges to incentivize and reward authors for reporting complete and detailed histories of their experiments, data, and analyses. Lindsay (2015) extended Eich's initiative by introducing *Preregistered Direct Replications* and encouraging preregistration of experiments when relevant. The effort (as well as Lindsay's energetic lecture circuit to promote a stronger publication model) is an exemplar of good editorial stewardship and, as a result, the journal has regained its prestige.

In our estimation, Lindsay was the right person for the job. However, the issues forced by the replication crisis set the stage for an improved publication model. Where editors were once rewarded for publishing risky discoveries and punished for publishing safe reports, the replication crisis has re-incentivized editors by rewarding them for publishing safe incremental science and, more critically, punishing them for publishing risky discoveries.

The positive changes initiated at *Psychological Science* are now increasingly reflected across the publication landscape (e.g., Jamieson et al., 2019; Pexman, 2017). Still, *impact factors* and *altmetrics* continue to influence and thus distort publication patterns behind the scenes—whether explicitly or implicitly. Nevertheless, making those changes will continue to come at a cost. In a final editorial, Lindsay (2019) acknowledged that introducing more rigorous publication practices and policies to the journal was correlated with a decline in submissions to *Psychological Science*—a pattern that suggests that changing our publication models for the better might require a follow up effort to re-align the distorted career and professional incentive system that underlies practices in scientific publication. So, how can we as reviewers, editors, and professional stewards stem those influences?

Curb your enthusiasm. Reviewers are generally more enthusiastic about manuscripts that report new and exciting discoveries than manuscripts that report conservative and uninventive tests of theory. The imbalance influences editorial decisions and, consequently, exposes the experimental record to replication problems. Reviewers and editors should pay attention to balancing their enthusiasm for manuscripts that report exciting discoveries versus manuscripts that report experimentally uninventive but valuable tests of theory.

Killing ourselves with kindness. Reviewers know the first-hand disappointment of receiving an action letter that requests additional experiments and, based on our experience, we surmise that sympathy leads reviewers to suggest without demanding additional experiments needed to verify conclusions in a revision. Although we encourage criticizing with kindness at the *Canadian Journal of Experimental Psychology*, a more important goal is to ensure that conclusions are sound and that published experiments replicate. Reviewers can help editors to address the replication crisis by being clear and direct about any additional experiments required to resolve uncertainties and concerns. They can also help by more liberally raising additional experiments that would help to clarify conclusions, leaving the editor with the responsibility of deciding if those suggested experiments are necessary before those conclusions become part of the lasting experimental record.

Single experiment reports. Anything can happen once and, consequently, single experiment reports make editors nervous. Nevertheless, single experiment reports are common in our journals. Reviewers can help editors to manage the risks associated with single experiment papers by being explicit and forthright about whether the single experiment provides certain conclusions without additional experimental corroboration. That service is particularly valuable to an editor when the data demonstrate but do not force the motivating hypothesis (see our recommendation on *Converging Evidence*).

Professional incentives. Science is a noble pursuit of drilling down to truth by logical scrutiny and interrogation. Every scientist has dedicated their life to that philosophy. However, the reality of *publish or perish* raises tensions and can incentivize scientists to publish data quickly to win jobs, tenure, promotions, and funding (see Pennycook & Thompson, 2018). Naturally, the speed-accuracy trade off leads to reporting of false positives and other kinds of errors.

Although it is unclear how reviewers and editors can correct the publish or perish culture and thereby alleviate some pressure in the replication crisis, the academic incentive system needs to be re-considered and re-designed to allow researchers the necessary time to ensure their data and conclusions stand up to scrutiny and replication before entering those data and conclusions in the empirical record. For that to happen, it is crucial that the forces

that shape our current incentive systems—the same ones that encourage a compromise of accuracy for speed—rebalance the game to favour the pursuit of rigour and truth rather than production and metrics. As the replication crisis has made clear, more data are not necessarily better data.

Although the responsibility of re-designing incentive systems falls to agencies, institutions, and professional organizations that encourage behaviours that put professional and scientific incentives into compromise, editors can contribute by rewarding theory development in the same way that they reward empirical discovery.

General discussion

The replication crisis is a defining moment of our time. The initial reaction was unproductive: exasperation, denial, disappointment, and a good deal of deflection, accusation, and scapegoating. However, as the weight of the situation set in, we shifted towards a productive program of identifying and seeking solutions to our problem. Early responses were focused on changes to research practice and a rebalancing of the incentive systems that underlie our publication model (Benjamin et al., 2018; Cohen, 1990, 1994; Nickerson, 2000; Simmons, Nelson, & Simohnson, 2011; Wagenmakers, 2007). More recently, we have begun to reconsider the nature of our theories, the role they play in causing the replication crisis, and how they can be improved in the pursuit of addressing our problems (Muthukrishna & Henrich, 2018; Oberauer & Lewandowsky, 2019; Szollosi & Donkin, 2019; Yarkoni, 2020). We have put those concerns in context by discussing how they interface with the review process.

In light of the efforts already underway, we are optimistic that psychology will emerge stronger from its efforts (Lilienfeld, 2017; Rodgers & Shrout, 2018). Part of our optimism is grounded in history. In the 1960s, psychology dealt with the embarrassment of *experimenter effects* (Rosenthal, 1966). In the 1970s, we faced the *file drawer effect* (Rosenthal, 1973). In the 1980s and 1990s, we revived the interrogation of the statistical frameworks that we use for scientific decision (Cohen, 1994; Nickerson, 2000). Taking a long view, the replication crisis is a new but consistent part in our ongoing history of self-evaluation and improvement. Perhaps we are too optimistic, but we predict a positive outcome from facing our most recent dilemma.

However, we agree with Oberauer and Lewandowsky's (2019) diagnosis. Reforming research practice will not be enough and we must also attend to the theory crisis. To move beyond the toothbrush problem and thereby begin the pursuit of a cooperative and interactive science guided by the goal of investigating and refining shared dominant theories, psychology needs to pause, reflect on what kinds of theories it is trying to develop, and proceed again from that shared frame. That will involve a collaborative and discipline-wide commitment that extends the discussion from how we conduct experiments and analyze data to how we express our theories and build a collaborative, mature, and cumulative experimental record in which different laboratories play with rather than beside one another (see also Lakens, 2017).

Efforts in the domain of experimental method prove that our proposal is not fantasy. As an example, take the study of memory. The discipline shares a number of articulate formal memory theories that explain data from a range of experimental tasks and manipulations. MINERVA 2 is one such theory (Hintzman, 1986, 1988).

Informally, MINERVA 2 is a theoretical framework that articulates representation, storage, and retrieval from memory. A central assumption of the model is that each experience is represented in memory by a unique trace. Another central assumption is that retrieval is probe-specific, similarity-driven, and parallel. Because retrieval is similarity-driven, a probe retrieves traces to which it is similar; this is how the model accomplishes recognition. Because a probe retrieves whole traces from memory, and because whole traces record all events of a trial, a probe also retrieves events it has co-occurred with in the past; this is how the model simulates cued-recall, prediction, and categorization. Formally, MINERVA 2 is a computational model of memory expressed in linear algebra in which memory traces are data structures (i.e., vectors), memory is a matrix, retrieval is a matrix operation, and decision follows from the information retrieved from the matrix operation. Critically, the model predicts behaviours from a range of phenomena including recognition (Hintzman, 1986; Jamieson, Hockley, & Mewhort, 2016), frequency judgement (Hintzman, 1988), cued recall (Hintzman, 1986), classification (Hintzman, 1986), function learning (Kwantes & Neal, 2006), judgement and decision (Dougherty et al, 1999; Thomas et al, 2008), speech normalization (Goldinger, 1998), confidence/accuracy inversions in recognition (Clark, 1997), language processing (Thiessen &

Pavlik, 2012), false remembering (Arndt & Hirshman, 1998; Johns, Jones, & Mewhort, 2012), memory dissociations in aging and amnesia (Benjamin, 2010; Curtis, 2019; Curtis & Jamieson, 2019; Jamieson, Holmes, & Mewhort, 2010), implicit learning (Jamieson & Mewhort, 2009a, 2010, 2011), speeded choice (Jamieson & Mewhort, 2009b), associative learning (Jamieson, Crump, & Hannah, 2012; Jamieson, Hannah, & Crump, 2010), evaluative conditioning (Aust, Haaf, & Stahl, 2019), embodied cognition (Versace, Vallet, Riou, Lesourd, Labeye, & Brunel, 2014), and semantic memory (Jamieson, Avery, Johns, & Jones, 2018; Kwantes, 2005). Importantly, for our thesis, model predictions are examined by several research groups over a range of questions and phenomena. In isolation, any particular application represents a minor advance. However, the collective effort amounts to something greater than the sum of its parts. Secondly, the theory's predictions are testable. Even if the theory has not been applied to a particular domain of investigation, it might already make predictions in that domain; or if it mispredicts data from the domain, it provides a framework to examine what additional assumptions or operations are necessary to extend into that domain. Of course, MINERVA 2 is not the only model framework that has been used to organize a collective research program (see also Brown et al., 2007; Murdock, 1982; Nosofsky, 1986; Shiffrin & Steyvers, 1997). But, it provides one active demonstration of formal, theory-driven research.

On the empirical front, the *ManyLabs* (https://www.manylabs.org) and *ManyBabies* (https://manybabies.github.io) projects demonstrate that (and show how) a collective replication effort can be organized. Those projects also provide positive evidence that a collective and coordinated research strategy can be effective. A similar initiative in the domain of theory seems like an excellent goal. Perhaps compiling a collection of models and contributing a collaborative database of tests on the theories to amalgamate and sort through the candidates would lead the field to some beginning degree of consensus on the structure of function of psychological processes.

Naturally, it is hard to predict what our shared theories might be and given the diverse nature of our discipline, it is unlikely we will arrive at a grand theory. For example, the Canadian Psychological Association has 32 content sections and it is unlikely or at least extremely difficult to imagine a theory that would predict behaviours examined across all of those sections.

However, the different branches of our discipline should make efforts to decide on the best candidates and get down to the work of investigating and testing those theories. Without a collective effort, it is unlikely we will arrive at a collective solution to the problem of psychology and behaviour.

In closing, we framed our discussion in relation to Newell's (1973) criticism of psychological research and theory. In his now famous paper, he noted that he was of two minds:

> "Half of me is half distressed and half confused…. Maybe we should cooperate in working on larger experimental wholes than we now do. My positive suggestions in the prior section were proposals of how to do that. They all have in common forcing enough detail and scope to tighten the inferential web that ties our experimental studies together. This is what I think would be good for the field. [Or] Maybe we should all simply continue playing our collective game of 20 questions. Maybe all is well, as my other half assures me, and when we arrive in 1992 (the retirement date I pick might as well be my own) we will have homed in to the essential structure of the mind."

Like Newell (1973), we are also of two minds. Our one half is hopeful, but our other is skeptical. We are convinced that efforts to develop a shared theoretical framework for investigation will benefit psychology and contribute to solving the replication crisis. However, Newell's call from 47 years ago has not exerted the strong influence it deserves and so maybe we are overly optimistic. Nevertheless, we remain hopeful that refreshing people's memory about Newell's argument will reinvigorate his cause and that in 30 years' time we will have finally homed in to the essential structure of the mind.

References

Aust, F., Haaf, J. M., & Stahl, C. (2019). A memory-based judgment account of expectancy-liking dissociations in evaluative conditioning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*, 417-439.

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language*, *39*, 371–391.

Benjamin, A. S. (2010). Representational explanations of "process" dissociations in recognition: The DRYAD theory of aging and memory judgments. *Psychological Review*, *117*, 1055–1079.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10.

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407-425.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*, 539-576.

Clark, S. E. (1997). A familiarity-based account of confidence-accuracy inversions in recognition memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *25*, 232–238.

Curtis, E. T. (2019). Interactive processes in an instance model of memory: A computational analysis of Jacoby's (1983) dissociation between perception and recognition. *Canadian Journal of Experimental Psychology*, *73*, 288–294.

Curtis, E. T., & Jamieson, R. K. (2019). Computational and empirical simulations of selective memory impairments: Converging evidence for a single-system account of memory dissociations. *Quarterly Journal of Experimental Psychology*, *74*, 798-817.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*, 7-29.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA- DM: A memory processes model for judgments of likelihood. Psychological Review, 106, 180–209.

Eysenck, H. J. (1991). Personality, stress, and disease: An interactionist perspective. *Psychological Inquiry*, 2, 221–232.

Franklin, D. R., & Mewhort, D. J. K. (2015). Memory as a hologram: An analysis of learning and recall. *Canadian Journal of Experimental Psychology*, *69*, 115-135.

Gigerenzer, G. (2010). Personal reflections on theory and Psychology. *Theory & Psychology, 20,* 733-743.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.

Grimes, D. R., Bauch, C. T., Ioannidis, J. P. A. (2018). Modelling science trustworthiness under publish or perish pressure. *Royal Society Open Science*, *5*, 171511.

Grossarth-Maticek, R., & Eysenck H. J. (1991). Creative novation behaviour therapy as a pro-phylactic treatment for cancer and coronary heart disease: Part I – Description of treatment. *Behaviour Research and Therapy*, *29*, 1-16.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528–551.

Jamieson, R. K., Bodner, G. E., Saint-Aubin, J., & Titone, D. (2019) Editorial: Registered reports. *Canadian Journal of Experimental Psychology*, *73*, 3-4.

Jamieson, R. K., Crump, M. J. C., & Hannah, S. D. (2012). An instance theory of associative learning. *Learning & Behavior*, *40*, 61-82.

Jamieson, R. K., Hannah, S. D., & Crump, M. J. C. (2010). A memory-based account of retrospective revaluation. *Canadian Journal of Experimental Psychology*, *64*, 153–164.

Jamieson, R. K., Holmes, S., & Mewhort, D. J. K. (2010). Global similarity predicts dissociation of classification and recognition: Evidence questioning the implicit/explicit learning

distinction in amnesia. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*, 1529–1535.

Jamieson, R. K., & Mewhort, D. J. K. (2009a). Applying an exemplar model to the artificial-grammar task: Inferring grammaticality from similarity. *Quarterly Journal of Experimental Psychology*, *62*, 550–575.

Jamieson, R. K., & Mewhort, D. J. K. (2009b). Applying an exemplar model to the serial reaction time task: Anticipating from experience. *Quarterly Journal of Experimental Psychology*, *62*, 1757–1783.

Jamieson, R. K., & Mewhort, D. J. K. (2010). Applying an exemplar model to the artificial-grammar task: String-completion and performance for individual items. *Quarterly Journal of Experimental Psychology*, *63*, 1014–1039.

Jamieson, R. K., & Mewhort, D. J. K. (2011). Grammaticality is inferred from global similarity: A reply to Kinder (2010). *Quarterly Journal of Experimental Psychology*, *64*, 209–216.

Jamieson, R. K., Mewhort, D. J. K., & Hockley, W. E. (2016). A computational account of the production effect: Still playing twenty questions with nature. *Canadian Journal of Experimental Psychology*, *70*, 154-164.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology, 65*, 486-518.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*, 524-532.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment* (pp. 279–296). New York: Appleton-Century-Crofts.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196-217.

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*, 178-206.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin and Review*, *12*, 703–710.

Kwantes, P. J., & Mewhort, D. J. K. (1999). Modeling lexical decision and word naming as a retrieval process. *Canadian Journal of Experimental Psychology*, *53*, 306–315.

Kwantes, P. J., & Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*, 1019–1030.

Lakens, D. (2017). Towards a more collaborative science with StudySwap. http://daniellakens.blogspot.com/2017/08/

Lee, S. W., & Schwarz, N. (2012). Bidirectionality, mediation, and moderation of metaphorical effects: The embodiment of social suspicion and fishy smells. *Journal of Personality and Social Psychology*, *103*, 737-749.

Lindsay, D. S. (2015). Editorial: Replication in psychological science. *Psychological Science*, *26*. 1827-1832.

Lindsay, D. S. (2018). Editorial: Preregistered direct replications in psychological science. *Psychological Science*, *28*, 1191-1192.

Lindsay, D. S. (2019). Swan song editorial. *Psychological Science, 30,* 1669-1673.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla–Wagner model. *Psychological Bulletin*, *117*, 363– 386.

Mischel, W. (2008). The toothbrush problem. *APS Observer, 21 (11).*

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*, 609-626,

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behavior*, *3*, 221–229.

Neath, I., VanWormer, L. A., Bireta, T. J., & Surprenant, A. M. (2014). From Brown-Peterson to continual distractor via operation span: A SIMPLE account of complex span. *Canadian Journal of Experimental Psychology, 68*, 204–211.

Nosofsky, R. M. (1986). Attention, similarity, and the identification– categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). New York, NY: Academic Press.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, *5*, 241.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . .Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422-1425.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), Article aac4716. doi:10.1126/science.aac4716

Pelosi, A. J. (2019). Personality and fatal diseases: Revisiting a scientific scandal. *Journal of Health Psychology*, *24*, 421-439.

Pennycook, G., & Thompson, V. A. (2018). An analysis of the Canadian cognitive psychology job market (2006–2016). *Canadian Journal of Experimental Psychology, 72*, 71–80.

Pexman, P. M. (2017). CJEP will offer open science badges. *Canadian Journal of Experimental Psychology*, *71*, 1.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). New York: Appleton- Century-Crofts.

Rodgers, J. L., & Shrout, P. E. (2018). Psychology's replication crisis as scientific opportunity: A precis for policymakers. *Behavioral and Brain Sciences*, *5*, 131-141.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. NY: Appleton-Century-Crofts.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638-641.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145-166.

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*, 1359–1366.

Surprenant, A. M., & Neath, I. (2009). *Principles of memory*. New York, NY: Psychology Press.

Szollosi, A., & Donkin, C. (2019). Neglected sources of flexibility in psychological theories: From replicability to good explanations. *Computational Brain & Behavior*, *2*(3-4), 190-192.

Szollosi, A., Kellen, D., Navarro, D., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Science*, *24*, 94-95.

Thiessen, E. D., & Pavlik, P. I. (2013). iMinerva: A mathematical model of distributional statistical learning. *Cognitive Science*, *37*, 310-343.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*, 155–185.

Tukey, J. W. (1977). *Exploratory data analysis*. Pearson.

Tukey, J. W. (1980). We need both exploratory and confirmatory research. *The American Statistician*, *34*, 23-25.

Versace, R., Vallet, G. T., Riou, B. Lesourd, M., Labeye, E., & Brunel, L. (2014) Act-In: An integrated view of memory mechanisms. *Journal of Cognitive Psychology*, *26*, 280-306.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779-804.

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*, 632–638.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task, *Quarterly Journal of Experimental Psychology*, *12*, 129-140,

Watkins, M. (1984). Models as toothbrushes. *Behavioral and Brain Sciences, 7,* 86.

Yarkoni, T. (2020). The generalizability crisis. https://doi.org/10.31234/osf.io/jqw35

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in Psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12,* 1100-1122.