aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2021, Vol. 16(4) 789–802

DOI: 10.1177/1745691620970585
www.psychologicalscience.org/PPS
SAGE

How Computational Modeling Can Force Theory Building in Psychological Science

Olivia Guest[1,2,3] and Andrea E. Martin[1,4]
[1]Donders Centre for Cognitive Neuroimaging, Radboud University; [2]Research Centre on Interactive Media, Smart Systems and Emerging Technologies (RISE), Nicosia, Cyprus; [3]Department of Experimental Psychology, University College London; and [4]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Abstract
Psychology endeavors to develop theories of human capacities and behaviors on the basis of a variety of methodologies and dependent measures. We argue that one of the most divisive factors in psychological science is whether researchers choose to use computational modeling of theories (over and above data) during the scientific-inference process. Modeling is undervalued yet holds promise for advancing psychological science. The inherent demands of computational modeling guide us toward better science by forcing us to conceptually analyze, specify, and formalize intuitions that otherwise remain unexamined—what we dub *open theory*. Constraining our inference process through modeling enables us to build explanatory and predictive theories. Here, we present scientific inference in psychology as a path function in which each step shapes the next. Computational modeling can constrain these steps, thus advancing scientific inference over and above the stewardship of experimental practice (e.g., preregistration). If psychology continues to eschew computational modeling, we predict more replicability crises and persistent failure at coherent theory building. This is because without formal modeling we lack open and transparent theorizing. We also explain how to formalize, specify, and implement a computational model, emphasizing that the advantages of modeling can be achieved by anyone with benefit to all.

Keywords
computational model, theoretical psychology, open science, scientific inference

Psychology is a science that attempts to explain human capacities and behaviors. This results in a wide range of research practices, from conducting behavioral and neuroscientific experiments to clinical and qualitative work. Psychology intersects with many other fields, creating interdisciplinary subfields across science, technology, engineering, mathematics, and the humanities. Here we focus on a distinction within psychological science that is underdiscussed: the difference in explanatory force between research programs that use formal, mathematical, and/or computational modeling and those that do not, or, more specifically, programs that explicitly state and define their models and those that do not.

We start by explaining what a computational model is, how it is built, and how formalization is required at various steps along the way. 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 an example in which intuition is insufficient in determining a quantity. Next, we present our path model of how psychological science should be conducted to maximize the relationship between theory, specification, and data. 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* that must step through theory, specification, and implementation before an interpretation can have explanatory force in relation to a theory. Our path-function model also enables us to evaluate claims about the process of doing psychological and cognitive science itself,

Corresponding Author:
Olivia Guest, Donders Centre for Cognitive Neuroimaging, Radboud University
E-mail: olivia.guest@ru.nl

pinpointing where in the path questionable ways of conducting research occur, such as *p*-hacking (biasing data analysis or collection to force statistical modeling to return significant *p* values; e.g., Head et al., 2015). Finally, we believe psychological science needs to use modeling to address the structural problems in theory building that underlie the so-called replication crisis in, for example, social psychology (see Flis, 2019). 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

Psychological scientists typically ascribe to a school of thought that specifies a framework, a theoretical position, or at least some basic hypotheses that they then set out to test using inferential statistics (Meehl, 1967; Newell, 1973). Almost every article in psychological science can be boiled down to introduction, methods, analysis, results, and discussion. The way we approach science is nearly 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 in the majority of psychology's scientific endeavors. Most psychological researchers are not trained in modeling beyond constructing statistical models of their data, which are typically applicable off the shelf.

In contrast, 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; they create semiformalized or formalized versions of scientific theories, often creating (or least amending) their accounts along the way. Computational modelers are researchers who have 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.

Involving modeling in a research program 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 be able to generate and test hypotheses using empirical data. Computational modeling does this by forcing scientists to explicitly document 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.

What Is a Computational Model? And Why Build One?

> Let us calculate, without further ado, and see who is right.
>
> —Gottfried Leibniz (Wiener, 1951)

Gottfried Leibniz 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; Cohen, 1954; Wiener, 1951). Computational modeling is the process by which a verbal description is formalized to remove ambiguity, as Leibniz correctly predicted, while also constraining the dimensions 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, "How do we understand brain and behavior in this context, and why?" By thinking through how to represent the data and model the experiment, scientists gain insight into the computational repercussions of their ideas in a much deeper and explicit way than by just collecting data. 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 theorizing, in other words explicitly stating and formalizing our theoretical commitments, is done by default as a function of the process.

Through modeling, even in, or especially in, failures we hone our ideas: Can our theory be formally specified, and if not, why not? Thus, we may check whether 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 those who may wish to apply these ideas to their work outside science (e.g., in industrial or clinical settings).

One of the core properties of models is that they allow us to "safely remove a theory from the brain of its author" (A. J. Wills, personal communication, May 19, 2020; see also Wills et al., 2017; Wills & Pothos,

2012), thus allowing the ideas in one's head to run on other computers. Modeling also allows us to compare models based on one theory with those based on another and compare different parameter values' effects within a model, including damaging models in ways that would be unethical in human participants (e.g., "lesioning" artificial neural network models; see Guest et al., 2020). One of the only situations in which multiple theories can be distinguished in a formal setting is when they can make sense of the available data (e.g., Levering et al., 2019; see also, however, Cox & Shiffrin, in press; Navarro, 2019; Wills & Pothos, 2012).

We now walk the reader through building a computational model from scratch to illustrate our argument and then present a path function of research in psychological science. 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.

The pizza problem

> All models are wrong but some are more wrong than others.
>
> —based on Box (1976) and Orwell (1945)

Imagine it is Friday night, and your favorite pizzeria has a special: two 12-in. pizzas for the price of one 18-in. pizza. Your definition of a good deal is one in which you purchase the most food. Is this a good deal?

A Twitter post (Fermat's Library, 2019) said "a useful counterintuitive fact: one 18 inch pizza has more 'pizza' than two 12 inch pizzas"—along with an image similar to Figure 1. The reaction to this tweet was largely surprise or disbelief. For example, one follower replied, "But two pizzas are more than one" (Sykes, 2019). Why were people taken aback?

When it comes to comparing the two options in Figure 1, although we all agree on how the area of a circle is defined, the results of the "true" model, that one 18-in. pizza has more surface and therefore is more food, are counterintuitive. Computational modeling can demonstrate how one cannot always trust one's gut. To start, one must create (a) a verbal description, a conceptual analysis, and/or a theory; (b) a formal (or formalizable) description, that is, a specification using mathematics, pseudocode, flowcharts, and so on; and (c) an executable implementation written in programming code (for an overview of these steps, see Fig. 2, red area). This process is the cornerstone of computational modeling and by extension of modern scientific thought, enabling us to refine our gut instincts through experience.

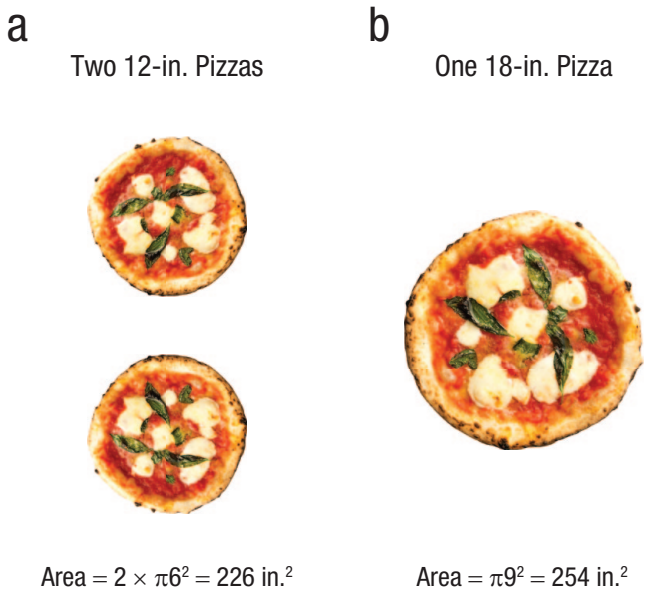

Fig. 1. The pizza problem. Something like comparing the two options presented here can appear counterintuitive, although we all learn the formula for the area of a circle in primary school. Compare (a) two 12-in. pizzas and (b) one 18-in. pizza (all three pizzas are to scale). Which order would you prefer?

Experience is seeing our ideas being executed by a computer, giving us the chance to debug scientific thinking in a very direct way. If we do not make our thinking explicit through formal modeling, and if we do not bother to execute (i.e., implement and run our specification through computational modeling), we can have massive inconsistencies in our understanding of our own model(s). We call this issue "the pizza problem."

Herein we model the most pizza for our buck—overkill for scientific purposes but certainly not for pedagogical ones. For any formalized specification, including that for the pizza orders shown in Figure 1, simplifications need to be made, so we choose to represent pizzas as circles. Therefore, we define the amount of food ϕ per order option i as

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

where i is the pizza-order option, N is the number of pizzas in the order, and the rest is the area of a circle. We also propose a pairwise decision rule:

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

where the output of the ω function is the order with the most food. This is the model that everyone would have claimed to have running in their heads, but they

still were surprised—an expectation violation occurred—when faced with the actual results. 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, we move from specification to implementation (consider where we are in the path shown in Fig. 2). We notice Equation 1 is not wrong, but could be defined more usefully as

$$\phi_i = \sum_{j=1}^{N} \pi R_j^2, \tag{3}$$

where j is the current pizza, allowing us to sum over all pizzas N within food order i.

This change allows generalization of the model (both in the specification above and the implementation below) to account for different radii per order (i.e., in the future we can compare an 11-in. pizza plus a 13-in. pizza with one 18-in. pizza). One possible implementation (in Python) of our pizza model looks 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))
```

Note that this implementation change, which we choose to percolate upward, editing 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 affect neither the specification nor the theory in this case. This is a core concept to grasp: the relationships between theory, specification, and implementation—consider our movements up and down the path as depicted in Figure 2.

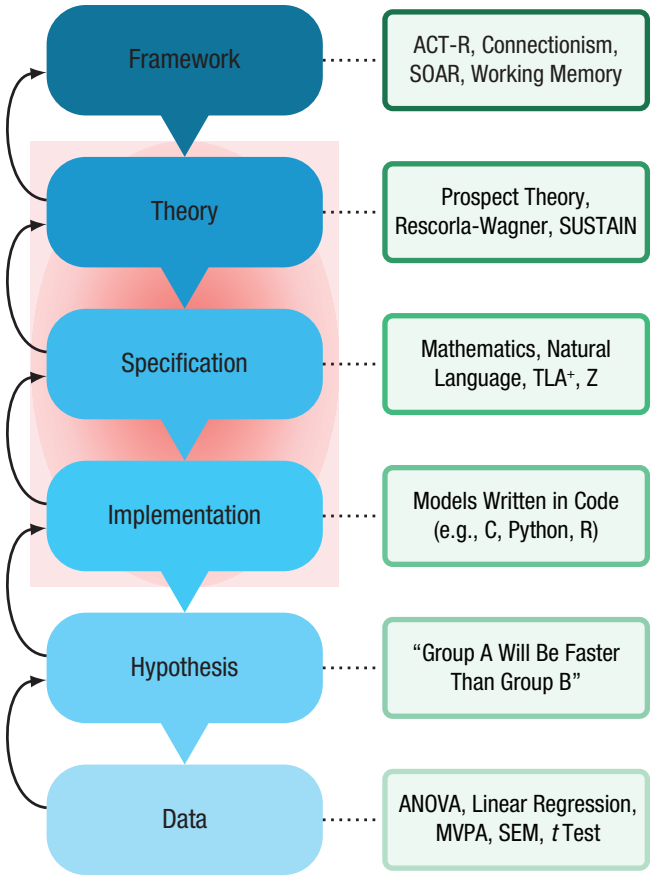

Fig. 2. One of many possible paths (in blue) that can be used to understand and describe how psychological research is carried out, with examples of models at each step shown on the left (in green). Each research output within psychological science can be described with respect to the levels in this path. The three levels superimposed on a red background (theory, specification, implementation) are those that are most often ignored or left out from research descriptions. ACT-R = adaptive control of thought–rational; SUSTAIN = supervised and unsupervised stratified adaptive incremental network; ANOVA = analysis of variance; MVPA = multivariate pattern analysis; SEM = structural equation modeling.

Computational modeling, when carried out the way we describe herein, is quintessentially open science: Verbal descriptions of science, specifications, and implementations of models are transparent and thus open to replication and modification. If one disagrees with any of the formalisms, they can plug in another decision rule or definition of the amount of food or even another aspect of the order being evaluated (e.g., perhaps they prefer more crust than overall surface). Computational modeling—when done the way we describe, which requires the creation of specifications and implementations—affords open theorizing to go along with open data, open-source code, and so on. In contrast to merely stating two 12-in. pizzas offer more food than one 18-in. pizza, a computational model can

be generalized and can show our work clearly. Through writing code, we debug our scientific thinking.

Model of Psychological Science

> Theory 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.
>
> —Planck (1936, p. 27)

In this section, we describe an analytical view of psychological research, shown in Figure 2. Although other such models exist for capturing some aspect of the process of psychological science (e.g., Haig, 2018; Haslbeck et al., 2019; Kellen, 2019; van Rooij & Baggio, 2021), our model proposes a unified account that demonstrates how computational modeling can play a radical and central role in all of psychological research.

We propose that scientific outputs can be analyzed using the levels shown in the left column of Figure 2. Scientific inquiry can be understood as a function from theory to data and back again, and this function must pass through several states to gain 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 enforces a degree of formalization. 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 their associated level. If a level is not well understood, making a model of that level helps to elucidate implicit assumptions, addressing "pizza problems."

A path function is a function in which the output depends on a path of transformations the input undergoes. Path functions are used in thermodynamics to describe heat and work transfer; an intuitive example is distance to a destination being dependent on the route and mode of transport taken. The path function moves from top to bottom in terms of dependencies, but the connections between each level and those adjacent are bidirectional (represented by large blue and small black arrows). Connections capture the adding or removing, loosening or tightening, of constraints that one level can impose on those above or below it.

In our model depicted in Figure 2, the directionality of transitions is constrained only when moving downward. Thus, at any point transitions moving upward are permissible, whereas moving downward is possible only if an expectation violation is resolved by first moving upward. Downward transitions can be thought of as functions in which the input is the current layer and the output is the next. Upward transitions are more complex and involve adjusting (e.g., a theory given some data) and can involve changes to levels along the way to obtain the required theory-level update. With respect to why we might want to move upward out of choice and recalling the case of the pizza model above, we updated the specification (changing Equation 1 to Equation 3) because we thought about the code/implementation more deeply and decided it is worth updating our formal specification (Equation 3). Downward motion is not allowed if a violation occurs (e.g., our model at the current step is not in line with our expectations). Once this violation is resolved by moving to any step above, we may move downward, respecting the serial ordering of the levels. For example, when the data do not confirm the hypothesis, we must move upward and understand why and what needs to be amended in the levels above the hypothesis. Attempting to "fix" things at the hypothesis level is known as hypothesizing after the results are known (or HARKing; Kerr, 1998). In the case of the pizza model, an expectation violation occurs when we realize that the one pizza is more food. At that point, we reevaluate our unspecified/implicit model and move back up to the appropriate level to create a more sensible account.

At least implicitly, every scientific output is model- and theory-laden (i.e., contains theoretical and modeling commitments). By making these implicit models explicit via computational modeling the quality, usefulness, and verisimilitude of research programs can be secured and ascertained. The three levels with a red background in Figure 2 (theory, specification, and implementation) are those that we believe are left implicit in most of psychological research—this is especially so in parts of psychological science that have been most seriously affected by the so-called replication crisis. This tendency to ignore these levels results from the same process by which theory and hypothesis are conflated (Fried, 2020; Meehl, 1967; Morey et al., 2018) and by which models of the data are taken to be models of the theory: "theoretical amnesia" (Borsboom, 2013). When models of the data are seen as models of the theory, potentially bizarre situations can arise—eventually forcing dramatic rethinkings of (sub)fields (e.g., Jones & Love, 2011).

Framework

A framework is a conceptual system of building blocks for creating facsimiles of complex psychological systems (see Fig. 2, topmost level). A framework is typically

described using natural language and figures but can also be implemented in code such as ACT-R (or adaptive control of thought–rational; Anderson & Lebiere, 1998) and SOAR (Newell, 1992). Some frameworks appear superficially simple or narrow, such as the concept of working memory (Baddeley, 2010) or dual-systems approaches (Dayan & Berridge, 2014; Kahneman, 2011), whereas others can be all-encompassing such as unified theories of cognition (Newell, 1990) or connectionism (McClelland et al., 1986).

In the simplest case a framework is the context—the interpretation of the terms of a theory (Lakatos, 1976). Frameworks usually require descending further down the path before they can be computationally modeled (Hunt & Luce, 1992; Vere, 1992). Although it is possible to avoid explicit frameworks, it is "awkward and unduly laborious" (Suppes, 1967, p. 58) to work without one and thus depends 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 such a model: We have to descend the path first, creating a specific theory, then a specification, then an implementation, and then generate hypotheses before any data can be collected (see Cooper, 2007; Cooper et al., 1996).

Theory

A theory is a scientific proposition—described by a collection of natural-language sentences, mathematics, logic, and figures—that introduces causal relations with the aim of describing, explaining, and/or predicting a set of phenomena (Lakatos, 1976; see Fig. 2, second level). Examples of psychological theories are prospect theory (Kahneman & Tversky, 1979), the Rescorla-Wagner model for Pavlovian conditioning (Rescorla & Wagner, 1972), and SUSTAIN (supervised and unsupervised stratified adaptive incremental network), an account of categorization (Love et al., 2004).

To move to the next level and produce a specification for a psychological theory, we must posit a plausible mechanism for the specification model to define. As can be seen from our path, direct comparisons to data can happen only once a model is at the correct level. However, not all psychological models must be (or can be) evaluated against data directly. Theoretical computational models allow us to check whether our ideas, when taken to their logical conclusions, hold up (e.g., Guest & Love, 2017; Martin, 2016, 2020; Martin & Baggio, 2020; van Rooij, 2008). If a theory cannot lead to coherent specifications, it is our responsibility as scientists to amend or, more rarely, abandon it in favor of one that does.

Specification

A specification is a formal (or formalizable) description of a system to be implemented on the basis of a theory (Fig. 2, third level). It provides a means of discriminating between theory-relevant (i.e., closer to the core claims of the theory) and theory-irrelevant auxiliary assumptions (Cooper & Guest, 2014; Lakatos, 1976). Specifications provide both a way to check whether a computational model encapsulates the theory and a way to create a model even if the theory is not clear enough, simply by constraining the space of possible computational models. Specifications can be expressed in natural-language sentences, mathematics, logic, diagrams, and formal specification languages, such as Z notation (Spivey & Abrial, 1992) and TLA+ (Lamport, 2015).

The transition to code from specification has been automated in some cases in computer science (Monperrus et al., 2008). In psychological science, creating an implementation typically involves taking the specification implicitly embedded in a journal article and writing code that is faithful to it. Without specifications, we can neither debug our implementations nor properly test our theories (Cooper et al., 1996; Cooper & Guest, 2014; Miłkowski et al., 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; Fig. 2, fourth level). A computational implementation is a codebase written in one or more programming languages that constitutes a software unit and embodies a computational model. Although the concept of an implementation is simple to grasp—perhaps what most psychologists think when they hear the term model—it might appear to be the hardest step. This is arguably not the case. Provided one follows the steps in Figure 2, a large proportion of the heavy lifting is done by all the previous steps.

In some senses, implementations are the most disposable and time-dependent parts of the scientific process illustrated in Figure 2. Very few programming languages stay in vogue for more than a decade, rendering code older than even a few months in extreme cases unrunnable without amendments (Cooper & Guest, 2014; Rougier et al., 2017). This is not entirely damaging to our enterprise because the core scientific components we want to evaluate are theory and specification. If the computational model is not reimplementable given the specification, it poses serious questions for the theory (Cooper & Guest, 2014). This constitutes an expectation violation and must be

addressed by moving upward to whichever previous level can amend the issue. However, it is premature to generalize from the success or failure of one implementation if it cannot be recreated according to the specification because we have no reason to assume it is embodying the theory. Whether code appropriately embodies a theory can be answered only by iterating through theory, specification, and implementation.

Running our computational model's code allows us to generate hypotheses. For example, if our model behaves in a certain way in a given task (e.g., has trouble categorizing some types of visual stimuli more than others), we can formulate a hypothesis to test this behavior. Alternatively, if we already know this phenomenon occurs, computational modeling is a useful way to check that our high-level understanding does indeed so far match our observations. If our implementation displays behavior outside what is permitted by the specification and theory, then we need to adjust something because this constitutes a violation. It might be that the theory is underspecified and that this behavior should not be permissible, in which case we might need to change both the specification and the implementation to match the theory (Cooper & Guest, 2014). 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. Loosening and tightening theory, specification, and implementation never ends—it is the essence of theory development in science.

Hypothesis

A hypothesis is a narrow testable statement (Fig. 2, fifth level). Hypotheses in psychological science focus on properties of the world that can be measured and evaluated by collecting data and running inferential statistics. Any sentence that can be directly tested statistically can be a hypothesis, for example, "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" (Hyde, 2005, p. 581).

Hypothesis testing is unbounded without iterating through theory, specification, and implementation and creating computational models. The supervening levels constrain the space of possible hypotheses to be tested. Testing hypotheses in an ad hoc way—what we could dub *hypohacking*—is to the hypothesis layer what *p*-hacking is to the data layer (Head et al., 2015). Researchers can concoct any hypothesis, and given sufficient data a significant result is likely to be found when comparing, for example, two theoretically baseless groupings. Another way to hypohack is to atheoretically run pilot studies until something works. When research is carried out this way, losing the significant *p* value (e.g., because of a failure to replicate) could be enough to destroy the research program. Any theories based on hypohacking will crumble if no bidirectional transitions in the path were carried out, especially within the steps highlighted in red in Figure 2. Having built a computational account researchers can avoid the confirmation bias of hypohacking, which cheats the path and skips levels.

Data

Data are observations collected from the real world or from a computational model (Fig. 2, sixth level). Data can take on many forms in psychological science, the most common being numerical values that represent variables as defined by our experimental design (e.g., reaction times, questionnaire responses, neuroimaging). Most undergraduate psychology students know some basic statistical modeling techniques. Tests such as analysis of variance, linear regression, multivariate pattern analysis, structural equation modeling, the *t* test, and mixed-effects modeling (e.g., Davidson & Martin, 2013) are all possible inferential statistical models of data sets.

Because data are theory-laden, they can never be free from, or understood outside of, the assumptions implicit in their collection (Feyerabend, 1957; Lakatos, 1976). For example, functional MRI (fMRI) data rest on our understanding of electromagnetism and of the blood-oxygen-level-dependent signal's association with neural activation. If any of the scientific theories that support the current interpretation of fMRI data change then the properties of the data will also change.

If the data model does not support the hypothesis (an expectation violation), we can reject the experimental hypothesis with a certain confidence. However, we cannot reject a theory with as much confidence. The same caution is advised in the inverse situation (Meehl, 1967). For example, a large number of studies have collected data on cognitive training over the past century, and yet consensus on its efficacy is absent (Katz et al., 2018). To escape these problems and understand how data and hypothesis relate to our working theory we must ascend the path and contextualize our findings using computational modeling. These violations cannot be addressed by inventing new hypotheses that conveniently fit our data (i.e., HARKing) but by asking what needs to change in our theoretical understanding.

Harking back to pizza

The pizza example (purposefully chosen in part because it is simple and devoid of psychological constructs, which bias reader's opinions toward one formalism over

another) can be decomposed readily into the six levels shown in Figure 2. At the framework level we have the concepts of pizza, food, and order because we want to compare the total amount of food per order. These are the building blocks for any account that involves deciding between orders of food made up of pizzas, even if we disagree on which aspects of the order (e.g., money, speed of delivery), food (e.g., calories, ingredients), or pizza (e.g., crust, transportability) we will eventually formally model and empirically test.

Then at the theory level, there are essentially two theories: the original (implicit) theory T_0 that the number of pizzas per order corresponds to the amount of food in that order and the post hoc corrected (explicit) T_0 that the surface areas of the pizzas per order correspond to the amount of food in that order. To get to T_1, we created a specification, created an implementation, and refined the specification—we discuss exactly how this happened using the path model of Figure 2.

Before obtaining T_1, we descended the path by going from basically framework to hypothesis (bypassing the red area completely; recall T_0 was not explicitly stated at all, let alone formalized, at the beginning) to generate the very clear prediction (and thus testable hypothesis) that the order with two pizzas is more food than the order with one. Because we skipped the parts of the path that required formalizing our ideas (shown in red in Fig. 2), we are faced with an expectation violation. We believed that two 12-in. pizzas are more food than one 18-in. pizza (recall Fig. 1), and we also believed that the food per order is a function of the surface area of the pizzas. Therefore, we realize our own ideas about the amount of food per order are incompatible with themselves (what we dub the pizza problem), as well as what we know about the world from other sources (imagine if we had weighed the pizzas per order, for example). Had this been a real research program (and not a fictional example), we would have descended all the way and collected empirical data on the pizzas by, for example, weighing them. This act of collecting observations would have further solidified the existence of an expectation violation because the two pizzas would have been found to, for example, have less mass, thus falsifying both our hypothesis and indirectly T_0.

At the point of an expectation violation, we decided to address the steps we skipped in the red area, so we moved upward to create a formal specification S_0 embodied by Equations 1 and 2. We then attempted to descend from S_0 to create an implementation I_0, which led to refining our specification, thus creating S_1 (Equations 2 and 3). We are now fully in the throes of formal and computational modeling by cycling through the steps shown in Figure 2 in red.

Arguably—and this is one of the core points of this article—had we not ignored the steps in red and created a theory, specification, and implementation explicitly, we would have been on better footing from the start. And so it is demonstrated that applying the path model adds information to the scientific-inference process. Still, we managed to document and update our less-than-useful assumptions by going back and formally and computationally capturing our ideas. We should all strive not to ignore these vital steps by directly focusing on them, either ourselves or by making sure the literature contains this explicit formal and computational legwork.

What our path-function model offers

We have denoted the boundaries and functions of levels within the scientific-inference process in psychological research—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 the path, whereas more complex descriptions of psychological science are lower down the path—for example, data are much less compressed as a description of an experiment than is a hypothesis. Each level is a renormalization, or coarser description, of those below (DeDeo, 2018; Flack, 2012; Martin, 2020). Higher levels contain fewer exemplars than lower levels. Moving through the path of scientific inference is a form of dimensionality reduction or coordinate transformation. Not only are there often no substantive nor formalized theories for some data sets in practice (causing chaos; see Forscher, 1963), but the principle of multiple realizability (Putnam, 1967) also implies that for every theory there are infinitely many possible implementations consistent with it and data sets that can be collected to test it (Blokpoel, 2018). This helps to contextualize studies that show divergence in data-modeling decisions given the same hypotheses (e.g., Botvinik-Nezer et al., 2020; Silberzahn et al., 2018).

Open theories (i.e., those developed explicitly, defined formally, and explored computationally, in line with Fig. 2) are more robust to failures of replication of any single study from which they might derive some support because of the specific way the path has been followed to develop and test them. For example, if the impetus or inspiration for theory development is a single study (that is thereafter found not to be replicable) because we then move to the red area, refine our ideas, and then drop back down to test them again, we will avoid dependence on a single (potentially problematic) study. Failures to replicate can not only be detected but also explained and perhaps even drive

theory creation as opposed to just theory rejection. Thus, building a theory explicitly as laid out in Figure 2, even if based on some hypo- and *p*-hacking, means once a phenomenon is detected we ascend the path and spend time formalizing our account (e.g., Fried, 2020). Our path model asks for formalization using specifications and implementations (or indeed anything more comprehensive than an individual study; see Head et al., 2015); thus, when our model is used, "sins" that are out of individual scientists' control—such as questionable research practices (QRPs; see John et al., 2012) committed by other labs or publication bias committed by the system as a whole—can be both discovered and controlled for in many cases.

Thinking about our science with reference to Figure 2 allows us to discuss and decide where in the path claims about science are being made (i.e., not only allowing us to evaluate claims about the phenomena being examined, modeled, and so on, but also to evaluate general claims about how we conduct research or about how not to conduct research). For example, the claim that "science is posthoc, with results, especially unexpected results, driving theory and new applications" (Shiffrin, 2018) is not incompatible with guarding against HARKing because one cannot have an account of a phenomenon without having access to some data—anecdotal, observational, and/or experimental—that guide one to notice the phenomenon in the first case.

Theories in psychology result from protracted thought about and experience with a human cognitive capacity. Scientists immerse themselves in deep thought about why and how their phenomena of study behave. This basic principle of developing theories is captured in the example of Wald's investigation into optimally (and thus minimally because of weight) armoring aircraft to ensure pilots returned safely during World War II (Mangel & Samaniego, 1984). Planes returned after engaging with the Nazis with a smattering of bullet holes that were distributed in a specific way: More holes were present in the fuselage than in the engines, for example. Wald explained post hoc why and how the holes were correlated to survival. Contrary to what one might expect, areas with the least holes would benefit from armor. Wald theorized that planes in general were likely hit by bullets uniformly, unlike the planes in the data set; aircraft hit in the engines did not make it home and so were not present in the data set; and, therefore, armor should be placed over the engines, the area with the fewest bullet holes. This is not HARKing—this is formal modeling. Wald moved upward from the data (distribution of bullet holes) to a theory (survivor bias) and created an explicit formal model that could explain and predict the patterns of the bullets in planes that made it back safely. In many cases theory development involves analysis at the data level, as an inspiration or impetus, and then a lot of scientific activity within the levels: theory, specification, and implementation. This is why we do not impose any constraints on moving upward in Figure 2, only on moving downward.

On the other hand, our path-function model allows us to pinpoint on which level QRPs are taking place and how to avoid them. Different QRPs occur at different levels, for example, *p*-hacking at the data level, HARKing at the hypothesis level, and so on. HARKing does not resolve expectation violations that occur when the data meet the hypothesis—it is not, for example, theorizing after the results are known, which is part of the scientific practice of creating modeling accounts. To retrofit a hypothesis onto a data set does not constitute resolving a violation because this de novo hypothesis is not generated directly or indirectly by a theory. If we start out with a hypothesis and collect data that reject our hypothesis, the violation has not occurred uniquely at the hypothesis level because the hypothesis has been generated (via the intervening levels) by the theory. This is essentially the opposite to conjuring a new hypothesis (HARKing) that exists only in the scientific literature because it has been "confirmed" by data—data collected to test a different hypothesis.

It is at the data and hypothesis levels that preregistration and similar methods attempt to constrain science to avoid QRPs (e.g., Flis, 2019; Szollosi et al., 2019). To ensure scientific quality, however, we propose that preregistration is not enough because it serves only to constrain the data and hypothesis spaces. Researchers who wish to develop their formal account of a capacity must ascend the path instead of, or in addition to, for example, preregistering analysis plans. Preregistration cannot on its own evaluate theories. We cannot coherently describe and thus cannot sensibly preregister what we do not yet (formally and computationally) understand. Indeed, theories can and should be computationally embodied and pitted against each other without gathering or analyzing any new data. To develop, evaluate, and stress-test theories, we need theory-level constraints on and discussions about our science. Figure 2 can serve as a first step in the right direction toward such an ideal.

By the same token, our path model allows us to delineate and discuss where computational modeling itself has been compromised by QRPs occurring at the specification and implementation levels. A typical case of this is when authors report only partial results of implementing a specification of their theory; for example, only some implementations show the required or predicted patterns of behavior (Guest et al., 2020). As

mentioned, the solution is to cycle within the red area of Figure 2 to ensure theory-, specification-, and implementation-level details are indeed assigned to and described at the correct level. Failing to do that, we propose, is a type of QRP.

Computational modeling can be seen as mediating between theory and data (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. Claims are typically not falsifiable—not usually directly testable at the framework or theory level—but become more so as we move downward. We thus iterate through theory, specification, and implementation as required until we have achieved a modeling account that satisfies all of the various constraints imposed by empirical data, as well as collecting empirical data based on hypotheses generated from the computational model. Is an implementation detail pivotal to a model working? Then it must be upgraded to a specification detail (Cooper et al., 1996; Cooper & Guest, 2014). Mutatis mutandis for details at the specification level and so on—meaning that details at every level can be upgraded (or downgraded) as required. This process is even useful in the case of "false" models; that is, computational accounts that we do not agree with can still improve our understanding of phenomena (e.g., Wimsatt, 2002; Winsberg, 2006).

As mentioned, cycling through the steps in Figure 2 shines a direct light on what our theoretical commitments are in deep ways. Mathematically specifying and/or computationally implementing models, for example, can demonstrate that accounts are identical or overlap even when their verbal descriptions (i.e., informal specifications) are seemingly divergent. This can result from (a) multiple theories being indeed more similar in essence than previously thought, paving the way for theoretical unification of a seemingly disjointed literature (e.g., Kelly et al., 2017), or (b) theories that are indeed different being less computationally explored and thus less constrained in their current state (e.g., Olsson et al., 2004). These kinds of discoveries about how we compartmentalize, understand, and predict human capacities are why iterating over—and thus refining—theory, specification, and implementation is vital.

Research programs light on modeling do not have a clear grasp on what is going on in the area highlighted in red in Figure 2. These areas of psychological science might have many, often informal, theories, but this is not enough (Watts, 2017). Neither is more data—however open, they will never solve the issue of a lack of formal theorizing. Data cannot tell a scientific story; that role falls to theory, and theory needs formalization to be evaluated. Thus, whereas modelers often use the full scale of the path, reaping the benefits of formally testing and continuously improving their theories, those who eschew modeling miss out on fundamental scientific insights. By formalizing a research program, we can search and evaluate in a meticulous way the space of the account proposed (i.e., "theory-guided scientific exploration"; Navarro, 2019, p. 31). As shown using the pizza example, nonmodelers remain unaware of pizza problems and may not realize they are implicitly running a different model (in their head) to what they specify.

Discussion

We hope to spark a dialogue on the radical role computational modeling can play by forcing open theorizing. We also presented a case study in building a basic computational model, providing a useful guide to those who may not have modeled before. Models, especially when formalized and run on a digital computer, can shine a light on when our scientific expectations are violated. To wit, we presented a path-function model of science, radically centering computational modeling at the core of psychological science. Computational models cannot replace, for example, data or verbal theories, but the process of creating a computational account is invaluable and informative.

There are three routes that psychology can take, mirroring Newell (1973): The first is bifurcating between research programs that use modeling and those that do not; the second is uniting research programs inasmuch as they contain some modeling to force the creation, refinement, and rejection of theories; and the third is continuing to ask questions that are not secured to a sound theoretical mooring via computational modeling. These are not completely mutually exclusive possibilities—some components from each can be seen in the present.

For bifurcation of the field, theoreticians, scientists who mostly inhabit the red area of Figure 2, will be free to practice modeling, for example, without having to run frequentist statistics on their models if inappropriate. No constraints will be put on individual scientists to pick a side (e.g., Einstein was active in theoretical and experimental physics). Unlike it is now, it will be easy to publish work containing only modeling at the theory level without direct reference to data (something rare currently, although possible; e.g., Guest & Love, 2017; Martin, 2016, 2020).

In the case of uniting research programs, mass cooperation to work on "larger experimental wholes" (Newell, 1973, p. 24) is perhaps realistic given projects that involve many labs are commonplace (e.g., Botvinik-Nezer et al., 2020; Silberzahn et al., 2018). We advise cautious optimism because these collaborations are

operating only at the data and hypothesis levels, which are insufficient to force theory building. Nevertheless, such efforts might constitute the first step in understanding the logistics of multilab projects. On the other hand, modelers often already currently work on a series of related experiments and publish them as a single experimental whole (Shiffrin, 2018).

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). Thus, the future holds more time wasting and crises. Some scientists will spend time attempting to replicate atheoretical hypotheses. However, asking nature 20 questions without a computational model leads to serious theoretical issues, even if the results are superficially deemed replicable (e.g., Devezer et al., 2019; Katz et al., 2018).

A Way Forward

Psychological science can change if we follow Figure 2 and radically update how we view the place of modeling. The first step is introspective: realizing that we all do some modeling—we subscribe to frameworks and theories implicitly. Without formalizing our assumptions in the same way we explicitly state the variables in hypothesis testing, we cannot communicate efficiently. Some have even started to demand this shift in our thinking (e.g., Morey et al., 2018; Oberauer & Lewandowsky, 2019; Szollosi et al., 2019; Wills et al., 2017).

The second step is pedagogical: explaining what modeling is and why it is useful. We must teach mentees that modeling is neither extremely complex nor requires extra skills beyond those we already expect that they master, for example, programming, experimental design, literature review, and statistical-analysis techniques (e.g., Epstein, 2008; Wills et al., 2017; Wilson & Collins, 2019).

The third step is cooperative: working together as a field to center modeling in our scientific endeavors. Some believe the replication crisis is a measure of the scientific quality of a subfield, and given that it has affected areas of psychological science with less formal modeling, one possibility might be to ask these areas to model explicitly. By extension, modelers can begin to publish more in these areas (e.g., in consumer behavior; see Hornsby et al., 2019).

To ensure experimental results can be replicated and reobserved, we must force theory building; replicability in part depends causally on things higher up the path (see also Oberauer & Lewandowsky, 2019). Data and experiments that cannot be replicated are clearly important issues. However, the same is true for theoretical accounts that cannot be instantiated or reinstantiated as code. Should results of preregistered studies count as stronger evidence than results of nonpreregistered studies? Should results of computationally modeled studies count as stronger evidence than those of studies with only a statistical model? Just as the first question here has been actively discussed (e.g., Szollosi et al., 2019), the second should be as well.

Thus, although it may superficially appear that we are at odds with the emphasis on the bottom few steps in our path model (hypothesis testing and data analysis; recall Fig. 2) 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 are part of the solution (e.g., Lakens & DeBruine, 2021; Poldrack et al., 2019). Such a division of labor might help to maximize the quality of theories and showcase the contrast—which Meehl (1967) and others have drawn attention to—between substantive theories and the hypotheses they generate. We imagine a "best of all possible" massively collaborative future in which scientists allow machines to carry out the least creative steps and thus set themselves free to focus exclusively on computational modeling, theory generation, and explanation.

Transparency

Action Editors: Travis Proulx and Richard Morey
Advisory Editor: Richard Lucas
Editor: Laura A. King
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Funding
O. 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. A. 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).

ORCID iD

Olivia Guest https://orcid.org/0000-0002-1891-0972

Acknowledgments

We thank Abeba Birhane, Sebastian Bobadilla Suarez, Christopher D. Chambers, Eiko Fried, Dermot Lynott, Esther Mondragón, Richard D. Morey, Karim N'Diaye, Nicolas P.

Rougier, Richard M. Shiffrin, Loukia Tzavella, and Andy J. Wills for useful discussions and input on previous versions of the article.

References

Anderson, J., & Lebiere, C. (1998). *The atomic components of thought*. Erlbaum.

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*. Open Science Collaboration. http://osc.centerforopenscience.org/2013/11/20/theoretical-amnesia

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., . . . Schonberg, T. (2020). Variability in the analysis of a single neuroimaging data set by many teams. *Nature*, *582*(7810), 84–88.

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., Fox, J., Farringdon, J., & Shallice, T. (1996). A systematic methodology for cognitive modelling. *Artificial Intelligence*, *85*(1–2), 3–44.

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

Cox, G. E., & Shiffrin, R. M. (in press). Computational models of event memory. In M. J. Kahana & A. Wagner (Eds.), *Oxford handbook of human memory*. Oxford University Press.

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 A. Aguirre, B. Foster, & Z. Merali (Eds.) *Wandering towards a goal: How can mindless mathematical laws give rise to aims and intention?* (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), Article e0216125. https://doi.org/10.1371/journal.pone.0216125

Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11*(4), Article 12. http://jasss.soc.surrey.ac.uk/11/4/12.html

Fermat's Library [@fermatslibrary]. (2019, January 7). *Here's a useful counterintuitive fact: One 18 inch pizza has more 'pizza' than two 12 inch pizzas* [Image attached] [Tweet]. Twitter. https://twitter.com/fermatslibrary/status/1082273172114862083

Feyerabend, P. K. (1957). An attempt at a realistic interpretation of experience. *Proceedings of the Aristotelian Society*, *58*, 143–170. https://www.jstor.org/stable/4544593

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.

Flis, I. (2019). Psychologists psychologizing scientific psychology: An epistemological reading of the replication crisis. *Theory & Psychology*, *29*(2), 158–181.

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*. PsyArXiv. https://psyarxiv.com/zg84s

Guest, O., Caso, A., & Cooper, R. P. (2020). On simulating neural damage in connectionist networks. *Computational Brain & Behavior*, *3*, 289–32. https://doi.org/10.1007/s42113-020-00081-z

Guest, O., & Love, B. C. (2017). What the success of brain imaging implies about the neural code. *eLife*, *6*, Article e21397. https://doi.org/10.7554/eLife.21397

Haig, B. D. (2018). An abductive theory of scientific method. In *Method Matters in Psychology. Studies: Essays in Applied Philosophy of Science* (pp. 35–64). Springer. https://doi.org/10.1007/978-3-030-01051-5_3

Haslbeck, J., Ryan, O., Robinaugh, D., Waldorp, L., & Borsboom, D. (2019). *Modeling psychopathology: From data models to formal theories*. PsyArXiv. https://doi.org/10.31234/osf.io/jgm7f

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), Article e1002106. https://doi.org/10.1371/journal.pbio.1002106

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*, *3*, 162–173. https://doi.org/10.1007/s42113-019-00064-9

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), 581–592.

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.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of bayesian models of cognition [Target article and commentaries]. *Behavioral and Brain Sciences*, *34*(4), 169–231. https://doi.org/10.1017/S0140525X10003134

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

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 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, USA*, *115*(40), 9897–9904.

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

Kelly, M. A., Mewhort, D. J., & West, R. L. (2017). The memory tesseract: Mathematical equivalence between composite and separate storage memory models. *Journal of Mathematical Psychology*, *77*, 142–155.

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 S. G. Harding (Ed.), *Can theories be refuted?* (pp. 205–259). Springer.

Lakens, D., & DeBruine, L. (2021). Improving transparency, falsifiability, and rigour by making hypothesis tests machine readable. *Advances in Methods and Practices in Pychological Science*, *4*(2). https://doi.org/10.1177/2515245920970949

Lamport, L. (2015). *The TLA+ hyperbook*. https://lamport.azurewebsites.net/tla/hyperbook.html

Levering, K. R., Conaway, N., & Kurtz, K. J. (2019). Revisiting the linear separability constraint: New implications for theories of human category learning. *Memory & Cognition*, *48*, 335–347. https://doi.org/10.3758/s13421-019-00972-y1–13

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332.

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, May). *From understanding computation to understanding neural circuitry* [AI Memo 357]. MIT Artificial Intelligence Laboratory. https://dspace.mit.edu/bitstream/handle/1721.1/5782/AIM-357.pdf

Martin, A. E. (2016). Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology*, *7*, Article 120. https://doi.org/10.3389/fpsyg.2016.00120. PMID: 26909051

Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience*, *32*(8), 1407–1427. https://doi.org/10.1162/jocn_a_01552

Martin, A. E., & Baggio, G. (2020). Modelling meaning composition from formalism to mechanism. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), Article 20190298. https://doi.org/10.1098/rstb.2019.0298

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 *ACM/IEEE 11th International conference on model driven engineering languages and systems* (pp. 505–519). Springer. https://doi.org/10.1007/978-3-540-87875-9_36

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. https://doi.org/10.1177/2515245918776023

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

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. In W. G. Chase (Ed.), *Visual information processing: Proceedings of the eighth annual Carnegie symposium on cognition, held at the Carnegie-Mellon University, Pittsburgh, Pennsylvania, May 19, 1972* (pp. 283–305). Academic Press. https://kilthub.cmu.edu/articles/journal_contribution/You_can_t_play_20_questions_with_nature_and_win_projective_comments_on_the_papers_of_this_symposium/6612977

Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.

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

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

Olsson, H., Wennerholm, P., & Lyxzèn, U. (2004). Exemplars, prototypes, and the flexibility of classification models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 936–994.

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

Planck, M. (1936). *The philosophy of physics*. W.W. Norton.

Poldrack, R. A., Feingold, F., Frank, M. J., Gleeson, P., de Hollander, G., Huys, Q. J. M., Love, B. C., Markiewicz, C. J., Moran, R., Ritter, P., Rogers, T. T., Turner, B. M., Yarkoni, T., Zhan, M., & Cohen, J. D. (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.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (Vol. 2, pp. 64–99). Appleton-Century-Crofts.

Rougier, N. P., Hinsen, K., Alexandre, F., Arildsen, T., Barba, L. A., Benureau, F. C.Y., Brown, C. T., de Buyl, P.,

Caglayan, O., Davison, A. P., Delsuc, M.-A., Detorakis, G., Diem, A. K., Drix, D., Enel, P., Girard, B., Guest, O., Hall, M. G., Henriques, R. N., . . . Zito, T. (2017). Sustainable computational science: The rescience initiative. *PeerJ Computer Science*, *3*, Article e142. https://doi.org/10.7717/peerj-cs.142

Shiffrin, R. M. (2018, November 16). Science should govern the practice of statistics. In R. M. Shiffrin (Chair), *Should statistics determine the practice of science, or science determine the practice of statistics?* [Symposium]. Annual Psychonomics Meeting, New Orleans, LA, United States.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., . . . Nosek, B. A. (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.

Suppes, P. (1967). What is a scientific theory? In S. Morgenbesser (Ed.), *Philosophy of science today* (pp. 55–67), Basic Books.

Sykes, M. [@MarkSykes15]. (2019, January 7). *But two pizzas are more than one* [Image attached] [Tweet]. Twitter. https://twitter.com/MarkSykes15/status/1082274473737359366

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*, *24*(2), 94–95.

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

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

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), Article 0015.

Wiener, P. P. (Ed. & Trans.). (1951). *Leibniz: Selections* (Vol. 1). Scribner Book.

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 B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 66, pp. 79–115). Elsevier.

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, *138*(1), 102–125.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, Article e49547. https://doi.org/10.7554/eLife.49547

Wimsatt, W. C. (2002). Using false models to elaborate constraints on processes: Blending inheritance in organic and cultural evolution. *Philosophy of Science*, *69*(Suppl. 3), S12–S24.

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