# How computational modeling can force theory building in psychological science

Olivia Guest
Research Centre on Interactive Media, Smart Systems and Emerging Technologies — RISE,
Nicosia, Cyprus &
Department of Experimental Psychology,
UCL, UK

Andrea E. Martin
Max Planck Institute for Psycholinguistics,
Nijmegen, The Netherlands &
Donders Centre for Cognitive Neuroimaging, Radboud University,
Nijmegen, The Netherlands

Psychology is a broad field that endeavors to develop explanatory theories of human capacities and behaviors based on a wide variety of methodologies and dependent measures. Here we argue that whether or not researchers choose to employ modeling (viz., choose to create computational models of their theories over and above their data during the scientific inference process) is one of the most important and divisive factors in our field. Modeling is underdiscussed and underemployed, yet, in our view, holds integrative promise for advancing the goals of psychological science. The inherent demands of computational modeling offer invaluable momentum towards a better, and more open, psychological science. These demands force the scientist to conceptually analyze, specify, and ideally, formalise intuitions and ideas which would otherwise remain implicit or unexamined — something we propose should be called "open theory". Constraining our inference process through specification and modeling is what will enable us as a field to meaningfully interpret data, and to build theories that explain and predict. In this piece, we present scientific inference in psychology as a *path function*, where each step shapes the next. Computational modeling can constrain the steps in the path, and has the potential to advance scientific inference over and above the stewardship of the experimental practice (e.g., preregistration, choosing frequentist or Bayesian statistics, power and sample size, and other estimation variables). If as a field we continue to eschew, inadvertently avoid, or remain ignorant of formal and computational modeling, we set ourselves up for a persistent lack of replicability and, moreover, for failure at coherent theory-building that includes explanatory force. We explain how the basic steps in the modeling process can be accomplished and we touch on the cultural and practical issues that need to be faced therein, emphasizing that the advantages of modeling can be achieved by anyone with benefit to all. The process of computational modeling promotes transparent theorising; "open science" should include open theory alongside, e.g., open data and open source code.

*Keywords:* computational model; path function; scientific inference; psychology; theoretical psychology; research methods; cognitive science; open science; replicability; reproducibility; philosophy of science; pedagogy

## Challenges for scientific inference in psychological science

Psychology is a science that endeavors to develop theories that explain the capacities and behaviors of the human organism. In practice, this results in a wide range of research modes, from designing and running behavioural and neuroscientific experiments (e.g., performing basic science to investigate a capacity or behavior), to carrying out clinical work (e.g., studying and treating patients in a clinical setting), to qualitative work (e.g., interpretative phenomenological analysis). Psychology intersects with many other fields, creating sub-fields that are highly interdisciplinary across the spectrum of science, technology, engineering, mathematics, and the humanities. In this article, we do not intend to provide an exhaustive definition of "psychology", but rather, we focus on a distinction within psychological science that is less often discussed: the difference in explanatory force between research programmes that use formal, mathematical, and computational modeling and those that do not.

We start by explaining in a basic way what a computational model is, and we illustrate how specifying a model naturally results in better-specified theories, and therefore in better science. We give an example of a specified, formalized, and implemented computational model and use it to model a cartoon example where intuition is insufficient in determining a quantity (viz., an area). Next, we present a

verbal and pictorial model of a characterization of how psychological science should be done in order to maximize the relationship between theory, specification, and data. Our claim here is that the scientific inference process is a function from theory to data — but this function must be more than a state function to have explanatory force — it is a *path function* which must step through theory, specification, and implementation before a interpretation can have explanatory force in relation to a theory. Finally, we outline the steps we believe the field needs to take to use modeling to address the structural problems in theory building that underlie the so-called replication "crisis" in, e.g., social psychology. We propose a core yet overlooked component of open science that computational modeling forces scientists to carry out: *open theory*.

## A fork in the path of psychological science

Regardless of the level(s) of analysis — from neuron to behaviour — that a psychological scientist's work can be placed, there are certain core meta-theoretical ideas or assumptions that all or the vast majority of modern psychology agrees on. For example, *a*) that the brain gives rise to behaviour, or that *b*) networks of neurons can perform computations. In turn, agreement with these default positions means researchers can then believe, e.g., for *a*) that by looking at behaviour we can understand something about how brains work and vice versa, and for *b*) that by looking at the inputs of outputs of the neural system we can understand important properties and principles of the system. Thus, psychological scientists endeavor to discover how and why things work the way they do — typically by ascribing to (implicitly or explicitly) a school of thought and by specifying (implicitly or explicitly) theoretical accounts, or at the least some basic hypotheses, to test using inferential statistics. From there, clinical practitioners, for example, use the outputs of this research, implicitly (or not) subscribing to the theoretical positions of the researchers and agreeing with the value of the hypotheses generated and tested at the end of this scientific pipeline.

On the one hand, a typical psychology experiment will gather data to test an explicit hypothesis and will analyse that data using (overwhelmingly frequentist) inferential statistics. And this is and has been true for all branches of psychology that deal with data for a long time (Meehl, 1967; Newell, 1973). While there are important differences between subfields of psychology, similarities are also highly apparent in terms of methodology. Almost every paper we publish can be boiled down to introduction, methods, analysis, results, and discussion. Save for differences in jargon and specific methods, the way we approach science is near identical: we ask nature questions by collecting data and then report *p*-values, more rarely Bayes-factors or Bayesian inference, or some qualitative measure. Computational models do not feature explicitly in the majority of psychology's scientific endeavours. Most papers do not include computational modeling, and most psychological researchers are not trained in modeling beyond constructing statistical models of their data. In fact, many, while respectful of formal modeling techniques, still assume GUIs or preset models are useful outside the simplest cases or outside pedagogical contexts (Cooper & Guest, 2014). This is something that causes friction and is the source of further misunderstandings and miscommunications — perhaps it is due to generalising from data models, i.e., inferential statistics, which are typically applicable off-the-shelf. On the other hand, a subset of researchers — formal, mathematical, or computational modelers — take a different route in the idea-to-publication pipeline. They construct models of something other than the data directly.

In our view, the true task of the modeler is to create semi-formalised or formalised versions of scientific theories, often creating (or least amending) their accounts along the way. A computational modeler is somebody who has the tools to be acutely aware of the assumptions and implications of the theory they are using to carry out their science. This awareness comes, ideally, from specification and formalization, but minimally, it also comes from the necessity of writing code during implementation. Thus, involving modelers in a research programme has the effect of necessarily changing the way the research process is structured. It changes the focus from testing hypotheses generated from an opaque idea or intuition (e.g., a theory that has likely never been written down in anything other than natural language, if that), to testing a formal model of the theory as well as continuing to also be able to generate and test hypotheses using empirical data. Computational modeling does this by forcing the scientists involved to explicitly write down (e.g., in programming code) an instance of what their theory assumes, if not what their theory is. In our view, the most crucial part of the process is creating a specification, but even just creating an implementation (programming code) leverages more explicitness than going from framework to hypothesis to data collection directly. In addition to formalisation, introducing computational modeling makes the process of science more transparent. Computational modeling is a canonical open

Olivia Guest was supported by the Research Centre on Interactive Media, Smart Systems and Emerging Technologies (RISE) under the European Union's Horizon 2020 programme (grant 739578) and the Republic of Cyprus through the Directorate General for European Programmes, Coordination and Development.

Andrea E. Martin was supported by the Max Planck Research Group "Language and Computation in Neural Systems" and by the Netherlands Organization for Scientific Research (grant 016.Vidi.188.029).

The authors would like to thank: Karim N'Diaye, Zach Shifrel, Eiko Fried, for useful input.

science methodology due to the light it shines on the usually opaque practice of theory building and testing (Nosek et al., 2015). The process of building a computational model, we propose, should be dubbed: *open theory*. We will unpack further what computational modeling has to offer psychological science more in the next few sections.

### What is a computational model? And why build one?

> Let us calculate, without further ado, and see who is right (Leibniz, 1685; translated by: Wiener, 1951)

Gottfried Wilhelm Leibniz (1646–1716) predicted computational modeling when he envisaged a *characteristica universalis* that allows scientists to formally express theories and data (e.g., formal languages, logic, programming languages) and a *calculus ratiocinator* that computes the logical consequences of theories and data (e.g., digital computers; also see: Cohen, 1954; Wiener, 1951). Modeling can thus be seen as a universal (formal) language by which scientists communicate. More specifically, computational modeling is the process by which a verbal (or pictorial, etc.) description is formalised (e.g., using logic or mathematics, or another formal language, pseudocode, or even a programming language) in ways that ground it. Computational modeling removes ambiguity, as Leibniz correctly predicted, while also constraining the space or dimensions in which a theory can span.

In the best of possible worlds, modeling makes us think deeply about what we are going to model, (e.g., which phenomenon or capacity), in addition to any data, both before and during the creation of the model, and both before and during data collection. It can be as simple as the scientist asking themselves, "what is it we are proposing? How do we understand the brain and behaviour in this context, and why?" By thinking through how to represent the data, model the experiment, etc., scientists gain insight into their ideas and intuitions, and the computational repercussions of their ideas, in a much deeper and explicit way than by just collecting data in relation to framework or implicit idea. Without such care we may end up wasting resources and time pursuing scientific goals that are based on an incomplete picture of the literature or are otherwise impaired.

Modeling allows us to automate, to an extent, hypothesis generation and even in some cases code generation. It guides us away from testing hypotheses that are implausible or irrelevant given existing knowledge, and from building a nonsensical or bogus implementation given what we already know about the system we are trying to approximate. Such grounding rules out vast swathes of research, saving researchers time and money. Furthermore, a model allows us to make clear cut and falsifiable predictions. By providing a transparent genealogy for where predictions, explanations, and ideas for experiments come from, the process of modeling stops us from atheoretically testing hypotheses — a core value of open science. Open theorising is done by default by many modelers as a function of the (computational and/or formal) modeling process.

Through modeling, even in, or especially in, failures we hone our ideas: can our theory be formally specified, and if not, why not? Making explicit what might have been completely implicit is extremely useful in scientific inquiry. Thus, we may check if what we have described formally still makes sense in light of our theoretical commitments. It aids both us as researchers communicating with each other, and it aids those who may wish to apply these ideas to their work outside science in e.g., industrial or clinical settings.

Modeling also allows us to perform model comparison — to compare different parameter values' effects within a model and compare models based on one theory to those based on another. When two (or more) theories can make sense of the present data, this is one of the only ways to dissociate between them in a formal setting (although also see: Navarro, 2019).

In the rest of this section we will walk the reader through building a computational model from scratch in order to illustrate our argument, and then present a path function of research in psychology. We emphasize that, often, 'merely' building a formal model of a problem is not enough – actually writing code to implement a computational model is required to understand the model itself. We call this gap in understanding the "pizza problem" for reasons that will become clear.

### The pizza problem

> All models are wrong but some are more wrong than others. (pastiche based on: Box, 1976; Orwell, 1945)

Imagine it is Friday night, you are hungry, and so decide to order pizza with a friend. You call up your favourite pizzeria and they tell you they have an offer: two 12" pizzas for the price of one 18". Your definition of a good deal is one in which you purchase the most food. Is this a good deal?

A Twitter user by the name of Fermat's Library (@fermatslibrary) posted as "a useful counterintuitive fact [that] one 18 inch pizza has more 'pizza' than two 12 inch pizzas"[1] — along with an image similar to Figure 1. The reaction to this tweet was largely surprise or disbelief, with user @MarkSykes15 replying: "But two pizzas are more than one".[2] What is happening here and why are people taken aback?

When it comes to comparing ceteris paribus the two options depicted in Figure 1, even if everybody agrees what the area of a circle is defined as (e.g., a mathematical model based on: Area = $\pi R^2$) there are some people unwilling to

[1] Archived tweet: archive.ph/Fb66R

[2] Archived tweet: archive.ph/BoyRs

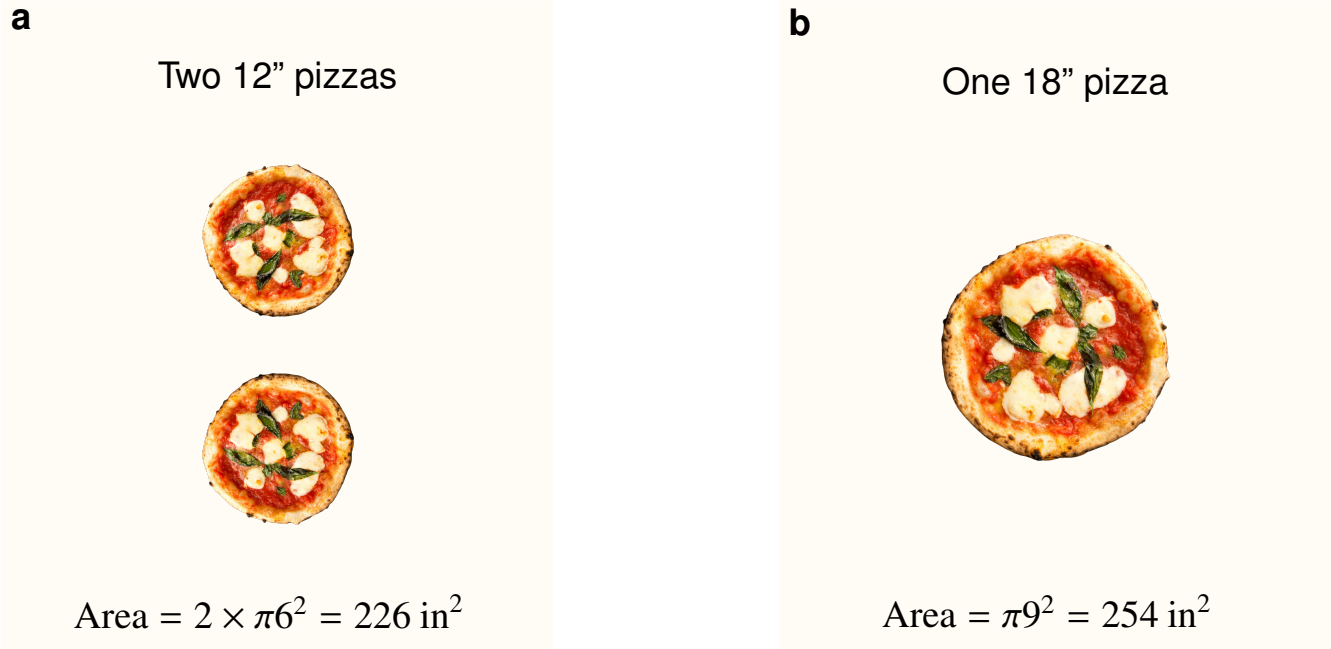


*Figure 1*. The pizza problem: something like comparing the two options above can appear “counterintuitive” even though we all learn the formula for the area of a circle in primary school. Compare **a**) two 12” pizzas with **b**) one 18” pizza (all three pizzas to-scale). Which order would you prefer ceteris paribus?

compute this model (e.g., in their head or on paper) and instead they run a model based on comparing the number of pizzas only. And thus the results of the “true” model, that one 18” pizza has more surface, and therefore is more food in Figure 1, are counterintuitive.

Modeling is able to demonstrate how one cannot always trust one’s gut even when it comes to something as simple as choosing how much pizza to order to maximize value. Moreover, computational modeling — actually implementing and running the model — can further highlight serious misunderstandings. To carry out computational modeling of a given problem or phenomenon one must have or create: *a*) a verbal description, a conceptual analysis, and/or a theory; *b*) a formal(isable) description, i.e., a specification using mathematics, pseudocode, flowcharts, etc.; and *c*) an executable implementation written using a programming language. This process is the cornerstone of computational modeling and by extension of modern scientific thought, enabling us to refine our gut instincts through experience. This experience is seeing our ideas being executed by a computer, giving us the chance to debug scientific thinking in a very direct way. This highlights an important difference between models’ specifications (here in mathematics) and their implementations, something that if ignored can introduce “bugs” in our scientific thinking. If we do not make explicit our thinking through formal modeling, and if we do not bother to execute, i.e., implement and run our formal(isable) specification, we can have massive inconsistencies in our understanding of our own model(s). We call this issue the “pizza problem.” Such misunderstandings - where a lot of the concepts are agreed upon and yet the models being run in people’s heads are dramatically different - are not at all rare. Another infamous online case is that of a person asking a body-building forum if it is “safe to do a full body workout every other day”[3]. Almost all 128 replies are dedicated to discussing what “every other day” means (3.5 or 4 days per week) even though everybody knows how many days a week has.

Let us go back to ordering the most pizza for our buck. And let us create a computational model of ordering pizza — overkill for scientific purposes, but certainly not for pedagogical ones. The verbal description of the problem (recall Figure 1) is that we need to pick an order option out of: one 18” pizza or two 12” pizzas. The formalised specification of our model is the next step and we choose to do this using mathematics. For any model, simplifications need to be made, so we choose to represent each individual pizza as a

[3] Archived webpage: archive.ph/9qQyT

circle. Therefore we define the amount of food $\phi$ per order option $i$ as:

$$\phi_i = N_i \pi R_j^2 \tag{1}$$

where $i$ is the pizza order option, $N$ is the number of pizzas in the order, and the rest is derived from the definition of the area of a circle. We also add in a decision rule that tells us what to order as a function of the food per order:

$$\omega(\phi_i, \phi_j) = \begin{cases} i, & \text{if } \phi_i > \phi_j \\ j, & \text{otherwise} \end{cases} \tag{2}$$

where the output of the $\omega$ function is the order index with the most food. Ta-da! We have just built a very basic mathematical model for deciding between the two order options — but we are not in the clear yet!

So far this is the model (or something very similar) that when explicitly asked everybody would have claimed to be running in their heads, but they still were surprised — an expectation violation occurred — when faced with the actual results: one 18" pizza is more food than two 12" pizzas. So how do we ensure we are all running the same model? We execute it on a computer that is not the human mind!

To make this model computational all we need is to program it using, e.g., Python. To be clear, we can run this model on paper. We can plug in the numbers as seen at the bottom of Figure 1, however for pedagogical purposes and because of course it is not always feasible to run models on paper, we will proceed to the final step computational modelers take: coding it up. So, using our favourite code editor, we start to create an implementation for the equations above. We notice that even though Equation 1 (a part of the specification) is not wrong per se, $\phi$ could be defined as:

$$\phi_i = \sum_{j \in N_i} \pi R_j^2 \tag{3}$$

with $N$ meaning exactly the same as before, the number of pizzas in the order. This change to the definition of food per order allows generalisation of the model (both in the specification and the implementation below) to account for different radii per order (i.e., in future we can compare an 11" pizza plus a 13" pizza with one 18" pizza). One possible implementation[4] of our pizza model could look like this:

```
import numpy as np
import math

def food(ds):
    '''
    Amount of food in an order as a function
    of the diameters per pizza (eq. 3).
    '''
    return (math.pi * (ds/2)**2).sum()
```

```
# Order option a in fig. 1, two 12'' pizzas:
two_pizzas = np.array([12, 12])

# Option b, one 18'' pizza:
one_pizza = np.array([18])

# Decision rule (eq. 2):
print(food(two_pizzas) > food(one_pizza))
```

However, it is extremely important to point out that this implementation change, which we choose to percolate upwards and thus edit our specification, does not affect the verbal description of the model. By the same token, a change in the code to use a for-loop in the definition of the `food()` function would neither affect the specification nor the theory in this specific case. This is a core concept to grasp when modeling: the properties of the relationships between theory/verbal description, specification, and implementation.

Is this whole exercise overkill for ordering pizza? Absolutely. But it serves as a valuable tool for showing somebody who might never have modeled, or never have thought deeply about the differences between verbal description, specification, and implementation, that they can be and should be dissociated. Despite the simplicity of our pizza model, it will likely fail to capture what each person actually wants: what if you value crust more, or consider two pizzas easier to share with your friend, or believe bigger pizzas are more likely to be damaged during transport, etc.? Every single modeling decision so far could have been made slightly (or dramatically) differently. This is one of the great things about modeling: it allows for full transparency. If one disagrees with any of the formalisms, they can easily plug in a different decision rule or a different definition of the amount of food or even a different aspect of the order being evaluated — perhaps they prefer more crust than overall pizza surface.

Formal modeling the way we have described above, and moreover, especially computational modeling, is quintessentially open science: verbal descriptions of science, specifications and implementations of models are totally transparent, open to be replicated, and open to be modified, i.e., open theory. Computational modeling is a step towards full open theorising to go along with open data, open source, etc. In contrast to merely stating "two 12" pizzas are more food than one 18"pizza", a computational model can be generalised and can show our work clearly. Through writing code, we can debug our scientific thinking.

In this section, we presented the, pedagogically valuable, process of creating a computational model. In the rest of the manuscript, we will present a bird's eye view of how research in our field is carried out in order to demonstrate where and how modeling fits in.

[4] Link to repo so others can use it easily.

## Model of psychological science

> [T]heory takes us beyond measurement in a way which cannot be foretold *a priori*, and it does so by means of the so-called intellectual experiments which render us largely independent of the defects of the actual instruments. (p. 27 Planck, 1936)

In this section, we describe an analytical view of psychological research, shown in Figure 2. Although other such models exist that aim to capture some aspect of the process of psychology (e.g., Haig, 2018; Haslbeck, Ryan, Robinaugh, Waldorp, & Borsboom, 2019; Kellen, 2019), ours proposes a unified account that demonstrates how computational modeling can play a radical and central role in all of psychological research. We propose that every scientific output in psychology can be analysed using the levels shown in the left column of Figure 2. The core of our claim is that scientific inquiry can be understood as a function from theory to data, and back to itself again, and this function must pass through several states in order to have explanatory force. The function can express a meaningful mapping, transformation, or update between a theory at time $t$ and that theory at time $t+1$ as it passes through specification and implementation, which ideally enforce a degree of formalisation. We note that each level (in blue) can, but does not have to, involve the construction of a (computational) model for that level, with examples of models shown in the left column (in green) connected by a dotted line to the level with which they are associated. If a given level is not well understood, making a model of that level can help elucidate the implicit assumptions therein and uncover so-called pizza problems.

A process function or *path function* is a function where the output or returned value of the function is dependent on the path, or the nature of transformations the input undergoes to become output. Path functions are used in thermodynamics to describe heat and work transfer; an intuitive example is distance to a destination being dependent on the road you take. Say you live in a country with decent and reliable public transport, and you can choose whether to take the train or drive a car from city A to city B. The distance on the freeway between A and B is 100 km, but the distance between A and B via the train system is 150 km. The time it takes to reach you destination is dependent on the path you take on your journey.

The path function moves from top to bottom in terms of dependencies, but the connections between each level and those below or above are bidirectional (represented by the large blue and small black arrows) capture the potential for adding or removing, loosening or tightening, of constraints that one level can impose on those above or below it. The nature of the connections from one layer to another takes on many scientifically pertinent forms which will be described in the following subsections. Our model, constrains the directionality of transitions in the following way: *a*) at any point transitions moving upwards are permissable and *b*) moving downwards is only possible if an expectation violation is resolved by first moving upwards. Transitions, when moving downwards, from one layer to the next can be thought of as functions where the input is the current layer and the output is the next layer. Transitions going upwards, are more complex and involve adjusting, e.g., a theory given some data, and can involve adjustments of many levels along the way to obtain the required theory-level update. Downwards motion is not allowed if a violation occurs, e.g., our model at the current step is not inline with our expectations. Once this violation is resolved by moving to any step above, we may move downwards respecting the serial ordering of the levels. For example, when the data does not confirm the hypothesis, we must move upwards and understand why and what needs to be amended in the levels above the hypothesis. Attempting to "fix" things at the hypothesis level is hypothesising after results known (HARKing, Kerr, 1998): scientific dishonesty.

We do not believe that every psychological study must contain models explicitly, but we propose that at least implicitly every research output is model- and theory-laden (i.e., carries with it theoretical and modeling commitments). And in addition, we believe that by making these implicit models explicit via computational modeling (writing code-based implementations) that the quality, usefulness, and truthiness of research programmes can be secured and ascertained. The three levels with a red background, theory, specification, and implementation, are those which we believe are left implicit to a greater or lesser extent in most of psychological research, especially parts of our field that have been most seriously affected by the so-called replication "crisis". This tendency to ignore the levels in red is a result of the same process by which theory and hypothesis are also conflated (Fried, 2020; Meehl, 1967; Morey, Homer, & Proulx, 2018), and by which models of the data are taken to be models of the theory: "theoretical amnesia" (Borsboom, 2013). In the rest of this section, we will first discuss what each level encompasses, defining how we intend these words to be used in the context of Figure 2, and then discuss the properties of the path function in general.

### Framework

A framework is a conceptual system of building blocks for creating facsimiles of complex psychological systems, see topmost level of Figure 2. A framework is typically described using natural language and figures, but can also be implemented in code like ACT-R (Anderson & Lebiere, 1998) and Soar (Newell, 1992). Some frameworks appear superficially simple or narrow, like the concept of working memory (Baddeley, 2010) or dual-systems approaches (Dayan & Berridge, 2014; Kahneman, 2011), while others

can be all-encompassing such as unified theories of cognition (Newell, 1990) or connectionism (McClelland, Rumelhart, & the PDP Research Group, 1986).

In the simplest case a framework is the context, the interpretation of the terms of a theory (Lakatos, 1976). Many framework-level ideas usually require descending further down the path before they can be computationally modeled (Hunt & Luce, 1992; Vere, 1992). While it is possible to avoid frameworks, it is "awkward and unduly laborious" (Suppes, 1967, p. 58) to work without one and thus depend on the next level down in the path to do all the heavy lifting.

It is not the case that all psychological models are or can be evaluated against data directly. For example, ACT-R is certainly not evaluatable using data directly, we have to move down the path first, thus creating a specific theory, then a specification, then an implementation, and then generate hypotheses, before any data can be collected. Even then perhaps ACT-R (due to its distance in the path from data) is in some sense unfalsifiable, but the path provides a clear pathway to evaluating modeling accounts within this framework (e.g., see: Cooper, 2007).

### Theory

A theory is a scientific proposition — described by a collection of natural language sentences, mathematics, logic, and figures — that (implicitly or explicitly) introduces causal relations with the aim of describing, explaining, and/or predicting a set of phenomena (Lakatos, 1976), see second level of Figure 2. Examples of psychological theories are prospect theory (Kahneman & Tversky, 1979), classical conditioning (Pavlov, 2010), and SUSTAIN, an account of categorisation (Love, Medin, & Gureckis, 2004).

To move to the next level and produce a specification for a psychological theory, we must be able to posit a plausible mechanism for the specification model to define based on the theory. As can be seen from our path direct comparisons to data can only happen once a model is at the right level. However, it is not the case that all psychological models must be evaluated against data directly. Theoretical computational models allow us to check if our ideas when taken to their logical conclusions hold up and also help us generate more theoretical knowledge (e.g., Guest & Love, 2017; Martin, 2016, 2020; Van Rooij, 2008). If a theory is scientifically stunted and thus cannot lead to coherent specifications or implementations, it is our responsibility as scientists to amend or in rare cases abandon it in favour of one that does.

### Specification

A specification is a formal(isable) description of a system to be implemented based on a theory, see third level of Figure 2. It provides a means of discriminating between theory-relevant, closer to the core claims of the theory, and theory-irrelevant, auxiliary assumptions (Cooper & Guest, 2014;

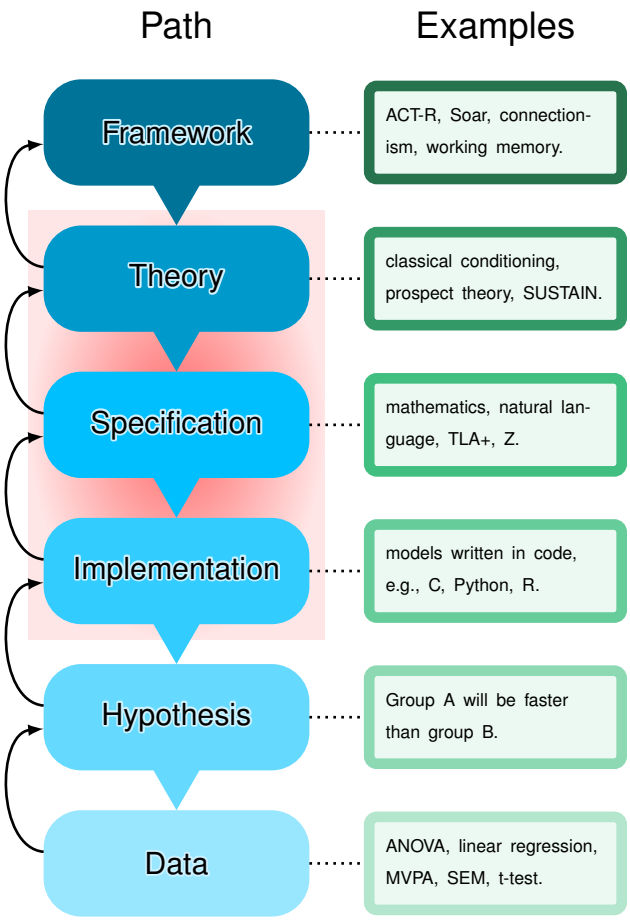


*Figure 2*. One of many possible paths (in blue) that can be used to understand and model how psychological research is carried out with examples at each step shown on the left (in green). Each research output within psychology can be described with respect to the levels in this path. The three levels with a red background (theory, specification, implementation) are those that are most often ignored or left out from research descriptions.

Lakatos, 1976). Specifications provide both a way to check if a computational model encapsulates the theory and a way to create a model if the theory is not clear enough by constraining the space of possible computational models. Specifications can be expressed in natural language sentences, mathematics, logic, flowcharts and other diagrams, and formal specification languages, such as Z notation (Spivey & Abrial, 1992) and TLA+ (Lamport, 2015) used in computer science.

The transition to code from specification has been automated in some cases in computer science (Monperrus, Jézéquel, Champeau, & Hoeltzener, 2008). In psychology,

creating an implementation typically involves taking the specification implicitly embedded in a journal article and writing code that is faithful to it. Specifications are invaluable because when the time comes to debug our implementation, it will be impossible to do so without a specification — and thus by extension impossible to properly test our theory (Cooper & Guest, 2014; Miłkowski, Hensel, & Hohol, 2018).

## Implementation

An implementation is an instantiation of a model created using anything from physical materials, e.g., a scale model of an airplane (Morgan & Morrison, 1999), to software, e.g., a git repository, see fourth level of Figure 2. A computational implementation is a codebase, a collection of programming code written in one or more languages that constitutes a software unit and embodies a computational model. While the concept of an implementation is simple to grasp — perhaps what most psychologists think of when they hear "model" — it might appear to be the hardest step in a research programme. This is arguably not the case. Provided one follows the steps in Figure 2, a large proportion of the heavy lifting is being done by all the previous steps and enables the part that requires coding to be equally if not less difficult than, e.g., theory creation or data collection.

In some senses, implementations are the most disposable and the most time-dependant parts of the scientific process of Figure 2. Very few programming languages stay in vogue for more than a decade and thus even though the raw text files of the code itself might survive bit-rot (digital entropy) and other problems set in, rendering code older than even a few months in extreme cases un-runnable without amendments (Cooper & Guest, 2014; Rougier et al., 2017). This is not entirely damaging to our enterprise since the core components of the science we want to evaluate are the theory and specification. If the computational model is not truly replicable, i.e., re-implementable given the specification, then it poses serious questions for the theory and specification it is based on (Cooper & Guest, 2014). This constitutes an expectation violation and must be addressed by moving upwards to whichever previous level we believe can amend the issue. However, we would be premature to generalise from the success or failure of one implementation if it cannot be recreated based the specification, since we have no reason to assume it is embodying the theory. This latter point of code appropriately embodying a theory can only be answered by iterating through theory, specification, implementation.

When we run our computational model's code, we can start to generate hypotheses. For example, if our model behaves in a certain way in a given task, e.g., it has trouble categorising some types of visual stimuli more than others, we can formulate a hypothesis to test if this holds in the empirical world. Alternatively, if we already know this phenomenon happens, it is a useful way to check that our high-level understanding does indeed so far match our observations. If our implementation displays behaviour outside what is permitted by the specification and theory, then we need to adjust something as this constitutes a violation. It might be that the theory is under-specified and this behaviour should not be permissible ever. In which case we might need to change both the specification and the implementation to match the theory (Cooper & Guest, 2014).

On the other hand, if our implementation displays behaviour outside what is known about the world, then we also need to adjust something. It could be that the theory instantiated by the implementation is again too loose and allows, e.g., behaviours that are not found in empirical experiments. Alternatively, it could be that we need to go and collect data to see if such behaviours are seen in the "real world" — thus this is not an expectation violation (yet) but a prediction we wish to test. Such a cycle of adjustments until the theory is captured by the code and the code is a strict subset of the theory are necessary parts of the scientific process. This cycle of loosening and tightening theory, specification, and implementation never ends — it is the sine qua non of scientific computational modeling and mutatis mutandis theory development in science.

## Hypothesis

A hypothesis is a narrow testable statement, see fifth level of Figure 2. Hypotheses in psychology focus on a set of properties of the world that can be measured and evaluated by collecting data and running inferential statistics. A hypothesis is scientifically valuable when embedded in the theoretical context from which it was derived. Any sentence that can be directly translated into a statistical test can be called a hypothesis, e.g., "the gender similarities hypothesis which states that most psychological gender differences are in the close to zero ($d \leq 0.10$) or small ($0.11 < d < 0.35$) range" (p. 581 Hyde, 2005). Hypothesis in psychology are directly amenable to data collection usually through highly-controlled lab- or web-based experiments, in part because they are formulated by scientists who are aware of the next step descending the path.

Hypothesis testing is unbounded without iterating through theory, specification, implementation and creating a computational model. The supervening levels constrain the space of possible hypotheses to-be-tested. Testing hypotheses in an ad hoc way — what we could dub *hypo-hacking* within out model — is to the hypothesis layer what *p*-hacking is to the data layer (Head, Holman, Lanfear, Kahn, & Jennions, 2015). Researchers can come up with any hypothesis and given big enough data a significant result is likely to be found when comparing, e.g., two theoretically-baseless groupings. Another way to hypo-hack is to atheoretically run pilot studies until something "works". When research is carried out this way "losing" the significant *p*-value, e.g., due to a fail-

ure to replicate, could be enough to destroy the research programme. Any theories built upon such hypo-hacked results will crumble if no bidirectional transitions in the path were carried out, especially within the redzone. Having built a computational account researchers can avoid the confirmation bias of hypo-hacking. Hypo-hacking, cheats the path and skips levels. While building a theory using a computational modeling approach, even if on data that includes some hypo-hacking and *p*-hacking means once a phenomenon is seen we ascend the path and spend time formalising a model (e.g., see: Fried, 2020; Head et al., 2015).

### Data

Data are observations generated by and collected from the "real world" or from a computational model, see sixth level of Figure 2. Data can take on many forms in psychology, the most common being numerical values that represent variables as defined by our experimental design, e.g., reaction times, questionnaire responses, neuroimaging, etc. Because of how theory-laden data is it can never be completely free from the theoretical assumptions implicit in its collection (Lakatos, 1976). For example, functional magnetic resonance imaging (fMRI) data rests on belief in the theory of electromagnetism, and in the theory of the BOLD signal's association with neural activation, etc. If any of these scientific theories that support the current interpretation of fMRI data change then the properties of the data collected will also change. In addition, data only has meaning as understood through the lens of the experiment (which is a product of theory) during which it was collected (Feyerabend, 1957).

Statistical models are the kinds of models most psychological scientists have been exposed to — every student who has been through a research methods class in a psychology undergraduate degree knows some basic statistical modeling techniques. Tests such as analysis of variance (ANOVA), mixed-effects modeling (e.g., Davidson & Martin, 2013), linear regression, multivariate pattern analysis (MVPA), structural equation modeling (SEM), and the t-test are all possible inferential statistical models of datasets.

If the data model does not support the hypothesis (an expectation violation), this allows us with a certain confidence to reject the hypothesis for our data. This does not however given us licence to reject a theory with as much confidence. The same caution is advised in the inverse situation when our statistical testing supports our experimental hypothesis: "there is subtle tendency to "carry over" a very small probability of a Type I error into a sizeable resulting confidence in the truth of the substantive theory" (Meehl, 1967, p. 107). For example, there have been a large number of studies that collected data on cognitive training over the past century and yet it is still not accepted as a scientific fact that it works (Katz, Shah, & Meyer, 2018). To escape these problems and understand how data and hypothesis relate to the theory we need to ascend the path and contextualise our data and hypotheses given our theory using our computational model. These violations cannot be addressed by plucking a new hypothesis out of thin air that conveniently fits our data, i.e., HARKing, but by moving back to theory and asking what needs to change in our theoretical understanding in order to explain the current results.

### What our path function model offers

We have clearly denoted the boundaries, which are often in reality not so clear-cut, between the levels of understanding in psychological research. Many of us do not explicitly think or work in way that facilitates separating our research into these chunks, thus it is often difficult to tease these layers of description apart. Notwithstanding, modelers often do this by definition — many should be familiar with similar layers of abstraction from computer science and levels of analysis from Marr and Poggio (1976). Simpler more abstract descriptions appear higher up, while more complex descriptions of psychological science are lower down the path — e.g., data is a much less "compressed" as a description of an experiment than a hypothesis. As such, each level is a renormalisation, coarser description, of the level below (DeDeo, 2018; Flack, 2012; Martin, 2020). The higher levels contain fewer exemplars than the lower levels. In this sense, moving through the path of scientific inference can be seen as a form of dimensionality reduction, or of coordinate transform. Not only are there fewer theories than datasets in practice (arguably causing chaos, Forscher, 1963), but also the principle of multiple realisability (Putnam, 1967) means that for every theory there are infinitely many possible implementations consistent with it and datasets that can be collected to test it (Blokpoel, 2018). This helps contextualise studies that show the divergence in data modeling decisions given the same hypotheses (e.g., Botvinik-Nezer et al., 2019; Silberzahn et al., 2018).

Figure 2 allows us to discuss and decide where in the path claims about psychological science are being made, thus contextualising descriptions of research and of science in general. For example, the claim that "[s]cience is posthoc, with results, especially unexpected results, driving theory and new applications" (Shiffrin, 2018) is not incompatible, if understood in context, with guarding against hypothesising after results known (HARKing, Kerr, 1998). The reason "science is post hoc with respect to the data" (to paraphrase Shiffrin, 2018) is because arguably one cannot have a theoretical account of a phenomenon without having access to some data, anecdotal, observational, and/or experimental, that guides one to notice said phenomenon in the first instance. Abraham Wald, for example, explained post hoc why the bullet holes found in fighter planes that returned home were correlated to their survival — this is not HARKing. Wald moved upwards from the data (distribution of bullet holes) to a the-

ory (survivor bias) and created a model at the theory level that could explain and predict the patterns of the bullets in the planes that made it back safely (Mangel & Samaniego, 1984). In other words, in many cases theory development involves some analysis (formal or informal) at the data level, as an inspiration or impetus, and then a lot of scientific activity within the levels we highlight in red in Figure 2: theory, specification, and implementation.

On the other hand, our path function model allows us to pinpoint on which level questionable research practises (QRPs, e.g., see: John, Loewenstein, & Prelec, 2012) or scientific misconduct are taking place in a specific instance and contextualise why and how to avoid them. Different QPRs occur at different levels, e.g., *p*-hacking at the data level, HARKing at the hypothesis level, and so on. HARKing is flawed because it does not resolve violations that occur when the data meets the hypothesis — it is not, e.g., TARKing (theorising after results known) which is part of the scientific practice of creating modeling accounts, as mentioned by Shiffrin (2018). To retrofit a hypothesis onto a dataset does not constitute resolving a violation because this de novo hypothesis not generated directly or indirectly by a theory. If we start out with a hypothesis (generated by going through the levels of theory, specification, implementation) and collect data that rejects our hypothesis, the violation has not only occurred at the hypothesis level since the hypothesis has been generated (via the intervening levels) by the theory. If this occurs, we need to move back up the levels to understand where the violation needs to be addressed. This is essentially the opposite to conjuring a new hypothesis (HARKing) that only exists in the scientific literature because it has been confirmed by data — data that was collected to test a completely different hypothesis. Importantly, it is at the data and hypothesis levels that preregistration and similar methods (Nosek et al., 2015) attempt to constrain science to avoid, e.g., HARKing. In our model, to ensure scientific quality researchers must ascend the path, instead of or in addition to, preregistration and other constraints on data modeling.

In terms of understanding the role of modeling explicitly: if we can move from a theory to a computational model, we are on the right track. Thus models can be seen as acting as mediators between theory and data (e.g., see: Morgan & Morrison, 1999; Oberauer & Lewandowsky, 2019). Asking if we can build a model of our theory allows us to understand where our theoretical understanding is lacking. Importantly, claims are typically not falsifiable — not usually directly testable at the framework or theory level — but become more so as we move downwards. This is why we need models, to shine a light on how to move downwards and what to test. By going through these motions we can understand the difference between auxiliary assumptions and core assumptions both in general but specifically for a given research programme as well (Lakatos, 1976). We thus iterate through theory, specification, and implementation as required until we have achieved a modeling account that satisfies all the various constraints using empirical data as well as collecting empirical data based on hypotheses generated from the computational model. Is an implementation detail in fact pivotal to a model working? Then it must be upgraded to a specification detail (Cooper & Guest, 2014). Mutatis mutandis for details at the specification level, etc. Importantly, this process is even useful in the case of "false" models, i.e., computational accounts that we do not agree with but can still improve our conceptual understanding of phenomena (e.g., Wimsatt, 2002; Winsberg, 2006).

On the other hand, research programmes which are light on modeling do not have a clear grasp on what is going in the area highlighted in red of Figure 2. These areas of psychology might have many, often informal, theories, but this is not enough (Watts, 2017). More data alone, however open, will never solve the issue of a lack of formal theorising. Data cannot tell a scientific story, that role falls to theory and theory needs formalisation to be evaluated, which can only be accomplished through computational modeling. Thus, while modelers are using the full scale of the path (often explicitly), reaping the benefits of formally testing and continuously improving their theories, those who eschew modeling miss out on these fundamental scientific insights. By formalising a research programme, we can search and evaluate the space of the account proposed in a meticulous way, i.e., "theory-guided scientific exploration" (Navarro, 2019, p. 31). We may locate missing parts of the research programme, its strengths, weaknesses, etc., and address them. To hearken back to the pizza example, non-modelers will be ignorant of pizza problems in their understanding and will potentially order two pizzas without realising they might be implicitly running a different model (in their head) to what they specify they will run. Scientific enterprises are vulnerable when they have not engaged in transparent formal open theory that is afforded by computational modeling, thus scrutinising their research questions and their understanding thereof.

## Discussion

We hope to spark dialogue on the radical role of computational modeling can play within open psychological science in forcing open theorising. We also presented a case study in building a basic computational model, providing a useful guide to those who may not have undergone such a process previously. Models, especially when formalised and run on a digital computer, can shine a light on when our scientific expectations are violated. To wit, we presented a high-level model of how science is carried out as a path function and radically centering computational modeling at the core of psychology. Computational models cannot replace, e.g., data or verbal theories, but that the process of creating a computational account is invaluable and informative.

There are three routes that psychology can take as a field, mirroring what Allen Newell said half a century ago in 1973: *a*) the field might bifurcate along the lines we have proposed herein between research programmes that use modeling and those that do not; *b*) the field might unite in so much as research programmes will contain some modeling to force the creation, refinement, and rejection of theories; and *c*) we carry on by asking questions that are not secured to a sound theoretical mooring via computational modeling. These are not completely mutually exclusive possibilities and some components from each of them can be seen in the present.

For *a*) bifurcation of the field, theoreticians, scientists who mostly inhabit the red area of Figure 2, will be free to practice modeling, e.g., without having to run frequentist statistics on their models if it is not appropriate. Much as is done in physics, no constraints will be put on individual scientists to pick a side, e.g., Einstein was both a theoretical and an experimental physicist. The importance is that the distinction will be highlighted and the scientific methods used will be slightly different. Unlike in the present in psychology, it will be easy to publish a paper with, e.g., only modeling at the theory level with no direct reference to data (something rare currently, although possible, e.g., Guest & Love, 2017).

In the case of *b*), mass cooperation to work on "larger experimental wholes" (Newell, 1973, p. 24), this is something perhaps realistic given projects that involve many labs have become commonplace (e.g., Botvinik-Nezer et al., 2019; Silberzahn et al., 2018). Although, we advise cautious optimism since these collaborations are only operating at the data and hypothesis levels at the moment, which are not enough to force theory building. Notwithstanding, such efforts might constitute the first step in understanding the logistics of multi-lab projects that aim to answer theoretical questions using computational modeling. On the other hand, as Richard Shiffrin mentioned in his talk (2018), computational and mathematical modelers often work on a series of related experiments, construct modeling accounts of the phenomena being studied, and publish this as one "larger experimental whole".

The third possibility — more of the same — is the most dire: "Maybe we should all simply continue playing our collective game of 20 questions. Maybe all is well [...] and when we arrive in 1992 [...] we will have homed in to the essential structure of the mind." (Newell, 1973, p. 24) The future, in such a case, holds more time-wasting and crises if we do not change. Some scientists will spend time (re)testing atheoretical hypotheses by e.g., replicating experiments. In reality such atheoretical work, regardless of replicability, could never have entered the literature had due process as outlined in Figure 2 been followed. Asking nature 20 questions without a computational model leads to serious theoretical issues even if results superficially are deemed replicable (e.g., see: Devezer, Nardin, Baumgaertner, & Buzbas, 2019; Katz et al., 2018).

## A way forward

We can only change if we all accept Figure 2 is how other sciences work, albeit in some cases implicitly, and radically update how we view the place of modeling in psychology. The first step is introspective: realising that we all already do some modeling even if we are not aware of it. We all use modeling, perhaps not formal, to some extent since we all ascribe to frameworks and theories even implicitly. Without formalising our assumptions, in the same way we explicitly state the variables in traditional hypothesis testing, we can never actually communicate efficiently. Thankfully, some have started to demand for this kind of shift in our thinking (e.g., Morey et al., 2018; Oberauer & Lewandowsky, 2019; Szollosi et al., 2019; Wills, O'Connell, Edmunds, & Inkster, 2017).

The second step is pedagogical: to teach mentees using open materials that this is neither extremely complex nor requires any extra skills over those we already ask them to master: programming, experimental design, literature review, statistical analyses techniques (e.g., Wills et al., 2017; Wilson & Collins, 2019). Modeling is a combination of the above mixed with a conscious understanding of what the scientist is subscribing to: the theory. Implementing assumptions about which psychological theory one agrees with and is currently working within is a natural conclusion of years of research and dedication to one's field.

The third step is cooperative: we need to work together as a field to insert modeling into more of our scientific endeavours. For those who believe the replication crisis is a measure of the scientific quality of a field and given that it has mostly affected areas of psychology with less formal modeling, it might be time to ask these areas explicitly to do modeling. And by extension for modelers to publish more in these areas — something which lately has been slowly happening, due in part to the rise of data science as a field (e.g., consumer behaviour: Hornsby, Evans, Riefer, Prior, & Love, 2019; Riefer, Prior, Blair, Pavey, & Love, 2017).

In order to ensure experiments result in data that can be recollected with similar effects we must force theory building, because replicability in part depends causally on things higher up the path (also see: Oberauer & Lewandowsky, 2019). Data that cannot be recollected and experiments that cannot be replicated are important issues. However, the same is true for theoretical accounts that cannot be instantiated as code. In the same way that questions such as "should results of preregistered studies count as stronger evidence than results of not preregistered studies?" questions like "should results of computationally modeled studies count as stronger evidence than those of studies with only a statistical model?" should also be actively discussed by the whole field (e.g., see: Szollosi et al., 2019).

Thus, while it may superficially appear that we are at odds with the great emphasis placed on the bottom few steps in the path by those who are investigating replicability, we are comfortable with this emphasis. We believe the proposals set out by some to automate or streamline the last few steps (hypothesis testing and data analysis) are part of the same solution (e.g., Lakens & DeBruine, 2020; Poldrack et al., 2019). We imagine a "best of all possible" massively collaborative future where scientists allow machines to carry out the least creative steps, and thus, set themselves free to focus wholly on computational modeling and theory generation.

## References


Anderson, J., & Lebiere, C. (1998). The atomic components of thought lawrence erlbaum. *Mathway, NJ*.

Baddeley, A. (2010). Working memory. *Current biology*, *20*(4), R136–R140.

Blokpoel, M. (2018). Sculpting computational-level models. *Topics in cognitive science*, *10*(3), 641–648.

Borsboom, D. (2013). *Theoretical amnesia.* `http://osc.centerforopenscience.org/2013/11/20/theoretical-amnesia/`. (Accessed: 2020-02-16)

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., ... Schonberg, T. (2019). Variability in the analysis of a single neuroimaging dataset by many teams. *bioRxiv*. Retrieved from `https://www.biorxiv.org/content/early/2019/11/15/843193`

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

Cohen, J. (1954). On the project of a universal character. *Mind*, *63*(249), 49–63.

Cooper, R. P. (2007). The role of falsification in the development of cognitive architectures: Insights from a lakatosian analysis. *Cognitive science*, *31*(3), 509–533.

Cooper, R. P., & Guest, O. (2014). Implementations are not specifications: Specification, replication and experimentation in computational cognitive modeling. *Cognitive Systems Research*, *27*, 42–49.

Davidson, D., & Martin, A. E. (2013). Modeling accuracy as a function of response time with the generalized linear mixed effects model. *Acta psychologica*, *144*(1), 83–96.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free pavlovian reward learning: revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(2), 473–492.

DeDeo, S. (2018). Origin gaps and the eternal sunshine of the second-order pendulum. In *Wandering towards a goal* (pp. 41–61). Springer.

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PloS one*, *14*(5).

Feyerabend, P. K. (1957). An attempt at a realistic interpretation of experience. In *Proceedings of the aristotelian society* (Vol. 58, pp. 143–170).

Flack, J. C. (2012). Multiple time-scales and the developmental dynamics of social systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1597), 1802–1810.

Forscher, B. K. (1963). Chaos in the brickyard. *Science*, *142*(3590), 339.

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature.

Guest, O., & Love, B. C. (2017). What the success of brain imaging implies about the neural code. *Elife*, *6*, e21397.

Haig, B. D. (2018). An abductive theory of scientific method. In *Method matters in psychology* (pp. 35–64). Springer.

Haslbeck, J., Ryan, O., Robinaugh, D., Waldorp, L., & Borsboom, D. (2019). Modeling psychopathology: From data models to formal theories. *PsyArXiv*.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, *13*(3).

Hornsby, A. N., Evans, T., Riefer, P. S., Prior, R., & Love, B. C. (2019). Conceptual organization is revealed by consumer activity patterns. *Computational Brain & Behavior*, 1–12.

Hunt, E., & Luce, R. D. (1992). Soar as a world view, not a theory. *Behavioral and brain sciences*, *15*(3), 447–448.

Hyde, J. S. (2005). The gender similarities hypothesis. *American psychologist*, *60*(6), 581.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5), 524–532.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 263–292.

Katz, B., Shah, P., & Meyer, D. E. (2018). How to play 20 questions with nature and lose: Reflections on 100 years of brain-training research. *Proceedings of the National Academy of Sciences*, *115*(40), 9897–9904.

Kellen, D. (2019). A model hierarchy for psychological science. *Computational Brain & Behavior*, *2*(3-4), 160–165.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In *Can theories be refuted?* (pp. 205–259). Springer.

Lakens, D., & DeBruine, L. (2020). Improving transparency, falsifiability, and rigour by making hypothesis tests machine readable. *PsyArXiv*.

Lamport, L. (2015). The tla+ hyperbook. *Dostopno na: http://research. microsoft. com/enus/um/people/lamport/tla/hyperbook. html [31. 8. 2015]*.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological review*, *111*(2), 309.

Mangel, M., & Samaniego, F. J. (1984). Abraham wald's work on aircraft survivability. *Journal of the American Statistical Association*, *79*(386), 259–267.

Marr, D., & Poggio, T. (1976). *From understanding computation to understanding neural circuitry*.

Martin, A. E. (2016). Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in psychology*, *7*, 120.

Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience*.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group. (1986). Parallel distributed processing. *Explorations in the Microstructure of Cognition*, *2*, 216–271.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of science*, *34*(2), 103–115.

Miłkowski, M., Hensel, W. M., & Hohol, M. (2018). Replicability or reproducibility? on the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of computational neuroscience*, *45*(3), 163–172.

Monperrus, M., Jézéquel, J.-M., Champeau, J., & Hoeltzener, B. (2008). A model-driven measurement approach. In *International conference on model driven engineering languages and systems* (pp. 505–519).

Morey, R. D., Homer, S., & Proulx, T. (2018). Beyond statistics: accepting the null hypothesis in mature sciences. *Advances in Methods and Practices in Psychological Science*, *1*(2), 245–258.

Morgan, M. S., & Morrison, M. (1999). *Models as mediators*. Cambridge University Press Cambridge.

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.

Newell, A. (1973). *You can't play 20 questions with nature and win: Projective comments on the papers of this symposium.*

Newell, A. (1990). *Unified theories of cognition, harvarduniv*.

Newell, A. (1992). Soar as a unified theory of cognition: Issues and explanations. *Behavioral and Brain Sciences*, *15*(3), 464–492.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... others (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic bulletin & review*, *26*(5), 1596–1618.

Orwell, G. (1945). *Animal farm*.

Pavlov, P. I. (2010). Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. *Annals of neurosciences*, *17*(3), 136.

Planck, M. (1936). *The philosophy of physics*. W. W. Northon & Company Inc.

Poldrack, R. A., Feingold, F., Frank, M. J., Gleeson, P., de Hollander, G., Huys, Q. J., ... others (2019). The importance of standards for sharing of computational models and data. *Computational Brain & Behavior*, *2*(3-4), 229–232.

Putnam, H. (1967). Psychological predicates. *Art, mind, and religion*, *1*, 37–48.

Riefer, P. S., Prior, R., Blair, N., Pavey, G., & Love, B. C. (2017). Coherency-maximizing exploration in the supermarket. *Nature human behaviour*, *1*(1), 1–4.

Rougier, N. P., Hinsen, K., Alexandre, F., Arildsen, T., Barba, L. A., Benureau, F. C., ... others (2017). Sustainable computational science: the rescience initiative. *PeerJ Computer Science*, *3*, e142.

Shiffrin, R. M. (2018). Science should govern the practice of statistics. *Symposium: Should Science Determine the Practice of Statistics or Should Statistics Determine the Practice of Science? at the Annual Psychonomics Meeting 2018*.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... others (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356.

Spivey, J. M., & Abrial, J. R. (1992). *The z notation*. Prentice Hall Hemel Hempstead.

Suppes, P. (1967). What is a scientitic theory? *Philosophy of Science Today*, 55–67.

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R. M., van Rooij, I., Van Zandt, T., & Donkin, C. (2019). Is preregistration worthwhile? *Trends in Cognitive Sciences*.

Van Rooij, I. (2008). The tractable cognition thesis. *Cognitive science*, *32*(6), 939–984.

Vere, S. A. (1992). A cognitive process shell. *Behavioral and Brain Sciences*, *15*(3), 460–461.

Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, *1*(1), 0015.

Wiener, P. P. (1951). *Leibniz: selections* (Vol. 1). Scribner Book Company.

Wills, A. J., O'Connell, G., Edmunds, C. E., & Inkster, A. B. (2017). Progress in modeling through distributed collaboration: Concepts, tools and category-learning examples. In *Psychology of learning and motivation* (Vol. 66, pp. 79–115). Elsevier.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547.

Wimsatt, W. C. (2002). Using false models to elaborate constraints on processes: Blending inheritance in organic and cultural evolution. *Philosophy of Science*, *69*(S3), S12–S24.

Winsberg, E. (2006). Models of success versus the success of models: Reliability without truth. *Synthese*, *152*(1), 1–19.