This manuscript has been invited for revision at Psychological Review after an initial round of peer review.

If God Handed Us the Ground-Truth Theory of Memory, How Would We Recognize It?

Vencislav Popov

Department of Psychology, University of Zurich, Switzerland

Correspondence should be addressed to: vencislav.popov@gmail.com. The manuscript is available as a preprint on Psyarxiv: https://psyarxiv.com/ay5cm/. The ideas in this manuscript were presented at the 2023 Annual Summer Interdisciplinary Conference and the 2023 Psychonomic Society's One World Seminar Series. I am grateful to Klaus Oberauer, Gidon Frischkorn-Bartsch, Vladimir Sloutsky, Mark Steyvers, Robert Goldstone, Richard Shiffrin, Michael Kalish, Jonn Dunn, and the students of my class "Explanations of Brain and Behavior" at the University of Zürich for stimulating discussions about this topic and valuable feedback.

Abstract

What makes a scientific theory convincing? We have many formal models of human memory, but no agreement about which is the right one. If anything, we agree that they are all wrong. Despite important theoretical milestones, models of human memory are fragmented, most often explaining single paradigms or memory types; they are incompatible with each other; they lack precision in their predictions, and due to our current parameter fitting approach, they are often little more than proof of concept exercises. I fear that our current way of developing and evaluating theoretical models of memory and cognition is getting us nowhere, because we have no clear goal. To address this, I propose the Principle of Completeness: we will be convinced by a theory of memory only when it is able to make precise point predictions for individual people's behavior in any new memory task, manipulation, or paradigm we could construct, without refitting parameters to do so or only by estimating its parameters for each individual on an independent standardized battery of tests. Such a theory would not only be able to accurately describe lab-based empirical effects but would also be practically useful. I highlight how some of our current model development and evaluation practices might be holding us back and outline some important empirical steps necessary to evaluate theories by this standard. Whether readers agree with my proposal or not, it is crucial that we discuss where we are going and how to get there.

Keywords

cognitive modeling, generalization, individual differences, mathematical psychology, memory, philosophy of science, psychological theory

Contents

1. Where are we going?

On September 23, 1846, Johann Gottfried Galle, an assistant astronomer at the Berlin Observatory, woke up in a world quite different from the one he went to bed in. Later that day he would become the first person to knowingly observe Neptune – an observation made ever more remarkable by how it came to be.

For, you see, Neptune was discovered not with telescopes, but with math. That morning Galle received a letter from a French mathematician named Urbain Le Verrier who had been struggling with a baffling puzzle. Uranus, considered at the time to be the outermost boundary of the solar system, stubbornly refused to follow the orbit dictated by Newton's law of universal gravitation. Although some doubts about Newtonian mechanics might have crept up here and there, it was widely regarded as correct – its other successes were too difficult to deny. Another solution was needed, and Le Verrier had deduced, on a piece of paper, not only that a previously unknown planet must exist past Uranus, but exactly where and how big it needed to be to explain the discrepancies. While we can only speculate about how Galle must have felt when he observed Neptune a mere 1° away from the position Le Verrier had predicted, the public recognition that something momentous had occurred was unmistakable. In a letter to Le Verrier, Sir George Airy, the Astronomer Royal, wrote the following soon after: "With these things, the produce not only of a mathematical but also of a philosophical mind we have nothing which we can put in competition"[1]. To discover a celestial object not by accident, as had always been the case before, but through mathematical derivation, was an unprecedented human achievement (Smart, 1946).

It was also a startling testament to the magnetic power that Newton's mechanics held over people's minds. Despite Uranus's obstinate behavior, to doubt the law of universal gravitation would have been tantamount to sacrilege. Learned men and women, having long discarded the papal crown, now consecrated bread and wine at the empiricist's altar and could not deny that Newton's elegant equations accepted every sacrifice that was asked of them. The law of universal gravitation (Box 1) had explained with pioneering mathematical precision such diverse puzzles as to why the planets move the way that they do, why some places on Earth experience only one high tide per day instead of two, and why apples fall from trees instead of being hurled towards the sky. Aristotle was out, Newton was in, and all was one in the heavens and the Earth.

Explanatory breadth, mathematical precision, and parameter-free derivations ruled the day. A vivid example of this comes from Newton's "Moon Test" – a thought experiment, the accuracy of which is nowadays disputed (Westfall, 1973), but one that nevertheless illustrates an important point. In the canonical version of the story, Newton set out to calculate how far an object such as an apple on a tree will fall in one second, based on how far the moon travels across the sky in one minute. A silly endeavor as far as his contemporaries must have been concerned. Yet, the number derived from his theory came down within 1% of the measured value (Kollerstrom, 1991) – a stunning demonstration given that moons and apples seem to be such different beasts.

[1] Airy's role in the history of science is inglorious – not only did he largely ignore a similar plea for the search of a new planet a year earlier by John Couch Adams, a young Cambridge graduate who is nowadays recognized as a co-discoverer of Neptune, but he initially spurned Adams in public, preferring to extol Le Verrier's derivations instead (Smart, 1946).

Even though Newton's law was eventually displaced by Einstein's theory of relativity, the fact is that, at least in principle, it could have been correct – in its time it accounted for all possible observations with such precision that it remained undisputed for centuries and is even still used today as a useful approximation. What is more – it became and remains to this day the gold standard for the *form* of physical theories. The transformation of astronomy into the first modern science might have begun with Tycho Brache, who introduced the very idea of accurate measurements, but it was Newton who completed the project by providing a thorough mathematical account of these observations (Taton et al., 2003; Wootton, 2015).

What about psychological theories and particularly theories of human memory? Do we have or are we even approaching the explanatory breadth and precision of a discarded 350-year-old physical theory? While many readers might disagree with many of the claims in this paper, I have no doubt that the answer will be a resounding "No".

So where are we at? When I peruse the pages of our most prestigious journals, I find many mathematical theories of memory (to name but a few prominent examples, although many others exist - Anderson et al., 1998; Brown et al., 2007; Hintzman, 1984; Murdock, 1993; Oberauer & Lewandowsky, 2011; Oberauer & Lin, 2017; Osth & Dennis, 2015; Polyn et al., 2009; Popov & Reder, 2020; Reder et al., 2000; Shiffrin & Steyvers, 1997). These models are important milestones in our field, work that I greatly admire. At the same time, even though on the surface many of them have the same form as successful theories in the physical sciences, their weaknesses almost perfectly mirror the strengths of the Newtonian achievement. Lest I annoy you too much from the outset, let me rush to point out that my own modeling work (Popov & Reder, 2020) is not immune and suffers from the same problems: models of human memory are fragmented, most often explaining single paradigms or memory types; they are incompatible with each other; they lack precision in their predictions, and due to our current parameter fitting approach, they are often little more than proof of concept exercises. This may sound too harsh but give me the benefit of the doubt until I expand on these issues in the next sections. We will all surely agree that we have not solved memory quite yet – otherwise we would

Box 1. Newton's law of universal gravitation and Kepler's three laws of planetary motion

Newton's law: "Every particle attracts every other particle in the universe with a force that is proportional to the product of their masses and inversely proportional to the square of the distance between their centers." In its modern notation:

$$F = G\frac{m_1 m_2}{r^2}$$

where F is the gravitational force, m_1 and m_2 are the masses of the objects, r is the distance between their centers, and G is the gravitational constant.

Kepler's three laws of planetary motion:

1. Each planet's orbit is an ellipse with the Sun at one of the two foci
2. A line connecting a planet to the sun sweeps out equal areas in equal times
3. The square of the orbital period is proportional to the cube of the orbit's semimajor axis

not have such a cornucopia of models. My goal is not to criticize specific models, but to look at the road ahead and see where it might take us.

How fundamental are these problems? Like Alen Newell exactly 50 years ago (1973), I am a man of two minds. Half of me thinks that our field is in fine shape – every year we discover new phenomena or boundary conditions of existing phenomena. Every year we expand our understanding of the theoretical landscape; important new theories and extensions of existing theories are published. This is excellent work by people I respect. This half of me thinks that our current theoretical limitations are temporary growing pains, and that if we keep going the way we are, we will nevertheless eventually reach an all-encompassing, parameter-free, precise final theory of human memory on par with theories in the physical sciences.

My other more pessimistic and skeptical half plagues my waking moments with concerns that our current approach is hopelessly misguided; what's even more – that we don't really know where we are headed, let alone how to get there. This half of me shares Newell's (1973) troubling thoughts from half a century ago:

> *Suppose that in the next thirty years we continued as we are now going.... Will psychology then have come of age? Will it provide the kind of encompassing of its subject matter – the behavior of man – that we all posit as a characteristic of a mature science? And if so, how will the transformation be accomplished ...? (pp. 5-6)*

On the golden jubilee of Newell's "You can't play 20 questions with nature and win" paper, I can't help but ask myself: has anything changed? To translate these thoughts to the specific issue of human memory – would we ever achieve a final theory of human memory? What would such a theory look like and how would we recognize it? What is the end goal of our field? How would we know that we are "done" and it's time to throw in the towel and focus on other problems? The rest of this article is penned by my skeptical half, and it attempts to answer these questions, to diagnose the theoretical problems further, and to propose a way forward.

2. A brief, but important tangent – what is this paper (not) about?

The scientific process consists of an iterative cycle of theory development, measurement and theory evaluation (Box, 1976; Guest & Martin, 2021; Kellen, Davis-Stober, et al., 2021; Oberauer & Lewandowsky, 2019). Much work has gone into elucidating how to build better theories in psychology (Borsboom et al., 2021; Guest & Martin, 2021; Meehl, 1967; Navarro, 2021; Newell, 1973; Robinaugh et al., 2021; van Rooij & Baggio, 2021). Equally, significant efforts have been put into defining proper methods of measurement (Brady et al., 2021; Kellen, Davis-Stober, et al., 2021; Rotello et al., 2015; Starns et al., 2019; Van Fraassen, 2008), and into evaluating the validity of our theories (Busemeyer & Wang, 2000; Kellen, Winiger, et al., 2021; Platt, 1964; Roberts & Pashler, 2000; Robinson & Steyvers, 2023; Shiffrin et al., 2008). Most of this work focuses on the technical aspects of the theoretical development-evaluation cycle.

However, there is also a long tradition of asking what we want from theories of cognition and behavior. For example, Newell (1973) advocated for the development of large-scale models that encompass the entire cognitive process; Meehl (1967) emphasized the importance of theories making precise point-predictions of observations; van Rooij and Baggio (2021) stressed that theories should aim to explain cognitive capacities, and use empirical effects only to adjudicate between different

theories; Busemeyer and Wang (2000) encouraged the field to depend less on model fitting, and more on generalizing predictions to novel tasks. The current paper builds on this prior work by bringing such criteria into concert to address the thought experiment posed in the title – if God handed us the ground-truth theory of memory, how would we recognize it?

My goal is not to provide explicit step-by-step instructions for how we should develop theories of memory or theories of cognition more generally. For those interested in such instructions, there are excellent tutorials available (see Guest & Martin, 2021; Robinaugh et al., 2021; van Rooij & Baggio, 2021; van Rooij & Blokpoel, 2020). For now, suffice to say that I think the question of *how* to build the kinds of theories I urge for in this article is extremely important. However, to keep a tight focus on my goal, and to respect space constraints, any practical discussion of how to build such theories necessarily must be treated only in sketch form and must come after a discussion of what we want such theories to achieve. It is this question of what we want our theories to look like and what they should be able to do that I see as the gap in current meta-scientific discussions I am addressing.

This paper is also not about the link between theory and measurement, although no comprehensive discussion of how to do better science can go without it. I take it for granted that proper measurement depends on a set of theoretical assumptions linking a latent unobservable variable with observable behavior (Dunn & Anderson, 2022; Fraassen, 2012; Kellen, Davis-Stober, et al., 2021). Moreover, it is evident that theory development and evaluation can go off-track when this connection is ignored (Brady et al., 2021; Rotello et al., 2015) or misspecified (Kellen, Winiger, et al., 2021; Robinson et al., 2022; Starns et al., 2019). I also take it for granted that tests of theories always involve a test of the conjunction between core and auxiliary assumptions. Failures to pass a test do not necessarily falsify the core assumptions of a theory (Lakatos, 1976; Meehl, 1990), unless auxiliary assumptions are extensively documented and independently tested (Robinson et al., 2022).

Neither is this text intended as a comment on the replication crisis. That said, I agree with several researchers who have argued that one of the deeply rooted causes of the replication crisis is the weak link between theories, hypotheses and measurement; that formalizing theories is a necessary step towards solving the crisis (Fried, 2020; Guest & Martin, 2021; Oberauer & Lewandowsky, 2019).

I am also not advocating abandoning mechanistic theories in search of black box models that can do prediction at the cost of explanation (like, for example, Yarkoni & Westfall, 2017). I find no intellectual satisfaction in theories that do not provide a mechanistic understanding of memory or any other cognitive process. Being able to formally account for the relationship between experimental conditions and observable behavior do not automatically sum up to a theory (Devezer, 2023). Even though I argue later that precise quantitative predictions are a necessary property of some good explanatory theories, they are not sufficient – good explanatory models explicate the mechanisms involved in generating the behavior of interest (Craver, 2006). Prediction is not the end goal, but rather a useful by-product and diagnostic of a good theory. The two goals of prediction and explanation are not in conflict and should be integrated (Hofman et al., 2021).

Finally, the best measure to quantify how well a theory fits data, be it old or new, is a question best left to statisticians (Busemeyer & Wang, 2000; Shiffrin et al., 2008). However, we clearly should not blindly surrender scientific judgement to statistical inference (Gigerenzer, 1998; Navarro, 2019; Singmann et al., 2023).

So, if this article is not about any of the issues above, then what is it about?[2] My most fundamental objective is to present a thought experiment – if we were, by whatever means, to suddenly stumble upon the true model of human memory, how would we know it? By what standards will we determine this? As far as I am aware this question has not been asked or answered before. As I noted above, there exist a few recommendations about what makes a theory a good theory. I draw on such recommendations to push this idea to the extreme – to ask not what makes a theory good, but what would make it "complete"[3]. My thesis is that the field of memory research seems theoretically stuck (Hintzman, 2011; Watkins, 1990), because we have neither asked ourselves, nor agreed upon what we want to achieve. I argue that if we follow the principles outlined here, we would be able to recognize the true model of memory (or any other cognitive capacity), or a good enough approximation of it. We should take this standard, which I term later "The principle of completeness", as an end goal in our theory development practices.

An obvious question arises: if this standard doesn't hand-hold a scientist through the process of properly building theories, measuring latent variables, or statistically evaluating goodness of fit, then how is it, if at all, useful? The proposed standard can guide theory development on a more abstract level, not unlike how the phrase "Form follows function" guides aesthetic principles in modernist architecture. It doesn't tell you how to build a structure so that it doesn't collapse, but rather guides your choice of engineering and architectural solutions to fulfil a certain vision. It sets an agenda for memory modelers, should they choose to agree with it. And should they choose not to, the question posed in my title would still require an answer.

3. What makes a theory convincing?

Before I go any further, I would like to delve deeper into what made the universal law of gravitation so successful because "discussions in the abstract rarely provide much guidance to the would-be theoretician" (Navarro, 2021). Wittgenstein wrote that "at the end of reasons comes persuasion" (Wittgenstein, 1972), and even though some argue he meant it as a rebuke to the idea that there are any objective reasons to prefer one worldview over another (Wootton, 2015), it is this very notion that interests me: what makes a theory persuasive? What makes a theory so convincing, so widely accepted, that conflicting observations cause people not to doubt the theory but to look for new planets? Several features of Newton's achievement stand out.

[2] In my desire to streamline the exposition of this essay, I have purposefully chosen to not dwell on the issues described above, though needless to say I find all of them valuable – many of the cited articles have shaped my thinking on the philosophy of psychological science. My choice to forgo an extended discussion of measurement and theory-building issues might inevitably annoy some readers who know and work within those areas. As it stands, attempting to present an integrated view of the cycle between theory development, formalization, measurement, and theory evaluation, would require a far longer and different type of article; it would also unfortunately obscure my main goal, which is to raise the question of what it would take for a theory of memory to convince us that our job is done.

[3] A careful reader might at this point object that the very idea of a complete or true theory is fundamentally debatable, due to the problem of underdetermination (Duhem, 1954; Maatman, 2021); although I am a realist, none of my arguments rest on the assumption of realism – just substitute "complete/true" with "providing the best possible description of reality" – even if that best is merely one of several possible theories.

3.1. Explanatory breadth

One of the beautiful aspects of the law of universal gravitation is that despite the large variety of phenomena it applies to, it is very easy to designate them all with a single phrase – it applies to the motion of anything with mass. Without the understanding provided by Newton's law, the elliptical orbits of planets and the vertical fall line of objects towards earth seem as unrelated as humans and fungi, whose last common ancestor lived about a billion years ago. Similarly, even though the idea that matter is composed of indivisible atoms was proposed by Democritus c. 430 BC, it was not until the early 20th century for the theory to finally be fully developed and accepted: by that point it explained why chemical elements only form multiple compounds in whole number ratios (Dalton, 1808); the mechanical and thermal properties of gases (Bernoulli, 1738); and Brownian motion (Einstein, 1905). Explanatory breadth convinces because it is *a priori* unlikely and because of the large degree of compression involved – many observations are reduced to a few principles. At the other extreme, a theory that merely enumerates all possible observations has no value. Inversely, the larger the degree of compression, the more convincing the theory.

3.2. High precision and inference tickets

This may be an obvious and trivial point for a physicist, but for psychologists who mostly deal with ordinal scales and predictions such as "Condition A is greater than Condition B", it is worth discussing. The law of universal gravitation allows us to derive the exact numerical value of planetary orbit durations or the time necessary for an object to fall to the ground. Just like explanatory breadth, the precision of these derivations is convincing because of the very low prior probability that observations would agree with them. More importantly, these highly precise derivations extend not only to existing observations, but to new situations – successful theories serve as *inference tickets* (Ryle, 2000). They can tell us what will happen in a new situation, which we have not created or observed yet (Borsboom, 2013). The discovery of Neptune is a perfect example (Smart, 1946).

In contrast, psychological theories rarely predict anything better than directional differences between conditions or some monotonic relationship with a continuous variable. I am far from the first to point this out – Paul Meehl discussed the value of "risky point predictions" over 40 years ago when he wrote that "a theory that makes precise predictions and correctly picks out narrow intervals or point values out of the range of experimental possibilities is a pretty strong theory" (Meehl, 1978). Meehl specifically criticized theories in "soft" psychology, but despite the formalism of its "harder" counterparts, I struggled to think of any contemporary cognitive model that achieves this either.

3.3. No free parameters

"But our best theories do fit to numerical values of observations all the time" might respond some memory modelers, and indeed we often take pride that this is what gives an advantage of formalized over verbal theories. Until recently I would have gladly joined in this rebuke. But if we are honest with ourselves, we would agree that our enterprise of fitting models to data is very different from the way that Newton's law fits to data. We can fit numerical values of existing data by fiddling with parameters but give us a task or a manipulation not tried before and watch our predictive power plumet. We don't have inference tickets and a big reason for that is that we estimate free parameters anew for each task, paradigm, or population.

The law of universal gravitation has one parameter that is constant and estimated once – the gravitational constant. The accurate measurement of this constant by Henry Cavendish did not occur until 1798, 111 years after the publication of the *Principia,* but knowing its value was not even necessary to explain the most important observations about planetary motion, summarized by Kepler's three laws (see Box 1). Newton had demonstrated that Kepler's laws were an inevitable consequence of the law of universal gravitation.

The other variables are the masses of the relevant objects and the distance between them. The values of these variables are not estimated to fit the data – instead they are measured independently. In essence the model has no degrees of freedom. In contrast, in most psychological models, we fit and estimate many free parameters whenever we apply the model to a new task, population, or manipulation.

To drive the importance of this point further, imagine how much less impressive Newton's achievement would be if the masses of planets and the distances between them were not determined independently, but instead were estimated using state-of-the-modern-art model fitting techniques so that the equation would predict the correct orbits. This would already dramatically reduce the impressiveness of the theory but imagine further that you estimated different values for the mass of the earth to describe its orbit around the sun and to describe the orbit of the moon. It sounds silly, but this is often what we (including me) do with psychological mathematical models. Don't get me wrong – parameter fitting has an important role during theory building, but in the way it's practiced today it will never lead to a convincing final theory.

3.4. It explains individual behavior

Even though this point is clear in my mind, I struggled to formulate it because in the physical example it is almost moot – of course that what the astronomer wants to understand is why each particular planet has its particular orbital speed and orbital characteristics. This is not to say that general properties of planetary orbits are of no interest – indeed, Kepler's three laws, whose derivation from the law of universal gravitation provided its first strongest support, describe empirical generalization applicable to all planetary orbits. Yet as empirical generalizations they are insufficient to predict where a new planet should be located. They were also not simple aggregation of planetary orbits. Newton's law made it possible to explain general principles *through* the inclusion of individual characteristics (mass and distance).

Unfortunately, we often treat individual differences in psychology in the opposite way – not as the thing to be explained, but as the thing to be explained away – a nuisance on our way to construct a truly general theory of behavior. If people differ, well, that's what summary statistics are for.

3.5. Practical relevance

At a public lecture in 2013, the biologist Richard Dawkins was asked by an audience member "Why should we trust science?" His answer, crass though it may be, has become an internet legend, a meme by the father of memes – "If you base medicine on science, you cure people; if you base the design of planes on science, they fly; if you base the design of rockets on science, they reach the moon. It works, bitches" (ThinkWeekOxford, 2013).

I am usually the first to arrive to the frontlines to defend basic science from the onslaught of public attacks of the form "What is the point of this? What can we do with it?" Nevertheless, it is

difficult to deny the broader acceptance of scientific theories that can be used to answer practical questions. Newton's mechanics has enormous practical relevance – along with other scientific advances of the time, it arguably unleashed the beginnings of the industrial revolution (Gráda, 2016). It allowed people to build bridges and all kinds of new structures in a principled and secure manner, rather than through trial and error. When your new experimental bridge stands for decades, it is difficult to doubt the theory that made it possible. Practical relevance is not only for the public's benefit – I use the term in a broader sense – a successful theory often allows other scientists to make new discoveries that would have been inconceivable otherwise. Such theories allow the accumulation of knowledge because they sit on solid ground. Practical relevance may not be a necessary property of all successful theories, but if it is within reach, we should strive for it.

4. How does this apply to a theory of memory?

Accept for a moment that the features above are necessary and sufficient for any convincing theory that deals with quantitative phenomena and let me turn to how we can apply them in our search for a final theory of memory. My thesis is simple:

> **Principle of Completeness:** We will be convinced that a model is the final theory of memory when it is able to make *precise* point predictions for *individual* people's behavior in *any new* memory task, manipulation, or paradigm we could construct, *without refitting* parameters to do so or only by estimating its parameters for each individual on an *independent* standardized battery of tests.

I am not claiming that our current models don't possess any of these qualities; there are excellent examples of models doing one or more of these things right. Many of these recommendations have also been proposed before. My claim is that taken as a whole, the *Principle of Completeness* provides an evaluative standard by which we will recognize that the project for a comprehensive theory of memory is complete.

I suspect that to some this principle might sound trivial, while others might find it difficult to swallow. To spike the interest of the former, and to allay the fears of the latter, let us break down the mouthful Principle of Completeness into more digestible parts.

4.1. One theory to rule them all

The original title of this section was "Distinction between memory types and tasks don't matter", but of course they do – you only have to consider the many effects that appear one way in recall tasks, but the opposite in recognition tasks (e.g. the word-frequency paradox, Gregg, 1976; Popov & Reder, in press). Obviously, different memory tasks involve different, if overlapping, sets of mechanisms and processes. Because of this, our field has adopted a "divide and conquer" approach – with some notable exceptions (e.g. Anderson et al., 1998), researchers have dedicated their careers to understanding the functioning of a single (sometimes two) memory task, type of paradigm.

This is a useful strategy when a field is young – tasks like item recognition provide enough of a challenge on their own, without piling on top the complexities of recall; understanding serial recall has proved to be difficult enough without worrying about continuous reproduction; long-term and short-term memory have proved to have a much more puzzling relationship than anyone might have

expected; the question of how episodic context-bound memories are transformed to become the building blocks of our abstract semantic knowledge remains disputed to this day. And putting all these pieces together is clearly an even harder task.

The divide-and-conquer strategy has led to the proliferation of many different mathematical and computational models of human memory, each guarding a different turfdom within what is hopefully to become one day the United States of Memory. In conference hallways you can hear colleagues joking that "Models are like toothbrushes – everybody wants their own and no one wants to use anyone else's" – a sentiment that perfectly encapsulates how dissatisfied people are with this approach. After all, as Frederick Backman writes in Anxious People, "humor is the soul's last line of defense". When will the toothbrush joke stop being funny?

Clearly, a complete theory of memory will need to incorporate differences between memory types and tasks within its mechanistic assumptions. For example, when it comes to episodic memory, people are able to perform item recognition, associative recognition, cued recall, and free recall even when they do not know during encoding what kind of memory test awaits them (Cox et al., 2018). These tasks differ in the way people access the contents of their memories – so much is clear from the wildly different rates of successful remembering across the tasks. Yet the processes during encoding and the nature of the resulting memory representations are the same. There is an exception to this – when people know how their memory would be tested, they might use different encoding strategies and end up having different types or differently strong representation, depending on the test to follow. In both cases, a complete theory of memory would specify the encoding process, representational format and retrieval processes in any task used to measure memory performance and it would be able to predict the differences in memory performance across all these tasks with the same model[4].

When it comes to the distinction between working memory and long-term memory, regardless of whether you believe they are distinct systems or not, the only relevant difference for measurement is the filled or unfilled delay between the study and the test. If you believe that they are separate systems (e.g. Norris, 2017), you still need to recognize that at any short study-test delay it is likely that both systems contribute to memory performance. Perhaps many already accept this, but despite this we still model serial recall with pure models of working memory. If working memory and long-term are indeed separate systems that jointly contribute to performance, modeling serial recall data with a pure working memory model would surely derail model development.

If you instead believe that there is no structural distinction between long-term and working memory (e.g. Brown et al., 2007), the problem to be solved is easier, but you need to be able to account for data that is usually interpreted in support of the dual-system view. There is no escaping the fact that in most real-world or lab situations there is no clear-cut moment at which people retrieve a memory from a long-term instead of a short-term store.

What about explicit versus implicit memory tests? Just like the distinction between recognition versus recall, explicit and implicit memory tests might depend on different retrieval processes, but performance is still driven by the same representations (Reder et al., 2009), and it needs to be accounted for by the same theory. Furthermore, semantic and episodic memories interact in many complex ways

[4] For example, a preliminary attempt to do something like that, using the same set of parameters to model Cox et al's (2018) four-task data with the Source of Activation Confusion model appears in Popov & Reder (2020). But it was a minor point of the paper, it was limited to this one case, it ignored many complexities of the recall tasks, and plugging our model is not the goal of this article

(Greenberg & Verfaellie, 2010), in both directions, and clearly no theory of memory will be complete without explicating their relationship. Procedural memory, perhaps most distant to the other types I discussed, follows many of the same empirical regularities as declarative memory (Anderson et al., 1999), a fact betraying common mechanisms or, mechanisms shaped by common evolutionary, environmental, or developmental pressures.

I have presented these distinctions in the typical binary manner because most of the theoretical work that has even attempted any integration has typically focused on integrating these binary oppositions. But I do not want to give the impression that my concern is limited to binary oppositions of any flavor – Newell decried this approach far more elegantly that I could. We need to move beyond and attempt the grander integration – a theory that given any set of encoding and retrieval conditions, any set of materials and any population, predicts performance with precision.

Maybe I am preaching to the choir – after all, there are multiple models that attempt to unify two or more tasks or memory types (e.g. Anderson et al., 1998; Brown et al., 2007; Gillund & Shiffrin, 1984; Lin & Oberauer, 2022; Nelson & Shiffrin, 2013; Popov & Reder, 2020). This is work I admire. But it is done in relative isolation, and it is still applied only to idiosyncratically selected problems that each modeler considers as the most pressing (Oberauer et al., 2018a), leading to barely overlapping magisteria[5] across the theories.

Thus, we need to make grander integration a priority. Newell's own suggestion to progress the field forward was the development of cognitive architectures – large-scale models of the entire cognitive process. An obvious rebuke to my arguments above is that cognitive architectures such as ACT-R (Anderson et al., 2004) have since become common place and are widely used. But then why do we have other models? The most generous (and probably correct) answer would be that ACT-R's account of memory is not complete; that there are reasons to think other theoretical frameworks might be more fruitful to pursue; and there is an argument to be made about fully exploring the theoretical space before committing to any model (Shiffrin, 2010). Still, we need to make sure those other approaches are also headed in the right direction, and integration is part of it.

Objection: "Fully understanding individual tasks is more practical"

Conscious of the fact that these claims might be controversial, let me address an obvious objection. My less pessimistic half often tries to convince me that there is no reason to worry. That our current strategy has led to important theoretical insights (it has) and that if we continue as we are, once models of individual tasks are completed it would be trivial to combine them into a unified model of memory. The same voice tells me that breaking down the problem of human memory in the more manageable chunks of free recall and recognition, of working and long-term memory, of semantic and episodic memory, of implicit and explicit memory, of procedural and declarative memory, is a much more practical and productive approach. It leads to clearly defined problems such as whether global matching is a good description of how recognition occurs or whether the resource underling short-term storage is continuous or discrete (if it exists at all).

[5] Adapted from Stephen Jay Gould's phrase "non-overlapping magisteria" used to describe his position that science and religion are not in conflict because they represent different areas of inquiry, each with its own net of ideas and phenomena (Gould, 1997)

My skeptical half, however, worries that we are losing sight of the forest for the trees. It agrees that divide-and-conquer is a more practical starting approach when a young field needs to discover the possible building blocks of the theoretical landscape; yet it thinks that we run the risk of ending up with incompatible theories and that we will miss out on important theoretical insights. The first issue is already omnipresent – our best models make incompatible assumptions about how information is represented in memory. While it might be tempting, as Oberauer and colleagues put in the context of working memory models (2018a), to accept that different theories simply explain different sets of findings and leave it at that, this is unsatisfactory when the theories in questions hold mutually incompatible beliefs about how memory functions.

I fear that the gaps between theories might only widen as they become better at explaining individual tasks and that these theories might end up in local minima. That we'll remain with plenty of good models and no great ones. Ptolemey's, Copernicus's, and Brache's models of the solar system were all able to account for puzzling observations of planetary orbits, such as their apparent retrograde motion, despite vastly different assumptions – they were all observationally equivalent and all essentially wrong (Bunn, 2012). We might be "overfitting" individual tasks and by now surely, we have enough theoretical maturity to focus on the larger task at hand – integrating theoretical assumptions in a common framework.

4.2. The final theory is mathematical in form

I don't want to belabor this point since memory is one of the psychological fields richest with mathematical formalism, and much has been already written on the benefits of formal models over verbal theories (Hintzman, 1991; Lewandowsky & Farrell, 2011; Navarro, 2021; Oberauer & Lewandowsky, 2019; Smaldino, 2017) – although there are certainly dissenting voices (e.g. Maatman, 2021), and some might even go so far as to call this physics envy (Clarke & Primo, 2012). Nevertheless, there is plenty of verbal theorizing in memory, so a summary might be useful. Mathematical models allow precise derivations and predictions; they make the theory independent of its creator; they illuminate hidden assumptions; they allow us to develop better intuition of how complicated mechanisms interact; they reduce human bias and cognitive limitations; they tighten the link between theory and hypotheses. Perhaps most importantly – mathematics is simply the language, the best most precise language known to man, we use to describe patterns and regularities. As long as we believe there are any regularities in the functioning of human memory, mathematical or computational models are our best hope for understanding them.

4.3. Parameter fitting is not enough

Despite the existing formalism in the field, I am dissatisfied with our current approach to evaluating models. The standard practice is that of model fitting – to find a set of parameters for which the model reproduces the main patterns in the data as closely as possible. I won't deny that descriptive accuracy is an important step in the model development process (Shiffrin et al., 2008). A successful fit shows that the model could work in principle, but that is not enough to convince us that the model is complete – it could be merely one of many models that could be tinkered with to reproduce the data. Even though this is the finished product that often appears alone in our journals, every modeler knows that a much more important step occurs during the model development process, a step that is usually omitted in the

report – all the variations of the model for which no set of parameters allowed them to fit the data (Shiffrin & Nobel, 1997).

I see multiple issues with parameter fitting as it is often practiced today. One such issue was highlighted long ago by Roberts and Pashler (2000). As they argued, we don't learn anything from a good fit unless we know that many other plausible results would have been inconsistent with the model. A model that can fit any patterns provides no explanatory value and inversely the fewer patterns that are allowed by the model, the more impressive it is when it succeeds. Despite Roberts and Pashler's compelling commentary, the information necessary to evaluate model flexibility is rarely reported.

But for me there is a more fundamental problem with parameter fitting – the fact that we, with some exceptions, fit our models anew for each task or manipulation and that we often use slightly adjusted versions of the model in each publication, but nevertheless quote the model's previous successes to argue for its cumulative support, even though different parameters really mean a different model. Sometimes we hedge our bets, and we write that "parameters were consistent with previous fits of the model" or that "variation in parameter values reflect differences between tasks and or populations" – as it happens both are phrases taken from my own papers (Ma et al., 2022; Popov & Reder, 2020) – but honesty would compel us to admit that these are little more than convenient excuses during model development.

The prior paragraphs make it sound like modeling is pointless because you can make a model do anything you want if you fiddle with the parameters and assumptions long enough. Judging by common reactions to modelling work from non-modeler colleagues, there are many people in the field who believe this. Yet any modeler knows how difficult it is to find even one model that works reasonably well (Shiffrin & Nobel, 1997), especially if you consider more than one effect at a time. Because of this I understand why we have settled on parameter fitting as our dominant approach – during model development we use it to reject many unsuitable models, which is useful even if it doesn't make it to the published product because it increases our confidence in the viability of the model that eventually passes the test.

But in this article, I am not concerned with what makes a model viable, but with what would make us believe it's the best possible description of how human memory works. If we accept the standard set by Newton, the answer is clear – our final model needs to fit individual data points precisely without fitting parameters to do so. At first glance, in practice this standard in its pure form seems unsuitable for a theory of human memory. We know that concrete, semantically related materials are easier to learn than abstract materials; that people put more effort into learning when they expect a recall test rather than a recognition test; that some people are more skilled than others. How are we to account for these observations unless we let some parameters such as learning rate for example to vary across materials, conditions, and people?

A potential answer once again comes from examining Newton's achievement. In the law of universal gravitation, the key parameters – mass and distance – were never estimated to fit the data. Instead, they were obtained independently. If we subscribe to a realist view, the parameters of our psychological models stand for unobservable but real quantities and in principle we should be able to measure them. This is in essence what our standard parameter fitting procedure is doing, but the problem is that we are double-dipping – we use the data to estimate the model parameters and then we use the good fit of the model as support for it.

The standard solution to that problem in the statistical literature is to try out-of-sample validation or cross-validation – to fit a model to one set of data and to evaluate its fit to another set of data.

However, most of the time the held-out dataset comes from the exact same task/paradigm – just from a different sample. A final model of memory would be able to estimate its parameters from one task or manipulation and then be able to fit with precision data coming from another task or manipulation (Busemeyer & Wang, 2000). We can call this standard an "out-of-task" or "out-of-condition" prediction[6], or simply, generalization (also see, Lee et al., 2019; Shiffrin et al., 2008).

One of the best recent examples I know of such a practice comes from a recent paper by Oberauer and colleagues (2022) who set out to test one of the key assumptions of their SOB model of serial recall – novelty-gated encoding (Farrell & Lewandowsky, 2002; Oberauer et al., 2012). They designed a new task, and instead of fitting the model to the new data, they generated specific point predictions with SOB's default parameters. The model was refuted – to quote from the paper: "the results are inconsistent with the effect sizes predicted by SOB-CS. One *could* [emphasis added] change the parameters of SOB-CS to make the predicted effect smaller, but that would involve departing from the parameter values that we have consistently used for deriving predictions from SOB-CS in the past" (Oberauer et al., 2022). This was a difficult test of the model, and had it passed it, it would have received much stronger support than if it had simply fit the new data like we often do.

At this point you could object that some parameters will need to vary across conditions, like the example of the learning rate parameter I mentioned earlier, and I agree this seems inevitable. But there are two principled ways to do this without reverting to our current habits. One way I already mentioned – to estimate these parameters independently of the experimental results we are trying to fit. A second way would be for the theory itself to specify when and by how much parameters should vary. Maybe this is too tall an order during the model development process. But ultimately, whenever the project of a theory of human memory is completed, there should no free parameters at all – their values would be fixed and known, and the theory should be able to predict the precise memory performance given any set of conditions. One exception to this is that parameters inevitably need to vary across people, which is my next major point.

4.4. Individual differences are key

Imagine someone walks into your lab and says: "I have a foreign language exam next Friday. I need to learn this list of words. How many times, in what order and when do I need to study them to pass the test with flying colors?" They are not interested in how the average person would achieve this, but in how they specifically need to study. How would you go about answering this question? One way is obvious but theoretically uninteresting – you could directly measure it by asking them to study the words in question until they reach perfect recall criterion on the day of the test. Your answer would be correct, but cumbersome to achieve and completely atheoretical.

An alternative to this pure empirical approach would be to use the final theory of memory, if such existed, and to plug into it a few key pieces of information – the number and the identity of the words the person wants to learn, the type and date of the test, and the desired memory performance. Then the model would provide a precise answer given its default parameters – ten repetitions per word (less for some, more for others) spaced at such and such intervals.

---

[6] Note that in this context, a prediction doesn't mean that the data isn't gathered yet or that you as the experimenter don't know what the result is. It means that the theory itself is blind to it.

The person, amazed with this technology of the future, goes home; yet a week later you receive an angry call – they failed the test, and they want their money back. What happened?

Most of the time we pay lip service to individual differences in our modeling work – we all agree they are important, and much valuable empirical research has gone into understanding how and when people differ in their memory ability. Yet, despite regularly appearing articles that bring attention to the pitfalls of ignoring such differences in modeling (Estes, 1956, 2002; Gallistel et al., 2004; Musfeld et al., 2022), most of our efforts have gone into modeling aggregated effects.

Clearly, if we possess a final theory of human memory, it needs to be able to make predictions for individual people's behavior. But as I outlined in the previous section, it is insufficient to show that fitting a set of parameters to a particular individual's dataset successfully reproduces the data. Ideally, what we want is the ability to independently determine these parameters and then plug them in the model along with the task description to predict performance – a calibration of the model. This would be an extension of the generalization criterion described before, but at the individual level. A model that could answer with precision our imaginary person's question would truly be something to admire.

To illustrate how this might work, imagine the simplest yet unrealistic possibility that any set of model parameters would lead to a unique level of overall performance. If that were the case, then all we would need to do is to administer a short, standardized memory test and determine the percentage of information recalled. This would give us the parameters of the model we need to predict any future performance.

For any realistic model with multiple parameters more information would be needed for the calibration stage, because they cannot be uniquely determined from overall performance – we would need to measure the person's performance in a few key conditions that the research of the future has established as uniquely constraining the parameter space (e.g. serial positions curves, varying list length). This would be a standard battery of minimal tests needed to fix the parameters of the model, informed by priors established for the population.

Objection: "People are too complicated"

People have incredibly rich internal lives and memories. During any memory attempt, they bring so much to bear that differs wildly from one person to another – their existing knowledge, motivation, strategies, and neuropsychological condition are but a few things that distinguish them from one another. Is it possible to boil down these complex differences to variations in a few parameters?

It is a valid concern. Some relevant differences might not be quantifiable; however, I feel hopeful. Take for example word frequency effects, a topic I have spent a lot of work on. Normally, we measure word frequency at the population level based on large public corpora. People read different things and different amounts, so these population measures necessarily miss a lot of important information. With the widespread digitalization, however, it has become possible to use the content produced by individual people to generate personalized statistics (Yim et al., 2020). Word frequency is a small aspect of prior knowledge, but I suspect that in the next decades we will see similar techniques emerge to quantify more of people's existing knowledge.

Consider also what we know about different memorization strategies – they have reliable, quantifiable effects on key empirical benchmarks such as serial position curves and recall clustering (Unsworth et al., 2019). These benchmarks can be described with existing computational models, so it is not that far-fetched to presume that the effects of strategy use on memory could be captured in

parameter differences. Alternatively, we would need to understand how different strategies affect the format of the stored representations, creating chunks or associations where normally none would exist. The effects of some strategies such as chunking have already been implemented in some memory models (Brady & Tenenbaum, 2013; Nassar et al., 2018) and it is likely that such methods could be extended to account for individual differences in strategy use. Furthermore, empirical methods have been developed to track strategy use during encoding in real-time (Bartsch & Oberauer, 2021; Dunlosky & Kane, 2007), and I expect that future improvements would make this task even easier.

As for motivation, there is a long tradition in the field of modeling its effects on things like learning rate or response bias. And even neuropsychological differences and memory disorders can nowadays be decomposed into meaningful model parameters (Ratcliff & McKoon, 2022).

This is not to say that people can be reduced to a set of numbers without any loss of information. Reducing planets to their mass and distance ignores most of their properties – the questions is whether this loss of information has any bearing on observable behavior. Perhaps I am too optimistic, but I expect that most of what is relevant for memory performance can be quantified. My claim is that the structure of the memory systems, shaped by evolutionary, environmental and development pressures is fixed – an assumption I hope easy enough to swallow. Therefore, any variation in performance would have to boil down to variation in parameters and prior content, both of which we could potentially measure.

4.5. We need to be able to answer practical questions.

Consider again the question I posed in the previous section – "How should I study?" Many academic articles have been written choke full of practical advice for students: use vivid imagery; test yourself; repeat often; space out the repetitions; connect the material with what you already know; think deeply about the material and restate it with your own words. To the casual observer it may seem that academic psychology has made valuable practical contributions. But most of this advice has been known for thousands of years, discovered by laymen through trial and error. But even if it were a unique contribution of modern psychological science, it would be one based entirely on empirical work. Is there any mnemonic advice we could derive from our current theoretical understanding of human memory – advice that doesn't boil down to summarizing empirical effects? I struggle to think of any and this troubles me.

The strength of computational models of memory lies in their precision and that is where they can potentially go beyond the qualitative advice listed above – not just "space out the material", but how and when. Given a set of new vocabulary to learn they could instruct the learner precisely when each word should be repeated for maximum effect – something that simple empirical generalizations cannot hope to achieve, unless they have measured the entire task-parameter space. To do this, however, modeling efforts need to go beyond parameter fitting, as I argued.

The burgeoning field of cognitive tutoring (Anderson et al., 1995; Pavlik et al., 2008) is a testament to the potential of such techniques, but I want to stress that such practical relevance is more than just a useful byproduct of a good theory – that it should be one of the key markers by which we evaluate models against each other – not only by how well they describe empirical regularities, but by how successfully they can assist learning.

Should we today be handed down by God the ground-truth model of memory, given a specific set of materials, task parameters, knowledge of the learner's background and their learning goals, the model would be able to provide the optimal learning schedule and strategy.

Let me take up another angle. How many hours of practice does it take to reach expertise? Due to Malcom Gladwell's best-selling books, based on empirical work by Ericsson, Simon and Chase, most laymen would readily volunteer the number 10,000 as an answer. Had we not known this number empirically, could we have derived it from contemporary models of memory? I don't know, and this intellectual exercise deserves its own article, but I am surprised it has not occurred to us to ask this question before. Answering it would require addressing many smaller questions along the way (What do we define as expertise? In what domain? Does retrieval speed matter or only accuracy?); these details aside, what seems clear is that any serious contender for a model of memory would need to be able to derive such an important number.

We should keep our eye on the prize. Accounting for lab-based empirical effects is an important step in theory development, but such effects are only some of the tools our theory evaluation belt (Ranganath, 2022; van Rooij & Baggio, 2021). There is a commonly repeated phrase among mathematical psychologists – "all models are wrong, but some are useful" (Box, 1976). We need to ask – useful for what?

5. Putting it all together

"Look", said God, "This is how human memory works."

Faced with this improbable scenario, how would we know we are staring at the real deal and not at mere parlor trick? According to the Principle of Completeness, broken into the five claims above, we will be certain that the model is correct if we can estimate its parameters for a particular individual on one set of memory tests and use it to predict with precision performance in any other memory task (Figure 1). It would be nice, wouldn't it?

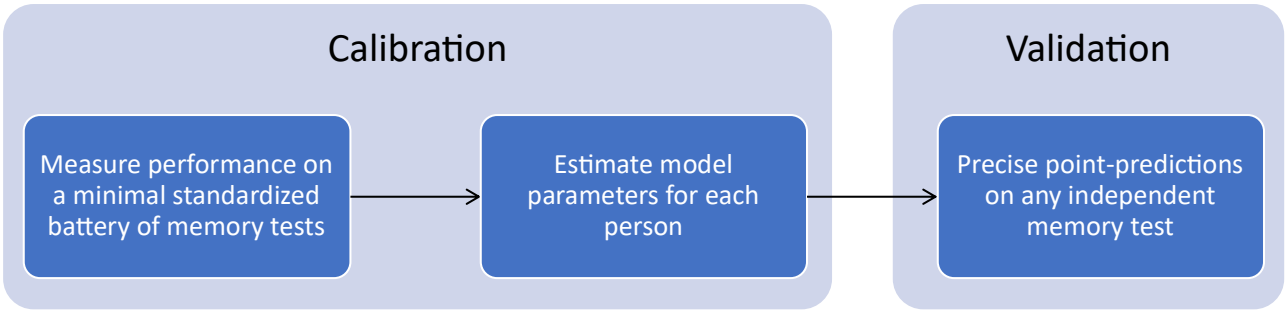

Figure 1. Two stage procedure of testing memory models

6. How do we get there?

What do we need to systematically evaluate whether any future model of memory satisfies the *Principle of Completeness*? Simply put, we need more and better data: a massive dataset from a representative sample of participants, each of which performs a large representative number of

memory tasks and manipulations that cut across theoretical divisions of human memory. A dataset of benchmark findings, not at the aggregate but at the individual level.

To better explain what I mean, I will discuss three existing projects that serve as an inspiration for this idea – community-based agreement on benchmark findings in working memory (Oberauer et al., 2018a), the Penn Electrophysiology of Encoding and Retrieval Study dataset (PEERS; Kahana et al., 2022) and the Cox and colleagues dataset (Cox et al., 2018).

Research on memory has generated a wealth of empirical regularities that any theory should be able to explain. In an effort to systematize this information for the working memory domain, Oberauer and colleagues compiled a list of "benchmarks" – replicable findings concerning various encoding and retrieval conditions such as set size effects, serial position curves, retention intervals, etc (Oberauer et al., 2018a). These benchmarks represent targets for explanation within the working memory domain. This is a commendable goal because such benchmarks provide "common ground for model development" (Oberauer et al., 2018b). However, as I discussed in Section 3.1, any model of memory would need to be able to explain effects across a variety of memory types, and we need to compile such benchmark findings not only for working memory, but for memory at large.

Compiling a list like the one provided by Oberauer and colleagues is, however, not sufficient. It depends on archival data, much of which is difficult or impossible to access. Whenever it is accessible, most often it is in aggregated form, both over conditions and participants. This presents two problems for model evaluation. First, if you are interested in modeling interactions between conditions, for example, set size and presentation rate, you can only do that if an existing study manipulated both factors and reported the data at a fine enough level of detail. This is often not the case, and modelers end up having to redo studies anyway (Oberauer, personal communication, 2022). Second, if we take the Principle of Completeness seriously, we need to 1) fit the model to a subset of data for one individual and 2) test it on other memory tests performed by the same individual. Because most studies reporting these benchmarks are done with different groups of people, and because they rarely provide individual-level data, this goal is impossible to achieve currently.

This is where a project like PEERS (Kahana et al., 2022) or Cox et al's dataset (2018) come in. Both projects provide publicly available datasets on memory tasks performed by hundreds of participants, each of which completed hundreds of different memory trials. In PEERS, participants studied multiple 16-word-lists for a subsequent free recall or recognition task on each of dozens of sessions. Thus, for each individual participant we could obtain metrics such as their serial position curves, overall recall or recognition rates, temporal and semantic clustering during recall, word frequency effects, etc. Similarly, in Cox et al's study, participants studied many lists of word pairs, and each list was randomly followed by one of five different tests – item recognition, associative recognition, cued-recall, free recall, or lexical decision. Again, for each individual we could obtain metrics for various established long-term memory benchmarks across a variety of test types.

Both projects are commendable in their scope, but they are limited in two ways. First, both only contain data on episodic memory. Second, both include idiosyncratically chosen manipulations and cannot serve the role I envision for systematic model evaluation using the principle of completeness. For example, neither project manipulates presentation rate, retention intervals, list length or a host of other factors that affect memory. Their strength lies in the amount of data available for each participant.

What we need is a dataset that combines the strengths of the two approaches – 1) the systematic selection of benchmark findings like the one provided by Oberauer and colleagues, and 2) fine-grained data for each of these benchmarks at the individual level like the one provided by PEERS and Cox's

datasets. The encoding and retrieval conditions need to span the gamut of memory task, manipulations, and paradigms.

This will clearly be an onerous task, but one that I believe necessary. The Newtonian achievement in physics would have been impossible without the decade-long effort by Tycho Brache to accurately record the daily positions of planets and stars. Such a dataset would allow us to fit comprehensive models to individual people's datasets, to make sure the models can describe more than just aggregate data and idiosyncratically picked effects, and to identify the minimum measurements required to calibrate the model. It would also allow us to quantify the explanatory breath of current models – which findings they apply to and what is left to be done.

7. Concluding thoughts

For several years now I have been troubled that we are running around in circles and that our theoretical efforts to understand human memory never really amount to much. Watkins (1990) put it bluntly when he wrote that "memory theorizing is going nowhere". On my most pessimistic days I am inclined to agree. And yet, I do know more about memory now compared to when I began my research career. Unfortunately, this knowledge seems more like a random collection of facts rather than the kind of cumulative, cohesive theoretical knowledge we have come to expect of mature sciences. Perhaps the field is in better shape than I portrayed it here and these complaints betray nothing more than my own ignorance, but I know many colleagues often feel the same. At best, I can describe my knowledge as meta-knowledge – knowing what various theories say about how various aspects of memory function. The question that often bothers me is how do we go from this meta-knowledge about the beliefs of memory theorists to actual knowledge about how memory works.

My solution is to start with the end goal in mind. As I was finally writing this article, my mood often oscillated between two opposing thoughts – "This is trivial" and "It will never work." I still cannot decide between them, but what I am convinced of is that we need to have a conversation about where we are going. Perhaps my proposal is misguided or too ambitious, but I welcome the discussion either way. My goal is not to defend the ideas presented here at all costs; although I believe there is merit to them (otherwise why would I write them?), my sincere hope is that this article brings the issue of what we want out of a theory of memory to the forefront of debate. Even if you reject my particular proposal, the question remains the same – how will we recognize the correct theory of memory?

Various previous papers have made specific proposals as to how we should change our research strategy in studying memory (Glenberg, 1997; Hintzman, 2011; Watkins, 1990). Hintzman advocated that we should focus less on established, seemingly objective lab-based paradigms and focus more on the qualitative aspects of memory. Watkins suggested that we should abandon mediationism, the idea that memory is mediated by memory traces, and instead just document empirical regularities. Glenberg proposed that we should ask what the purpose of memory is and start from there. I have no similar recommendations about our research strategy; a diversity of approaches is crucial to any scientific endeavor. My focus is instead on the end goal – to ask what we want to achieve, before thinking about how we get there.

This article comes on the heels of many papers discussing the "theory-crisis" in psychology (e.g. Eronen & Bringmann, 2021; Navarro, 2021; Oberauer & Lewandowsky, 2019; Maatman, 2021). Although the term is new, concerns about whether we are making sufficient theoretical progress date back at least five decades (Glenberg, 1997; Hintzman, 2011; Meehl, 1978; Newell, 1973; Watkins,

1990). A common proposal is that formal (mathematical or computational) models provide a ready solution to the theory crisis. Judging by the state of the memory literature, I fear that by itself a general appeal for formal modeling is not enough. As I have argued at length, formalism is necessary, but not sufficient. We need to clearly state what we want such models to achieve. The Principle of Completeness provides one such evaluative standard.

Is achieving a theory of memory that fulfills the Principle of Completeness hopeless? Is the complexity of the subject matter so great as to make this Mission Impossible? I conclude with the words of the physicist William Thomson (a.k.a Lord Kelvin), who in 1889, after struggling for decades with the question "What are the basic building blocks of the universe made of?", had to eventually abandon his vortex theory of the atom:

> *I am afraid I must end by saying that the difficulties are so great in the way of forming anything like a comprehensive theory that we cannot even imagine a finger-post pointing to a way that lead us towards the explanation. That is not putting it too strongly. I can only say we cannot now imagine it. But this time next year,– this time ten years, – this time one hundred years, – probably it will be just as easy as we think it is to understand that glass of water, which seems now so plain and simple. I cannot doubt but that these things, which now seem to us so mysterious, will be no mysteries at all; that the scales will fall from our eyes; that we shall learn to look on things in a different way – when that which is now a difficulty will be the only common-sense and intelligible way of looking at the subject. (Thomson, 2011, pp. 510–511)*

Less than a century later, the standard model of particle physics was completed, and it now indeed seems to us so plain and simple. So simple, that many readers have objected to my comparisons between physics and psychology in two ways: 1) the methods of physics are not well suited to the immensely more complicated phenomenon that is the mind; 2) any theory adhering to the Principle of Completeness might end up being as complex as the very phenomenon it tries to explain, thus losing its explanatory power. Hindsight is 20-20, so let us not forget that the only reason physical phenomena appear simple is that we have already discovered their building blocks and relations. A good theory distills chaos into clarity.

8. References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. L. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. https://doi.org/10.1037/0033-295x.111.4.1036

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341–380.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, Ray. (1995). Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences*, *4*(2), 167–207. https://doi.org/10.1207/s15327809jls0402_2

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1120.

Bartsch, L. M., & Oberauer, K. (2021). The effects of elaboration on working memory and long-term memory across age. *Journal of Memory and Language*, *118*, 104215. https://doi.org/10.1016/j.jml.2020.104215

Bernoulli, D. (1738). Hydrodynamica. *Dulsecker. Consultable En Ligne Http://Imgbase-Scd-Ulp. u-Strasbg. Fr/Displayimage. Php*, *1738*.

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

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory Construction Methodology: A Practical Framework for Building Theories in Psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Brady, T., Robinson, M. M., Williams, J. R., & Wixted, J. (2021). *Measuring memory is harder than you think: A crisis of measurement in memory research*. PsyArXiv. https://doi.org/10.31234/osf.io/qd75k

Brady, T., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*(1), 85–109. https://doi.org/10.1037/a0030779

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539.

Bunn, T. (2012). *Ptolemy and Copernicus*. https://blog.richmond.edu/physicsbunn/2012/09/13/ptolemy-and-copernicus/

Busemeyer, J. R., & Wang, Y.-M. (2000). Model Comparisons and Model Selections Based on Generalization Criterion Methodology. *Journal of Mathematical Psychology*, *44*(1), 171–189. https://doi.org/10.1006/jmps.1999.1282

Clarke, K. a, & Primo, D. M. (2012, March 31). Opinion | Overcoming 'Physics Envy.' *The New York Times*. https://www.nytimes.com/2012/04/01/opinion/sunday/the-social-sciences-physics-envy.html

Cox, G. E., Hemmer, P., Aue, W. R., & Criss, A. H. (2018). Information and processes underlying semantic and episodic memory across tasks, items, and individuals. *Journal of Experimental Psychology: General*, *147*(4), 545–590. https://doi.org/10.1037/xge0000407

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376.

Dalton, J. (1808). *A new system of chemical philosophy*. Manchester : Printed by S. Russell ... for R. Bickerstaff, ... London. http://archive.org/details/newsystemofchemi12dalt

Devezer, B. (2023). *There are no shortcuts to theory*. MetaArXiv. https://doi.org/10.31222/osf.io/umkan

Duhem, P. M. M. (1954). *The Aim and Structure of Physical Theory*. Princeton,: Princeton University Press.

Dunlosky, J., & Kane, M. J. (2007). The Contributions of Strategy Use to Working Memory Span: A Comparison of Strategy Assessment Methods. *Quarterly Journal of Experimental Psychology*, *60*(9), 1227–1245. https://doi.org/10.1080/17470210600926075

Dunn, J. C., & Anderson, L. M. (2022). *The Monotonic Linear Model: Testing for Removable Interactions*. PsyArXiv. https://doi.org/10.31234/osf.io/ydb45

Einstein, A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen Der Physik*, *322*(8), 549–560. https://doi.org/10.1002/andp.19053220806

Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134–140. https://doi.org/10.1037/h0045156

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, *9*(1), 3–25. https://doi.org/10.3758/BF03196254

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review*, *9*(1), 59–79. https://doi.org/10.3758/BF03196257

Fraassen, B. C. van. (2012). Modeling and Measurement: The Criterion of Empirical Grounding. *Philosophy of Science*, *79*(5), 773–784. https://doi.org/10.1086/667847

Fried, E. I. (2020). Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences*, *101*(36), 13124–13131. https://doi.org/10.1073/pnas.0404965101

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

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, *20*(1), 1–19.

Gould, S. J. (1997). Nonoverlapping Magisteria. *Natural History*, *106*, 16–22.

Gráda, C. Ó. (2016). Did Science Cause the Industrial Revolution? *Journal of Economic Literature*, *54*(1), 224–239.

Greenberg, D. L., & Verfaellie, M. (2010). Interdependence of episodic and semantic memory: Evidence from neuropsychology. *Journal of the International Neuropsychological Society : JINS*, *16*(5), 748–753. https://doi.org/10.1017/S1355617710000676

Gregg, V. (1976). Word frequency, recognition and recall. In *Recall and recognition* (pp. x, 275–x, 275). John Wiley & Sons.

Guest, O., & Martin, A. E. (2021). How Computational Modeling Can Force Theory Building in Psychological Science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Habgood-Coote, J., Watson, L., & Whitcomb, D. (2023). Can a good philosophical contribution be made just by asking a question?1. *Metaphilosophy*, *54*(1), 54–54. https://doi.org/10.1111/meta.12599

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, &amp Computers*, *16*(2), 96–101. https://doi.org/10.3758/BF03202365

Hintzman, D. L. (1991). Why are formal models useful in psychology? In *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 39–56). Lawrence Erlbaum Associates, Inc.

Hintzman, D. L. (2011). Research Strategy in the Study of Memory: Fads, Fallacies, and the Search for the "Coordinates of Truth." *Perspectives on Psychological Science*, *6*(3), 253–271. https://doi.org/10.1177/1745691611406924

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), Article 7866. https://doi.org/10.1038/s41586-021-03659-0

Kahana, M., Lohnas, L., Healey, K., Aka, A., Broitman, A., Crutchley, E., Crutchley, P., Alm, K., Katerman, B., Miller, N., Kuhn, J., Li, Y., Long, N., Miller, J., Paron, M., Pazdera, J., Pedisich, I., & Weidemann, C. (2022). *The Penn Electrophysiology of Encoding and Retrieval Study*. https://doi.org/10.31234/osf.io/bu5x8

Kellen, D., Davis-Stober, C. P., Dunn, J. C., & Kalish, M. L. (2021). The Problem of Coordination and the Pursuit of Structural Constraints in Psychology. *Perspectives on Psychological Science*, *16*(4), 767–778. https://doi.org/10.1177/1745691620974771

Kellen, D., Winiger, S., Dunn, J. C., & Singmann, H. (2021). Testing the foundations of signal detection theory in recognition memory. *Psychological Review*, *128*(6), 1022–1050. https://doi.org/10.1037/rev0000288

Kollerstrom, N. (1991). Newton's Two "Moon-Tests." *The British Journal for the History of Science*, *24*(3), 369–372.

Lakatos, I. (1976). Falsification and the Methodology of Scientific Research Programmes. In S. G. Harding (Ed.), *Can Theories be Refuted? Essays on the Duhem-Quine Thesis* (pp. 205–259). Springer Netherlands. https://doi.org/10.1007/978-94-010-1863-0_14

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., Matzke, D., Rouder, J. N., Trueblood, J. S., White, C. N., & Vandekerckhove, J. (2019). Robust Modeling in Cognitive Science. *Computational Brain & Behavior*, *2*(3), 141–153. https://doi.org/10.1007/s42113-019-00029-y

Lewandowsky, S., & Farrell, S. (2011). *Computational Modeling in Cognition: Principles and Practice*. https://doi.org/10.4135/9781483349428

Lin, H.-Y., & Oberauer, K. (2022). An interference model for visual working memory: Applications to the change detection task. *Cognitive Psychology*, *133*, 101463. https://doi.org/10.1016/j.cogpsych.2022.101463

Ma, S., Popov, V., & Zhang, Q. (2022). *A Neural Index Reflecting the Amount of Cognitive Resources Available during Memory Encoding: A Model-based Approach* (p. 2022.08.16.504058). bioRxiv. https://doi.org/10.1101/2022.08.16.504058

Maatman, F. O. (2021). *Psychology's Theory Crisis, and Why Formal Modelling Cannot Solve It*. PsyArXiv. https://doi.org/10.31234/osf.io/puqvs

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, *34*(2), 103–115. https://doi.org/10.1086/288135

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.

Meehl, P. E. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review*, *100*(2), 183–203. https://doi.org/10.1037/0033-295x.100.2.183

Musfeld, P., Souza, A. S., & Oberauer, K. (2022). *Repetition Learning: Neither a Continuous nor an Implicit Process*. PsyArXiv. https://doi.org/10.31234/osf.io/6xtwh

Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological Review*, *125*(4), 486–511. https://doi.org/10.1037/rev0000101

Navarro, D. J. (2019). Between the Devil and the Deep Blue Sea: Tensions Between Scientific Judgement and Statistical Model Selection. *Computational Brain & Behavior*, *2*(1), 28–34. https://doi.org/10.1007/s42113-018-0019-z

Navarro, D. J. (2021). If Mathematical Psychology Did Not Exist We Might Need to Invent It: A Comment on Theory Building in Psychology. *Perspectives on Psychological Science*, *16*(4), 707–716. https://doi.org/10.1177/1745691620974769

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, *120*(2), 356–394. https://doi.org/10.1037/a0032020

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*. Academic Press.

Norris, D. (2017). Short-term memory and long-term memory are still different. *Psychological Bulletin*, *143*(9), 992. https://doi.org/10.1037/bul0000108

Oberauer, K., Farrell, S., Jarrold, C., & Niklaus, M. (2022). Evidence Against Novelty-Gated Encoding in Serial Recall. *Journal of Cognition*, *5*(1), Article 1. https://doi.org/10.5334/joc.207

Oberauer, K., & Lewandowsky, S. (2011). Modeling working memory: A computational implementation of the Time-Based Resource-Sharing theory. *Psychonomic Bulletin & Review*, *18*(1), 10–45. https://doi.org/10.3758/s13423-010-0020-6

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018a). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958. https://doi.org/10.1037/bul0000153

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018b). Benchmarks provide common ground for model development: Reply to Logie (2018) and Vandierendonck (2018). *Psychological Bulletin*, *144*, 972–977. https://doi.org/10.1037/bul0000165

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review*, *19*(5), 779–819.

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21–59. https://doi.org/10.1037/rev0000044

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review*, *122*(2), 260.

Oude Maatman, F. (2021). *Psychology's Theory Crisis, and Why Formal Modelling Cannot Solve It* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/puqvs

Pavlik, P., Bolster, T., Wu, S.-M., Koedinger, K., & Macwhinney, B. (2008). Using optimally selected drill practice to train basic facts. *International Conference on Intelligent Tutoring Systems*, 593–602. http://link.springer.com/chapter/10.1007/978-3-540-69132-7_62

Platt, J. R. (1964). Strong Inference. *Science, New Series*, *146*(3642), 347–353.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156. https://doi.org/10.1037/a0014420

Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review*, *127*(1), 1–46. https://doi.org/10.1037/rev0000161

Popov, V., & Reder, L. M. (in press). Frequency effects in recognition and recall. In M. J. Kahana & A. D. Wagner (Eds.), *The Oxford Handbook of Human Memory*. Oxford University Press. https://doi.org/10.31234/osf.io/xb8es

Ranganath, C. (2022). What is episodic memory and how do we use it? *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2022.09.023

Ratcliff, R., & McKoon, G. (2022). Can neuropsychological testing be improved with model-based approaches? *Trends in Cognitive Sciences*, *26*(11), 899–901. https://doi.org/10.1016/j.tics.2022.08.015

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 294.

Reder, L. M., Park, H., & Kieffaber, P. D. (2009). Memory Systems Do Not Divide on Consciousness: Reinterpreting Memory in Terms of Activation and Binding. *Psychological Bulletin*, *135*(1), 23–49. https://doi.org/10.1037/a0013974

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible Hands and Fine Calipers: A Call to Use Formal Theory as a Toolkit for Theory Construction. *Perspectives on Psychological Science*, *16*(4), 725–743. https://doi.org/10.1177/1745691620974697

Robinson, M. M., & Steyvers, M. (2023). Linking computational models of two core tasks of cognitive control. *Psychological Review*, *130*(1), 71.

Robinson, M. M., Williams, J. R., & Brady, T. (2022). *What does it take to falsify a psychological theory? A case study on recognition models of visual working-memory*. PsyArXiv. https://doi.org/10.31234/osf.io/7an3x

Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, *22*(4), 944–954. https://doi.org/10.3758/s13423-014-0759-2

Ryle, G. (2000). *The Concept of Mind*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/C/bo3684918.html

Shiffrin, R. M. (2010). Perspectives on modeling in cognitive science. *Topics in Cognitive Science*, *2*(4), 736–750.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A Survey of Model Evaluation Approaches With a Tutorial on Hierarchical Bayesian Methods. *Cognitive Science*, *32*(8), 1248–1284. https://doi.org/10.1080/03640210802414826

Shiffrin, R. M., & Nobel, P. A. (1997). The art of model development and testing. *Behavior Research Methods, Instruments, &amp Computers*, *29*(1), 6–14. https://doi.org/10.3758/BF03200560

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166.

Singmann, H., Kellen, D., Cox, G. E., Chandramouli, S. H., Davis-Stober, C. P., Dunn, J. C., Gronau, Q. F., Kalish, M. L., McMullin, S. D., Navarro, D. J., & Shiffrin, R. M. (2023). Statistics in the Service of Science: Don't Let the Tail Wag the Dog. *Computational Brain and Behavior*, *6*(1), 64–83. https://doi.org/10.1007/s42113-022-00129-2

Smaldino, P. E. (2017). Models Are Stupid, and We Need More of Them. In *Computational Social Psychology*. Routledge.

Smart, W. M. (1946). John Couch Adams and the Discovery of Neptune. *Nature*, *158*(4019), Article 4019. https://doi.org/10.1038/158648a0

Starns, J. J., Cataldo, A. M., Rotello, C. M., Annis, J., Aschenbrenner, A., Bröder, A., Cox, G., Criss, A., Curl, R. A., Dobbins, I. G., Dunn, J., Enam, T., Evans, N. J., Farrell, S., Fraundorf, S. H., Gronlund, S. D., Heathcote, A., Heck, D. W., Hicks, J. L., … Wilson, J. (2019). Assessing Theoretical Conclusions With Blinded Inference to Investigate a Potential Inference Crisis. *Advances in Methods and Practices in Psychological Science*, *2*(4), 335–349. https://doi.org/10.1177/2515245919869583

Taton, R., Wilson, C., & Hoskin, M. (2003). *Planetary Astronomy from the Renaissance to the Rise of Astrophysics, Part A, Tycho Brahe to Newton*. Cambridge University Press.

ThinkWeekOxford (Director). (2013, March 13). *In Conversation with Richard Dawkins—Hosted by Stephen Law*. https://www.youtube.com/watch?v=zvkbiElAOqU

Thomson, W. (Ed.). (2011). ETHER, ELECTRICITY, AND PONDERABLE MATTER. In *Mathematical and Physical Papers: Volume 3: Elasticity, Heat, Electro-Magnetism* (Vol. 3, pp. 484–515). Cambridge University Press. https://doi.org/10.1017/CBO9780511996023.013

Unsworth, N., Miller, A. L., & Robison, M. K. (2019). Individual differences in encoding strategies and free recall dynamics. *Quarterly Journal of Experimental Psychology*, *72*(10), 2495–2508. https://doi.org/10.1177/1747021819847441

Van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Clarendon press.

van Rooij, I., & Baggio, G. (2021). Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. *Perspectives on Psychological Science*, *16*(4), 682–697. https://doi.org/10.1177/1745691620970604

van Rooij, I., & Blokpoel, M. (2020). Formalizing verbal theories. *Social Psychology*.

Watkins, M. J. (1990). Mediationism and the obfuscation of memory. *American Psychologist*, *45*, 328–335. https://doi.org/10.1037/0003-066X.45.3.328

Westfall, R. S. (1973). Newton and the Fudge Factor. *Science*, *179*(4075), 751–758. https://doi.org/10.1126/science.179.4075.751

Wittgenstein, L. (1972). *On certainty*. New York, Harper & Row. http://archive.org/details/oncertainty00witt

Wootton, D. (2015). *The Invention of Science: A New History of the Scientific Revolution*. Penguin UK.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yim, H., O'Brien, C., Stone, B., Osth, A., & Dennis, S. (2020). Using Emails to Quantify the Impact of Prior Exposure on Word Recognition Memory. *CogSci*.