To cite this article:

Vanhasbroeck, N.*, Yu, K.*, & Ariens, S.* (in press). No need to be indirect: On the role of data in the validation of theory. *Theory & Psychology*.

No need to be indirect: On the role of data in the validation of theoretical models

Niels Vanhasbroeck[1], Kenny Yu[2], and Sigert Ariens[2]

[1]Psychological Methods, University of Amsterdam, Amsterdam, the Netherlands

[2]Quantitative Psychology and Inidivdual Differences, KU Leuven, Leuven, Belgium

Author Note

All authors contributed equally. Authors can be contacted through their respective email addresses: niels.vanhasbroeck@gmail.com, kenny.yu@kuleuven.be, sigert.ariens@kuleuven.be

Abstract

In recent years, the case has been made to move from narrative accounts of psychological theories to rigorous mathematical descriptions thereof. Yet, a persistent tension remains in how the models that result from the formalization of psychological theories are to be evaluated – whether by their theoretical elegance, their ability to reproduce established phenomena, or their quantitative fit to empirical data. This paper challenges an exclusive use of validation paradigms that are based on phenomena or data alone as evaluative unit by reexamining the relationships between theory, phenomena, and data. We argue that testing the tenability of a theory or model while placing too much emphasis on only phenomena or data is inherently fallible. Instead, we argue that both phenomena and data should occupy a central role in theory testing. This perspective demands that models make substantive and explanatory claims while engaging with the full structure of empirical observations. This perspective not only clarifies the dynamic interdependence among theory, phenomena, and data, but also establishes a principled basis for cumulative, iterative scientific progress in psychological science.

Keywords: computational modeling; theory evaluation; phenomena; simulation; data

No need to be indirect: On the role of data in the validation of theoretical models

Introduction

Theories are the engines of scientific progress; They offer an explanation for how a system behaves and help predict as of yet unknown behavior that may occur in the future (Popper, 1959). While psychological researchers have a history of using natural language to communicate their theories, they are increasingly turning towards the language of mathematics for this purpose (Borsboom & Haslbeck, 2024; Farrell & Lewandowsky, 2010; Smaldino, 2017). While not uncontroversial (Eronen & Romeijn, 2020; Oude Maatman, 2025), we align with scholars who view this evolution as advantageous in several respects. For example, the use of mathematics allows for a greater precision in articulating explanatory mechanisms and in generating testable predictions. This increased precision, together with the tools of mathematics, provides researchers with rigorous methods to evaluate the adequacy of their theoretical frameworks and derive novel predictions that might otherwise have remained obscure in purely verbal accounts.

As efforts to formalize theories continue, so too must our efforts for the development of methods for evaluating them. Specifically, we should ask what constitutes as evidence for or against a theory, and how we can decide whether a model genuinely captures the psychological processes it purports to explain. Fortunately, a vast literature on model evaluation procedures exists, providing researchers with various tools to evaluate a model's explanatory and predictive power (see, for example Navarro et al., 2004; D. R. Roberts et al., 2017; Schwarz, 1978; Tashman, 2000; Wagenmakers et al., 2004; Wehrens et al., 2000).

However, researchers increasingly express skepticism towards the practice of evaluating a model's fit to empirical data, with some scholars arguing that such an analysis, when conducted in isolation, yields little meaningful insight (S. Roberts & Pashler, 2000). Instead of investigating fit, we should investigate a theory's ability to make "risky predictions", that is predictions that would be unlikely to hold if the theory

were false, and use this as the primary basis of evaluating a theory's validity (Lakatos, 1978; Navarro, 2019; S. Roberts & Pashler, 2000). This perspective has been voiced persuasively by Borsboom and colleagues, who, across a series of recent publications, have argued that psychological phenomena provide a more robust foundation for the development and evaluation of theories (Borsboom & Haslbeck, 2024; Borsboom et al., 2021; Haslbeck et al., 2022; van Dongen et al., 2025). According to this framework, (formal) theories should address the mechanisms underlying one or more psychological phenomena, and the testing of such theories should consequently focus on the predictions of phenomena under a prespecified set of conditions.

Importantly, they contrast this phenomenon-based approach to the dominant data-driven approach in which a model's fit to data is taken as evidence for or against the theory that has been formalized to the model. The validity of the data-driven approach is called into question on grounds that data serve as an unreliable basis for testing theory (Borsboom & Haslbeck, 2024; van Dongen et al., 2025). Specifically, data are particular, meaning that model-based results that are obtained through direct engagement with data may not generalize across different studies or may not hold up against later scrutiny. This is contrasted with phenomena, which are in themselves generalized patterns that have been repeatedly observed across multiple studies. Given that theories explain phenomena and that phenomena are more stable entities than data, phenomena are argued to serve as a more stable basis for the validation of theory.

We wholeheartedly agree that a crucial aspect of building and reasoning about the validity of a theory concerns it's ability to explain phenomena. However, when focusing on phenomena, one may start to underappreciate the role empirical data play in *validating* theory. In this article, we therefore aim to clarify the role data can and should play when validating a formal theory.

In what follows, we will first discuss the data-driven and phenomenon-based approaches to theory validation and show where a pure application of either of these approaches may go wrong. We then place data alongside phenomena in an integrated approach to theory testing, highlighting the complementary role both approaches play

in evaluating theory. We end this paper with a short discussion on the integrated approach an a historical example.

In this article, we focus primarily on theories that are expressed as mathematical equations, guided by our shared expertise in statistics and mathematical psychology. We furthermore want to emphasize that we fundamentally agree with many aspects of the phenomenon-based approach outlined in previous literature. We endorse the construction of theory through phenomena and the testing of theory through "risky" predictions, positions advocated by numerous philosophers and psychologists (Haslbeck et al., 2022; Lakatos, 1978; Navarro, 2019). Our contribution therefore does not aim to discredit the consideration of phenomena, but rather aims to provide a crucial missing element to the ongoing discussion on the evaluation of theory.

Definitions

Before we present our argument, it is useful to define the primary concepts of interest, namely those of *data*, *phenomena*, *theory*, *model*, and *theory validation*.

Data

In this paper, we define data as a collection of observations that are systematically structured and recorded (a definition aligned with van Dongen et al., 2025). These observations are typically encoded as variables that serve as the basis for subsequent analysis. Examples include self-reported ratings (Cloos et al., 2023; Krosnick & Fabrigar, 1997; Russel et al., 1989), spatial position recordings (Bastida-Castillo et al., 2019; Gamble et al., 2023; Hedrick, 2008), and psychophysiological measurements (Dawson et al., 2000; Mauss & Robinson, 2009; Read, 2017).

A fundamental characteristic of data is its particularity – each data point represents an observation obtained from a specific individual, within a specific context, at a specific moment in time. In their raw form, data lack inherent structure or generalizability: The process of identifying regularities within data constitutes the discovery of *phenomena*. Given that theories aim to explain regularities rather than idiosyncrasies, this particularity presents a challenge for using raw data directly to validate formal theories. The gap between particular observations and generalizable

theoretical principles requires bridging mechanisms that can connect these distinct epistemological domains.

Phenomena

Given the inherent limitations of data in their particularity, researchers have advocated for phenomena as more appropriate targets for validating theory (Bogen & Woodward, 1988; Haig, 2005; Woodward, 1989). Phenomena are conceptualized as stable, recurring patterns that manifest consistently across diverse datasets and contexts. Examples of phenomena are the speed-accuracy tradeoff (Brown & Heathcote, 2008; Dutilh et al., 2011; Ratcliff & Rouder, 1998), lowered mood at the end of multi-trial experiments (Jangraw et al., 2023), and the power law for learning (Logan, 1992; Newell and Rosenbloom, 1980, but see Heathcote et al., 2000). Critically, phenomena exist at a higher level of abstraction than individual data points or specific datasets – they represent generalizable regularities rather than context-dependent particulars. This generalizability makes phenomena especially suitable targets for theoretical explanation, as theories themselves aim to capture general principles rather than idiosyncratic observations (van Dongen et al., 2025).

Theory

In this paper, we define theory as a systematic and explanatory account of how aspects of the world work (following van Dongen et al., 2025). As convincingly argued by several authors (Bogen & Woodward, 1988; Borsboom et al., 2021; Guest & Martin, 2021), theories should be constructed to explain phenomena, either by describing how a collection of phenomena interrelate or, preferably, by targeting the mechanisms that underlie them. These mechanisms often focus on causal relationships between variables which, together, account for the manifestation of the phenomena. Examples of such theories are the kinetic theory of gases (Pathria & Beale, 2022; Schroeder, 2021), the theory of evolution (Darwin, 1859), and cognitive dissonance theory (Festinger, 1957).

An important characteristic of theories is that they can be used to predict under which conditions a given phenomenon should occur, thereby providing an opportunity to test a theory's adequacy to explain the phenomenon. However, deriving precise

predictions from theories often presents significant challenges, particularly when theories are formulated with insufficient specificity (Oude Maatman, 2025). This limitation has led researchers to advocate for the development of *formal theories* or *models* (Borsboom et al., 2021; Guest & Martin, 2021; van Rooij & Baggio, 2021), which we define in the next section.

Model

For the purposes of this article, we distinguish between a theory and its formalization as a mathematical or computational *model*. While a theory requires the specification of all theoretically relevant mechanisms that may give rise to a particular set of phenomena, a formal model attempts to convey these mechanisms by translating them, as accurately as possible, into mathematical equations. Formalizing a theory into a model requires both making implicit theoretical assumptions explicit as well as the specification of auxiliary assumptions that may not be relevant for the theory itself (Smaldino, 2020; Suppes, 1962; van Dongen et al., 2025). Examples of such auxiliary assumptions concern assumptions of statistical independence or normality, which are typically imposed to allow for a model's estimation on data. Throughout this article, we use "model" to refer to what we previously called a "formal theory", therefore considering models to represent a theory that has been formalized into mathematical language.

Interestingly, the formalization of theory has been argued to greatly increase the precision with which the mechanisms that generate a particular set of phenomena are described (Borsboom & Haslbeck, 2024; van Dongen et al., 2025; van Rooij & Baggio, 2021). Importantly, this statement is not uncontroversial. For example, some authors have argued that psychology is not ready for formalization yet, first requiring us to precisely define our constructs before we can meaningfully formulate a theory, let alone formalize one (Eronen & Bringmann, 2021; Kellen et al., 2021). Additionally, the translation of a theory into a model does not imply an increased precision in its specification (Oude Maatman, 2025): If a theory lacks specificity, then a model based on this theory will likely suffer the same fate. We agree with both of these arguments,

but at the same time believe that the act of formalization forces researchers to think their theory through. The fact that, in formalization, researchers have to engage with all aspects of a theory, greatly constrains researcher degrees of freedom and will, ultimately, lead to more rigorous and transparent scientific discourse.

Theory validation

Theory validation concerns the adequacy with which a theory captures the phenomena it aims to explain. We distinguish between two related aspects of theory validation that will clarify the argument presented in this article.

The first aspect of theory validation concerns *theory building*. During this phase, the researcher should identify the phenomena to-be-explained and construct a theory that can account for these phenomena. When formalizing a theory into a model, we can validate the model through examining whether it can adequately produce the phenomena that it has to explain (Borsboom & Haslbeck, 2024; Borsboom et al., 2021). Taking the recently proposed model of panic disorder as an example (Robinaugh et al., 2024), the validity of this model depends, in part, on its success in producing phenomena such as the avoidance of panic attacks through engaging in escape behavior and the efficacy of cognitive behavioral therapy in treating a patient with panic disorder. Ideally, a valid theory should explain all relevant phenomena for the construct of interest.

The second aspect of theory validation concerns *theory testing*. During this phase, the researcher is concerned with whether the theory's predictions are in accordance with actual observation. For this, one typically sets up a study to test whether a theory's predictions hold in empirical practice. Again taking the model of Robinaugh et al. (2024) as an example, one can validate this model by examining this model's predictions with regard to the success rate of different types of therapy and comparing these predictions to empirical observations. In these types of studies, the strength of evidence is often taken to depend on the "riskiness" of the prediction, that is on the a priori likelihood of the prediction and the prediction's uniqueness among competing theories (Vanpaemel, 2020).

Phenomenon-based Approach

As its name implies, the phenomenon-based approach places phenomena at the center of theory validation (Bogen & Woodward, 1988; Borsboom et al., 2021; Guest & Martin, 2021). On the one hand, this implies that observed phenomena inform theory, so that a theory should be constructed in such a way as to explain phenomena that are already known. For example, a new theory of gravity needs to explain all currently observed phenomena that are relevant to it, going from a simple apple falling down the tree to the bending of light around a large object. On the other hand, theory can also predict new, currently unobserved phenomena that, if discovered in the way the theory predicted, grants credence to the theory. Staying within the gravitation example, the discovery of Neptune due to irregularities in the orbit of Uranus is often cited as showing this principle in action (Gershman, 2019).

In summary, the phenomenon-based approach holds that phenomena constrain theory and that these phenomena should therefore serve as primary targets for theory validation. For theory building, this implies that one should assess the theory's success in reproducing a set of phenomena that are deemed interesting to the researcher. When using a model, the primary analysis then consists of repeatedly simulating the model's behavior under different types of conditions and different values for the parameters, providing one with a comprehensive overview of what phenomena the model is able to produce (Pitt et al., 2006). Here, we agree with proponents of the phenomenon-based approach in saying that this step is invaluable for those who aim to construct (and formalize) theory.

For theory testing, the phenomenon-based approach implies that one should test a theory's validity through an assessment of its predicted behavior. Again bringing this to the model level, one should derive a model's predicted behavior through either analytic means or through simulation studies and assess whether these predictions hold in empirical studies (e.g., Ariens et al., 2023; Navarro, 2019) Proponents of the phenomenon-based approach argue that phenomena provide sufficient, albeit indirect, evidence for theory testing (Borsboom & Haslbeck, 2024; Haslbeck et al., 2022).

However, we do not agree with this statement and believe that phenomena are not a sufficient basis to achieve this type of inference. In what follows, we will formulate three challenges to the phenomenon-based approach that show the insufficiency of phenomena for the testing of theory.

Challenges to the Phenomenon-based Approach

We now discuss three challenges when using phenomena to test a model. We will illustrate each of these challenges with an example that was chosen to be as simple as possible yet remaining as relevant as we can to real situations researchers may encounter in their investigations.

Phenomena Selection

This first issue concerns the problem of selecting phenomena for the validation of a particular model. It can be very difficult to find a set of *sufficient phenomena* – that is, a set of phenomena which can validate a theory in its entirety. To see this, we draw an instructive parallel to the concept of sufficient statistics in statistical theory. In statistics, a sufficient statistic contains all the information in a sample that is relevant for estimating the parameters of a model. For instance, when sampling from a normal distribution, the sample mean and variance together form the sufficient statistics for estimating the population parameters. However, the mean alone is insufficient: Multiple normal distributions can produce the same mean through variations in their variance. Similarly, in theory testing, a single phenomenon is often insufficient to discriminate between competing theories, as multiple theoretical mechanisms could generate the same phenomenon. Practically, this problem implies that researchers should be able to identify those phenomena that are sufficient in order to validate that theory. While we believe that there are sufficient phenomena to be found, identifying them becomes troublesome in more complex psychological theories where numerous underlying mechanisms could generate similar surface-level phenomena (e.g., Yu et al., 2023).

Example 1: To make this concrete, consider an example where researchers are testing competing theories of decision making. Through manipulation of an experimental variable (e.g., difficulty of an item), they want to investigate how different

models compare with regard to the capturing mean differences in response times between the conditions. Because such differences are consistently found over data particulars, the researchers consider these mean differences to be the crucial phenomenon-of-interest that any theory must be able to explain.

Now imagine four competing models, each specifying different expected response time distributions based on distinct underlying cognitive processes. Model A posits that response times follow a normal distribution, with differences between experimental conditions arising from a shift in the mean of this distribution—representing a simple additive effect. Model B posits a bimodal distribution, where the observed mean difference results from a strategic mixture between two response modes (e.g., fast guessing versus deliberate responding). Model C posits a skewed distribution, reflecting a threshold-based evidence accumulation process (such as in drift-diffusion models), where condition differences emerge from changes in the decision threshold or drift rate. Model D posits a heavy-tailed distribution, where rare but extreme response times play a substantial role, and mean differences are driven by changes in the tail behavior (e.g., increased likelihood of lapses or extreme delays). Through applying the phenomenon-based approach in isolation, we found that despite the fundamentally different assumptions and predictions of the four theories, we are not able to distinguish between them based on only this single phenomenon (see Figure 1). This shows how a selective focus on a single phenomenon can lead researchers to incorrectly confirm their preferred theory.

It is useful to repeat that we do think that there are sufficient phenomena to be found: In this example, we may have looked at other distributional parameters such as the skewness and kurtosis to distinguish between the four competing models (see Panel (c) in Figure 1). However, identifying such sufficient phenomena becomes increasingly difficult with more complex models that share prediction of many different phenomena.

Model Complexity

A single phenomenon can be implied by many formal theories, but the converse is also true, so that one model can imply a wide range of different phenomena. Indeed,

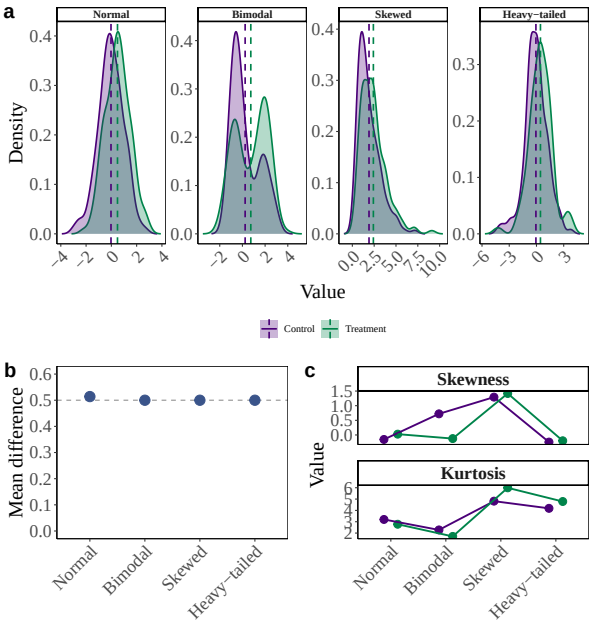


Figure 1

Illustration of the selective focus issue. Using the models described in the section, we simulated 150 data points for each model and each condition separately. Panel (a) shows the four distributions of the model, each having a distinctly different shapes representing different psychological processes: Normal (simple additive effects), bimodal (strategic choice between options), skewed (threshold-based accumulation), and heavy-tailed (processes with occasional extreme responses). Panel (b) shows that all four models produce identical mean differences of 0.5 between control and treatment conditions, which serves as the phenomenon-of-interest. Panel (c) visualizes the different skewness and kurtosis values of the models, providing clear evidence that they represent different underlying mechanisms.

while it is typically argued that verbal theories lack specificity, we should appreciate that mathematical models can suffer from the same limitation (Oude Maatman, 2025). In extreme cases, models can predict virtually any phenomenon if their parameters or boundary conditions are "tweaked" in a certain way. More generally, one can always come up with a more complex model that generates several phenomena equally well as long as it is used in a particular way, as we illustrate in the following example.

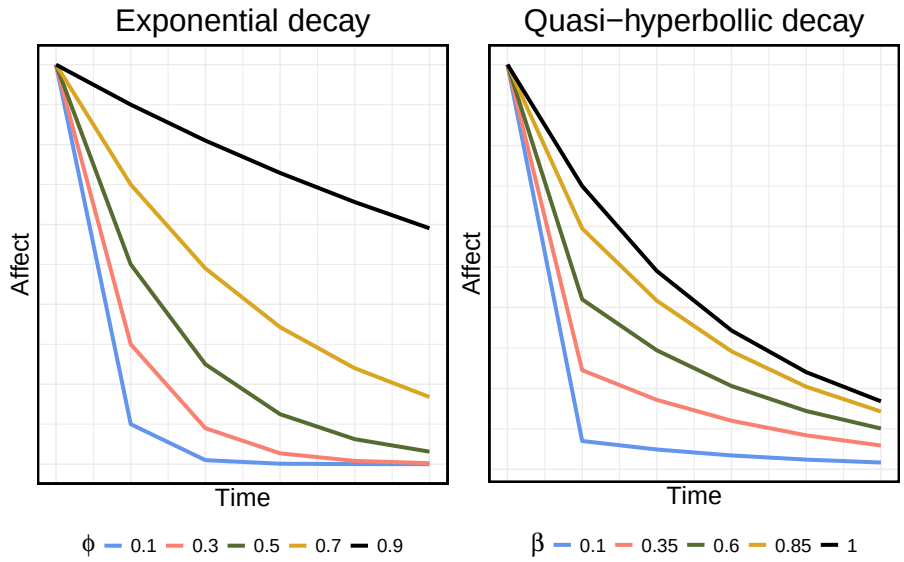


Figure 2

Illustration of the functional form of emotional inertia according to Model A (exponential, left) and Model B (quasi-hyperbolic, right). For the exponential model, we vary the defining parameter $\phi \in \{0.10, 0.30, 0.50, 0.70, 0.90\}$ *to show the strength of the regulation that is expected with each of these parameter settings. For the quasi-hyperbolic model, we fixed the parameter* $\delta = 0.70$ *and vary the value of the parameter* $\beta \in \{0.10, 0.35, 0.60, 0.85, 1.00\}$*. Observe that the quasi-hyperbolic expectation for the value of parameter* $\beta = 1$ *corresponds to the expectation of the exponential model for the value of* $\phi = 0.70$*.*

Example 2: In the field of affect dynamics, an important phenomenon is "emotional inertia" or the extent to which affective states linger on over time. Imagine that there are two models that attempt to formalize the functional form of this inertia, so that Model A assumes that affect decays exponentially over time (Ariens et al., 2020; Rutledge et al., 2014) while Model B states that affect decays in a quasi-hyperbolic fashion (Laibson, 1997; Vanhasbroeck et al., 2024, see Figure 2). In symbols, these models can be written down as:

$$\text{A:} \quad y_t = \phi^j y_{t-j} \qquad \phi \in [0, 1)$$

$$\text{B:} \quad y_t = \beta\delta^j y_{t-j} \quad \beta \in [0, 1], \delta \in [0, 1)$$

where $j \in \{1, \cdots, T\}$ represents the lag in the variable y.

Following the phenomenon-based approach, we have to find out which of the two

models best reproduces the observed decay rates. However, if we imagine that in reality affect decays exponentially, then we will not able to distinguish between these two models, as Model B is able to produce both exponential and quasi-hyperbolic decay through variation in its parameter β. In other words, reproducing the phenomenon-of-interest is not enough to differentiate between these two competitor models.

To get around this issue, one could consider the parsimony of a model to distinguish between alternatives (Borsboom et al., 2021). However, to our knowledge, the phenomenon-based approach provides no principled way of doing so, meaning there is no inherent limit to the complexity of the models we consider. Exacerbating this issue is that parsimony is difficult to define when based on phenomena alone, as even relatively simple models can produce complicated ranges of phenomena. One example is the linear differential equation which, under the right specification of the eigenvalues of its drift matrix, allows not only for the typical exponential decay function, but also for more complex oscillatory behavior (Strogatz, 2018). Should this be considered a "simple" model due to its limited number of parameters, or as a "complex" model due to the wide range of behavior that the model predicts? At this moment, the phenomenon-based approach does not answer this question.

The relevance of the complexity debate for modeling is further illustrated in an anecdote involving physicist Enrico Fermi, as retold by Freeman Dyson (Dyson, 2004). At the time, Dyson was investigating the strong nuclear force using mathematical equations that he had previously used to understand the weak nuclear force and which had their roots in the theory of quantum electrodynamics, which states that electrons and protons interact through weak electromagnetic forces. After finishing his calculations on the strong nuclear force, Dyson asked Fermi for his opinion on his work. Fermi responded by asking "How many arbitrary parameters did you use for your calculations?", afterwards concluding by quoting an adage of his friend John von Neumann: "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk" (p. 297, Dyson, 2004). Initially disheartened, Dyson abandoned the

project and only later appreciated the depth of Fermi's critique with the discovery of the quarks, which enhanced the understanding of the strong nuclear force more so than his complex model ever could have.

Circularity

The final limitation of the phenomenon-based approach is the fact that it takes phenomena as a given. If a phenomenon is indeed present in nature, a valid model should necessarily recover that phenomenon. However, this recovery becomes problematic if a phenomenon turns out to be non-existent (see also van Dongen et al., 2025) The strength of inference within the phenomenon-based approach therefore depends heavily on the phenomenon's existence.

Example 3: To make this issue concrete, imagine a research process where researchers attempt to uncover the true relationship between two quantities by building formal models and testing them using the phenomenon-based approach. Imagine that the true relationship between two variables X and Y is given by the following third-degree polynomial:

$$Y = \alpha + \beta X + \gamma X^2 + \delta X^3.$$

where $\alpha, \beta, \gamma \in \mathbb{R}$ and $\delta > 0$. Now imagine that researchers routinely fit a linear regression model to their data in order to estimate the relationship between X and Y:

$$Y = \alpha + \beta X + v_t$$
$$v_t \overset{iid}{\sim} N(0, \sigma^2)\ .$$

Across many studies, researchers would discover a seemingly linear relationship between X and Y, in the sense that the slope $\hat{\beta}$ would consistently be estimated to be positive. Because of the robustness of this phenomenon, theory builders conclude that their formal theories should accommodate the expected positive linear relationship between X and Y.

When it comes to theory testing, however, the ability of the model to reproduce a curve of the form $Y = \alpha + \beta X$ is clearly irrelevant, if not counterproductive: The more

accurately the model reproduces the phenomenon, the less likely it will be to conform to the true relationship $Y = \alpha + \beta X + \gamma X^2 + \delta X^3$. Reproducing the phenomena does not bring the literature closer to discovering, understanding, or explaining the true relationship between the variables Y and X. If the practice of validating a theory using the phenomena it was designed to explain becomes mainstream, it may be to the detriment of new ideas or findings which challenge the current way of thinking. More strongly, a pure application of the phenomenon-based approach creates an echo chamber where theory prescribes which phenomena to model without questioning whether these phenomena were correctly identified or without allowing for data to disagree.

Data-driven Approach

As an alternative to the phenomenon-based approach, researchers can instead turn to the dominant data-driven approach to theory validation. This approach holds that data serve as the primary targets for theory validation and that theory testing can be achieved through the use of model estimation and model selection techniques. The fit of the model to the data then serves as a tool to test the model and the theory it formalizes. Beyond simple fitting, one can also validate the model through prediction of unseen data. The cross-validation method levies on these types of inferences, for example, where a model is estimated on a particular part of the data called the "training set" and then used to predict the unseen data contained in the "test set" (Bates et al., 2023; Hastie et al., 2011; Stone, 1974) Importantly, both methods directly use empirical data to test the theories that are formalized with the models.

Note that these data-driven methods are not very useful when applied to only a singular model (S. Roberts & Pashler, 2000). Ideally, researchers compare the performance of multiple models that encompass different theories on the same dataset. Such a comparison is thought to inform one on the tenability of the theoretical assumptions that differ between models, therefore serving as a comparative form of theory testing.

Given our critique of the phenomenon-based approach, readers might expect us to champion the application of these pure data-driven methods. This would be

mistaken. While the data-driven approach provides us with some formal tools that lack in the phenomenon-based approach, it faces its own critical limitations, which we turn to now.

Challenges to the Data-driven Approach

Auxillary Assumptions

As we previously mentioned, researchers are required to make some auxiliary assumptions when formalizing a theory into a model (Suppes, 1962). These assumptions concern many different aspects, including the properties of the measurements (Krosnick & Presser, 2010; Stevens, 1946), structures that relate theoretical constructs to observables (Kellen et al., 2021; Regenwetter & Robinson, 2017; Robinson et al., 2025), or the structure of the error (Westfall & Yarkoni, 2016). Critically, when we use models to test theory, we never test theories in isolation. Rather, we test a theory plus auxiliary assumptions (Earp & Trafimow, 2015; Lakatos, 1978; Suppes, 1962). This creates a fundamental attribution problem. When a model fits data poorly, this failure could come from multiple sources: The core theory could be wrong, but so could any of the auxiliary assumptions of the model. Unfortunately, data alone cannot indicate the source of the misfit, meaning that current judgment of the performance of the model is largely determined by its theory-irrelevant auxiliary assumptions.

Estimation

A more practical issue concerns that of estimation. To be able to leverage on the data-driven approach, one requires their model to be estimable. This can be difficult to achieve for more complex models, especially given that classical estimation methods require the specification of a likelihood function (Spanos, 2024; Viscardi et al., 2025). Recent developments have tried to address this problem, using for example simulation-based methods (Mestdagh et al., 2019) or neural networks (Radev et al., 2020, 2023) to allow for model estimation without the explicit derivation of its likelihood function. Such developments are very useful, and allow researchers to focus on the structure of the models rather than how one should go about estimating and performing inference on the parameters. However, even with these developments, it still

holds that more complex models will be more difficult to estimate, hindering the application of these models in research.

Maybe more problematic is the implicit assumption that a model should be estimable for it to be useful. However, one can argue that models that are not estimable can still lead to important insights (Bramson et al., 2017; Campanella et al., 2014). For example, the boid model designed by Reynolds (1987) shows the possibility of flocks of birds emerging because of the tendency for these birds to account for their immediate neighbors rather than for the flock as a whole. Similarly, the model proposed by Robinaugh et al. (2024) allows for predictions that, despite its inestimability, may provide critical insight into the emergence of panic disorders. By requiring a model to be estimable, these insights would not see the light of day, severely limiting both the scope of the models that the data-driven methods apply to as well as what can be discovered through these methods.

Atheoretical Models

Finally, there exists a fundamental tension within the data-driven approach: Theoretical models often fit data less well than atheoretical models that were designed merely to predict. For example, machine learning models such as neural networks, random forests, or support vector machines routinely outperform theory-driven models on standard data-driven metrics (Yarkoni & Westfall, 2017), achieving superior fit by exploiting patterns in the data without requiring any understanding of the underlying mechanisms (Gelman et al., 2014; Piironen & Vehtari, 2017; Susko & Roger, 2020). This indicates that data-driven testing is mechanistically blind. Data-driven methods do not distinguish between mechanistic understanding and mathematical convenience, treating successful fit as equal evidence across these types of models. Taken to the extreme, applying a pure data-driven approach to model selection might lead to a proliferation of atheoretical models, a situation that is undesired when taking theory seriously (Eronen & Bringmann, 2021).

Integrative Approach

It should be clear that a sole focus on either data or phenomena will lead us into trouble. What has not been appreciated enough in the current debates on theory testing is that the strengths and limitations of both approaches are complementary. To test a theory, we need to consider both phenomena and data as they provide important counterweights to each approach's limitations. The way forward is thus to combine both approaches in an integrative approach of theory validation.

In the integrative approach, we distinguish between a top-down (phenomenon-based) and a bottom-up (data-driven) stream of analysis. The top-down stream is concerned with making and validating predictions from theory to phenomena. It starts at the theoretical framework and makes predictions about the patterns one should observe in data as well as about the conditions under which a given set of phenomena should occur. Key questions are therefore "whether a model can reproduce theoretically interesting patterns in the data" and "whether a model is able to make useful predictions"?

The bottom-up stream is concerned with evaluating how well a model fits the full structure of observed data. In other words, it starts at the data and examines how well a model can capture the statistical properties of the data. This allows not only for the comparison of different models with regard to their relative fit to a given dataset, but also for the discovery of new phenomena: By assessing the absolute fit of a model to data, one can find new data patterns that may turn out to be structural and need to be explained (van Dongen et al., 2025). Key questions are therefore "how closely a model fits the data relative to another model" and "whether the model captures all systematic patterns in the data, or whether there are signs of systematic misfit".

Importantly, both streams of analysis are necessary for theory testing. The top-down stream ensures that our tests remain theoretically grounded and produce theoretically interesting predictions. Meanwhile, the bottom-up approach ensures that we do not get stuck in an echo chamber, allowing us to more easily see what the model can and cannot explain yet.

Counterweight to the Challenges

While the integrative approach cannot fully resolve the fundamental limitations of the two approaches it is based on, it does provide an important counterweight to these challenges. We will illustrate this by discussing how the integrative approach deals with the previously identified challenges to the phenomenon-based approach.

Phenomena selection

In *Example 1*, we showed that four different formal theories produced identical mean differences of 0.5 between conditions in a decision-making task, making them indistinguishable based on this phenomenon alone. As we have explained in the section around this issue, one could try to alleviate this issue by trying to find another set of phenomena that is sufficient to distinguish between different theories. However, one could also address the issue more directly, namely through assessing the fit of the formal theories model to the data. Indeed, by fitting the model, we are imposing all of its assumptions on the data and we need not select individual phenomena.

In Table 1, we show the relative fit as assessed through the information criteria AIC and BIC. The results paint a clear picture: When a model was fit to data that it itself generated, it outperformed the alternative models. By considering data, we are now able to distinguish between the different models because all implications of the models, including their differences with respect to phenomena other than the mean, have been imposed on the data.

While one may be tempted to conclude that within this example, fit serves as a sufficient means of validation, we would disagree with such a conclusion. Fit only provides us with one piece of the puzzle and, as argued above, pursuing only fit to data would lead to a preference of atheoretical models. If our goal is to use models as a form of theory, fit to data as a sole criterion would be inadequate. Instead, it is the combination of the data-driven model fit together with the phenomenon-based predictions that lead us to believe one model is a better representation of reality than the other.

Table 1

Model comparison results using information criteria to solve the problem in Example 1. Best-fitting models for each dataset are indicated in bold.

Model	Information Criteria (AIC/BIC)			
	Normal	Bimodal	Skewed	Heavy-tailed
Normal	**850 / 865**	1026 / 1040	1040 / 1054	1006 / 1021
Bimodal	856 / 884	**909 / 938**	969 / 997	1006 / 1035
Skewed	956 / 978	1009 / 1030	**938 / 960**	1182 / 1204
Heavy-tailed	854 / 876	1030 / 1051	1019 / 1040	**993 / 1014**

Model complexity

One of the primary advantages of the integrative approach is that it allows for a principled way to handle the complexity of a model. In *Example 2*, we described a situation in which the reproduction of a phenomenon alone did not allow us to find out whether affect decays in an exponential or quasi-hyperbolic fashion. However, by assessing fit more directly and penalizing the models according to the number of parameters they have, we are able to get around this issue.

To illustrate this, we report the results of a simulation-based model comparison of the situation described in *Example 2*. We simulated 20 time-series of 5 datapoints with the exponential model, setting its parameter $\phi = 0.25$ and its error variance $\sigma^2 = 1$ (see Figure 2). When we then assess the AIC and BIC of the fit of the exponential and quasi-hyperbolic models, we find that the former ($AIC = -15.44$, $BIC = -9.63$) outperforms the latter ($AIC = -14.40$, $BIC = -5.98$). This again illustrates the value of using the integrative approach rather than an approach based on only phenomena.

Like in the previous example, we do not wish to suggest that the data-driven fit of the model is sufficient to assess the models' validity. If either Model A or Model B would have been unable to reproduce the phenomenon-of-interest (in this case, exponential decay), then we follow the phenomenon-based approach in saying that this would count as evidence against that particular model. Instead, our argument is that

only looking at this reproduction is not enough to validate a model, and that one should additionally use data-driven approaches to have a more comprehensive test of the model and the theory that it formalizes.

Additionally, we need to make explicit that in our example here, we use well-known fit metrics that provide a principled way of accounting for the complexity of a model. However, this complexity is only based on the number of parameters of the model and doesn't consider the complexity of the behavior implied by the model. Yet, considering the parsimony of a model should entail an investigation in the range of behavior that the model predicts, which may often but not always go together with the number of parameters of a model (see again the linear differential equation example Strogatz, 2018).

Unfortunately, we don't know of a principled way of accounting for the complexity of a model's qualitative behavior within the data-driven approach. Yet, one option may exist within the integrative approach: By investigating within which range of parameter values of the model a particular behavior occurs, one can leverage on the comparison of the fit of restricted and unrestricted versions of this model to get evidence for the occurrence of the more complicated behavior, as this behavior will "trade off" against parameter restrictions.

Circularity

In *Example 3*, we showed that the the ability of a model to reproduce a phenomenon can be counterproductive if the phenomenon itself is incorrectly specified. This is the case due to the absence of a principled correcting mechanism. Within the integrative approach, one is able to avoid this issue through a systematic study of misfit. Specifically, when one fits a model to data, one can examine the residuals of this model to identify missing or incorrectly specified phenomena (see Figure 3). Conversely, one may also spot overfit of a model to the data through the observation of extremely small residual variances, therefore also accommodating the issue of model complexity. Finally, one can check the values of the estimated parameters of the model, which may lie outside of the boundaries that are imposed by the theory.

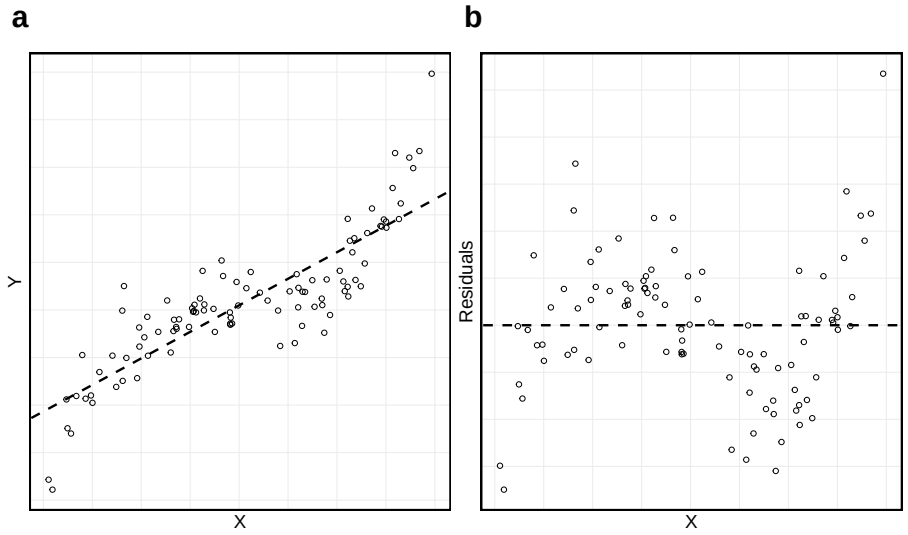


Figure 3

Inspection of the residuals of the linear regression model when fitted to $N = 100$ *data points generated from a third-degree polynomial with* $\alpha = 0, \beta = \gamma = \delta = 1$*; The situation we described in Example 3. Panel (a) shows the fit of the linear regression to the raw data. Panel (b) zooms in on the residuals of this fit, showing that residuals are not independently and identically distributed, which signals model misfit.*

Note that the use of data for the validation of a model could also be argued to be circular, as the models one fits to the data in principle need to be correctly specified to retrieve unbiased estimates of the model parameters (Ariens et al., 2023). However, a key advantage of the integrative approach is that it allows for the data to speak against the merits of the model we fit to it while keeping the model itself theory-informed and capable of making theoretically interesting predictions. These assumption violations are therefore important sources of information rather than nuisances and allow for researchers to detect that something has gone wrong. Unfortunately, this kind of correction mechanism is often neglected (for exceptions, see Revol et al., in press; Vanhasbroeck et al., 2022).

We must admit, though, that if a researcher encounters an assumption violation, this might indicate a failure of an auxiliary assumption rather than a theoretical one. In this case, the task of the model builder will be to reconsider either the method with which data has been collected, the auxiliary assumptions of the model, or the theory

itself in light of this mismatch.

Strengths and Limitations

We end this section with a discussion of several strengths and limitations of the integrative approach. Here, we focus on those features that have not been mentioned earlier in this article but which show how the phenomenon-based and data-driven approaches enrich each other.

Strengths

A major strength of the integrated approach is the direct testability of a model to data and, critically, the ability to estimate a realistic set of parameters from those data. The importance of a realistic set of parameters for evaluating a model can not be understated: Parameters govern the behavior of the model. The ability to estimate the parameters of a model is therefore useful, as it provides us with a realistic parameter set that can be used to predict phenomena that can be realistically expected to occur – the hallmark of strong scientific theories. Additionally, it allows for an empirical check of whether the parameter values satisfy theoretical constraints, rather than merely assuming that the parameters do. This type of analysis corresponds to what Pitt et al. (2006) referred to as local analysis, which focuses on evaluating a model's behavior at specific parameter values. In contrast, global analysis explores the full range of qualitative behaviors a model can produce across its entire parameter space. Critically, Pitt et al. (2006) argue that while global analysis offers valuable theoretical insights into the range of behaviors a model can produce, it should not stand alone. Instead, it must be complemented by local analysis based on parameter estimates derived from actual data. We fully agree with this position.

Parameter recoverability and identifiability can furthermore serve as critical diagnostics for whether theoretical constructs in a model can be empirically grounded. When different combinations of parameter values generate the same phenomena, or when parameter estimates fluctuate erratically across similar datasets, the correspondence between parameters and psychological constructs becomes vague, despite mathematical formalization. Pushing this argument a bit further, well-designed

models may use their parameters as indirect measurements of psychological constructs, extending the reach of psychological science beyond directly observable behavior (e.g., Brown & Heathcote, 2008; Heathcote et al., in preparation; Kahneman & Tversky, 1979; Yu et al., 2023, 2025). When a model includes parameters which represent mechanisms like learning rates, generalization gradients, or attentional focus, estimating these parameters allows researchers to investigate how specific psychological processes vary across individuals, developmental stages, or experimental conditions. These parameter-based insights provide a deeper understanding of psychological functioning than would be possible through analysis of behavioral measures alone.

Limitations

Combined Inference. While we have focused on how the data-driven and phenomenon-based approaches complement each other, we have not discussed what happens when the two methods fail to fully solve the limitations of each approach separately. For example, model fit may prefer models that do not necessarily reproduce all phenomena the researcher is interested in, leaving it to the researcher to determine which of the two pieces of information is more crucial. Additionally, there is some subjectivity in the metric that is chosen to assess model fit, a type of subjectivity that is also present in operationalizing the phenomena to be reproduced according to the phenomenon-based approach. Unfortunately, this subjectivity is not resolved in the integrative approach. In other words, while the integrative approach provides the researcher with a broader set of methods, it is not a catch-all remedy against problems in inferential reasoning.

Instead, one should view the use of the integrative approach as providing more direct feedback towards the modeler through both the data and the phenomena. To leverage on the strength of both approaches, one therefore has to consider this valuable feedback and integrate it into future iterations of the model.

Estimability. Like the data-driven approach, the integrative approach only applies to the testing of models that are estimable. Building a model that is both theoretically valid and parsimonious enough to be fit to data is no easy task, but we

believe a necessary one. In cases where this is infeasible, however, one can use the methods of the phenomenon-based approach to perform theory building. However, we caution readers to closely consider the limitations of this approach with regard to theory testing, asking them to remain cautious when claiming validity of a model.

Discussion

In this paper, we have provided a critical evaluation of the strengths and limitations of the phenomenon-based approach to theory testing. Additionally, we have argued that the inclusion of data in validating a model can mitigate the issues associated to the use of the phenomenon-based approach alone. Instead, we have argued for combining the phenomenon-based and data-driven approaches to theory validation into an integrative approach, which we believe provides the most promising avenue for future research.

It is important to reiterate that we do not disregard the importance of phenomena when building and testing a theory. The phenomenon-based approach represents a fresh breath of air in an otherwise data-driven field, allowing for the initial validation of theories that do not necessarily allow for estimation. However, we believe that there is a crucial piece missing from the methodology proposed by Borsboom et al. (2021) with regard to testing theory, and that data represents this missing piece. Similarly, we believe that purely data-driven methods have their own place in the literature, but that a focus on only data may equally harm theory validation attempts. Given the complementary nature of the strengths and weaknesses of both approaches when used in isolation, a combination of these approaches may yield the best results.

We close by returning to the previously mentioned discovery of Neptune, as it provides a beautiful historical example of the topics we have discussed in this article. As mentioned before, the discovery of Neptune is sometimes cited as an example of the phenomenon-based approach in action (Haslbeck et al., 2022). However, it is easy to overlook the decades of observation, prediction, and theory testing that led up to this discovery and which, when taken together, are indicative of the integrative approach to theory validation.

Uranus displayed "residual perturbations", that is deviations from the elliptic orbit predicted by the law of gravitation, which became apparent due to a systematic comparison of decades of carefully gathered observations (data) with the orbits (phenomena) implied by the theory of gravity (theory). To explain why these deviations were observed, astronomers put forward a few theories (Sheehan et al., 2021). First, some astronomers questioned the quality of the data, stating that old observations might be unreliable, while others argued that the theory of gravity was incorrect and put forward alternative theories that accommodated the phenomenon of residual perturbations (e.g., through selective attraction, see Sheehan et al., 2021). Yet other astronomers were less quick to seek to modify the theory of gravitation, writing that "the law of gravitation was too firmly established to be doubted till every other hypothesis had failed" (quote of John Adams taken from Bamford, 1996, p. 216).

Accepting both the validity of the model and the data, astronomers put forward the possibility that an undiscovered planet caused the deviations in the orbit of Uranus, indicating a problem with our understanding of the solar system at the time. Importantly, this hypothesis was supported through *fitting various orbits to the observed data*. For example, Urbain Leverrier attempted to explain the irregularities in the orbit of Uranus by correcting the relevant observations for the gravitational pull of other, known planets, concluding that "no ellipse would satisfy the range of observations, ancient and modern, even on the most favorable distribution of errors in them" (Bamford, 1996, p. 215). Following this result, Leverrier then approximated the location of Neptune by assuming the existence of another planet so that, when correcting the observations in Uranus for the pull of such a planet, the errors in the orbit of Uranus would no longer be systematic (Grant, 1852).

It is clear that the discovery of Neptune arose due to both a top-down stream, where theoretical predictions are derived from a theory mathematically, and a bottom-up stream, where data is used to judge the validity of these predictions. It is difficult to see how any stream in isolation would have led to the discovery of Neptune. Indeed citing Sheehan et al. (2021): "The discovery of Neptune was a story of two

parts. The first one was theoretical, belonging to what was then called ‘mathematical astronomy’ (...) The other part was empirical, and consisted in preparing the ground, such as charting the skies to facilitate location of transiting bodies, and the actual observations necessary for any genuine discovery of astronomical objects.” (p. 189).

References

Ariens, S., Ceulemans, E., & Adolf, J. K. (2020). Time series analysis of intensive longitudinal data in psychosomatic research: A methodological overview. *Journal of Psychosomatic Research*, *137*, 110191. https://doi.org/10.1016/j.jpsychores.2020.110191

Ariens, S., Adolf, J. K., & Ceulemans, E. (2023). One does not simply correct for serial dependence. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(5), 807–821.

Bamford, G. (1996). Popper and his commentators on the discovery of Neptune: A close shave for the law of gravitation? *Studies in History and Philosophy of Science Part A*, *27*(2), 207–232.

Bastida-Castillo, A., Gómez-Carmona, C. D., De La Cruz Sánchez, E., & Pino-Ortega, J. (2019). Comparing accuracy between global positioning systems and ultra-wideband-based position tracking systems used for tactical analyses in soccer. *European Journal of Sport Science*, *19*, 1157–1165. https://doi.org/10.1080/17461391.2019.1584248

Bates, S., Hastie, T., & Tibshirani, R. (2023). Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, *119*, 1434–1445. https://doi.org/10.1080/01621459.2023.2197686

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophical Review*, *97*(3), 303–352. https://doi.org/10.2307/2185445

Borsboom, D., & Haslbeck, J. M. B. (2024). Integrating intra- and interindividual phenomena in psychological theories. *Multivariate Behavioral Research*, *59*, 1290–1309. https://doi.org/10.1080/00273171.2024.2336178

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*, 756–766. https://doi.org/10.1177/1745691620969647

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Flocken, C., & Holman, B. (2017). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science*, *84*, 115–159. https://doi.org/10.1086/688938

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Campanella, M., Hoogendoorn, S., & Daamen, W. (2014). The Nomad model: Theory, developments and applications. *Transportation Research Procedia*, *2*, 462–467. https://doi.org/10.1016/j.trpro.2014.09.061

Cloos, L., Ceulemans, E., & Kuppens, P. (2023). Development, validation, and comparision of self-report measures for positive and negative affect in intensive longitudinal research. *Psychological Assessment*, *35*, 189–204. https://doi.org/10.1037/pas0001200

Darwin, C. (1859). *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. John Murray.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2000). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (2nd ed., pp. 200–223). Cambridge University Press.

Dutilh, G., Wagenmakers, E.-J., Visser, I., & van der Maas, H. L. J. (2011). A phase transition model for the speed-accuracy trade-off in response time. *Cognitive Science*, *35*, 211–250. https://doi.org/10.1111/j.1551-6709.2010.01147.x

Dyson, F. (2004). A meeting with Enrico Fermi. *Nature*, *427*, 297. https://doi.org/10.1038/427297a

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621. https://doi.org/10.3389/fpsyg.2015.00621

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*, 779–788. https://doi.org/10.1177/1745691620970586

Eronen, M. I., & Romeijn, J.-W. (2020). Philosophy of science and the formalization of psychological theory. *Theory & Psychology*, *30*, 786–799. https://doi.org/10.1177/0959354320969876

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science*, *19*, 329–335. https://doi.org/10.1177/0963721410386677

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Gamble, A. S. D., Bigg, J. L., Pignanelli, C., Nyman, D. L. E., Burr, J. F., & Spriet, L. L. (2023). Reliablity and validity of an indoor local positioning system for measuring external load in ice hockey players. *European Journal of Sport Science*, *23*, 311–318. https://doi.org/10.1080/17461391.2022.2032371

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin and Review*, *26*, 13–28. https://doi.org/10.3758/s13423-018-1488-8

Grant, R. (1852). *History of physical astronomy from the earliest ages to the middle of the nineteenth century*. Bohn.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*, 789–802. https://doi.org/10.1177/1745691620970585

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, *10*, 371–388. https://doi.org/10.1037/1082-989X.10.4.371

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2022). Modeling psychopathology: From data models to formal theories. *Psychological Methods*, *27*, 930–957. https://doi.org/10.1037/met0000303

Hastie, T., Tibshirani, R., & Friedman, J. (2011). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

737 Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The
738 case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*,
739 185–207. https://doi.org/10.3758/BF03212979
740 Heathcote, A., Vanhasbroeck, N., Anderson, A., Blanken, T., Borsboom, D., &
741 Matzke, D. (in preparation). A psychological modeling framework for individual
742 pedestrian decisions in complex environments.
743 Hedrick, T. L. (2008). Software techniques for two- and three-dimensional kinematic
744 measurements of biological and biomimetic sytems. *Bioinspiration &*
745 *Biomimetics*, *3*, 034001. https://doi.org/10.1088/1748-3182/3/3/034001
746 Jangraw, D. C., Keren, H., Sun, H., Bedder, R. L., Rutledge, R. B., Pereira, F.,
747 Thomas, A. G., Pine, D. S., Zheng, C., Nielson, D. M., & Stringaris, A. (2023).
748 A highly replicable decline in mood during rest and simple tasks. *Nature Human*
749 *Behaviour*, *7*, 596–610. https://doi.org/10.1038/s41562-023-01519-7
750 Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under
751 risk. *Econometrica*, *47*, 263–292. https://doi.org/10.2307/1914185
752 Kellen, D., Davis-Stober, C. P., Dunn, J. C., & Kalish, M. L. (2021). The problem of
753 coordination and the pursuit of structural constraints in psychology. *Perspectives*
754 *on Psychological Science*, *16*(4), 767–778.
755 https://doi.org/10.1177/1745691620974771
756 Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective
757 measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw,
758 C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process*
759 *quality*. Johnn Wiley & Sons, Inc.
760 Krosnick, J. A., & Presser, A. (2010). Question and questionnaire design. In
761 P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed.).
762 Emerald.
763 Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of*
764 *Economics*, *2*, 443–477.

765 Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge
766 University Press.
767 Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning
768 curves: A test of the instance theory of automaticity. *Journal of Experimental*
769 *Psychology: Learning, Memory, and Cognition*, *18*, 883–914.
770 https://doi.org/10.1037//0278-7393.18.5.883
771 Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and*
772 *Emotion*, *23*, 209–237. https://doi.org/10.1080/02699930802204677
773 Mestdagh, M., Verdonck, S., Meers, K., Loossens, T., & Tuerlinckx, F. (2019). Prepaid
774 parameter estimation without likelihoods. *PLoS Computational Biology*, *15*,
775 e1007181. https://doi.org/10.1371/journal.pcbi.1007181
776 Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between
777 scientific judgement and statistical model selection. *Computational Brain &*
778 *Behavior*, *2*, 28–34. https://doi.org/10.1007/s42113-018-0019-z
779 Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of
780 models and the informativeness of data. *Cognitive Psychology*, *49*, 47–84.
781 https://doi.org/10.1016/j.cogpsych.2003.11.001
782 Newell, A., & Rosenbloom, P. S. (1980). Mechanisms of skill acquisition and the law of
783 practice [Carnegie Mellon University].
784 Oude Maatman, F. J. W. (2025). Psychology's theory crisis, and why formal modelling
785 cannot solve it. *Meta-Psychology*, *9*, MP.2024.4224.
786 https://doi.org/10.15626/MP.2024.4224
787 Pathria, R. K., & Beale, P. D. (2022). *Statistical mechanics* (4th ed.). Academic Press.
788 Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian predictive methods for
789 model selection. *Statistics and Computing*, *27*(3), 711–735.
790 https://doi.org/10.1007/s11222-016-9649-y
791 Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by
792 parameter space partitioning. *Psychological Review*, *113*, 57–83.
793 https://doi.org/10.1037/0033-295X.113.1.57

Popper, K. (1959). *The logic of scientific discovery*. Routledge.

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., & Köthe, U. (2020). BayesFlow: Learning complex stochastic models with invertible neural networks. *IEEE transactions on neural networks and learning systems*, *33*(4), 1452–1466.

Radev, S. T., Schmitt, M., Schumacher, L., Elsemüller, L., Pratz, V., Schälte, Y., Köthe, U., & Bürkner, P.-C. (2023). BayesFlow: Amortized Bayesian workflows with neural networks. *Journal of Open Source Software*, *8*(89), 5702.

Ratcliff, R., & Rouder, J. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*, 347–356. https://doi.org/10.1111/1467-9280.00067

Read, G. L. (2017). Facial electromyography (EMG). In J. Matthes, C. S. Davis, & R. F. Potter (Eds.), *The international encyclopedia of communication research methods* (pp. 1–10). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118901731.iecrm0100

Regenwetter, M., & Robinson, M. M. (2017). The construct–behavior gap in behavioral decision research: A challenge beyond replicability. *Psychological Review*, *124*(5), 533–550. https://doi.org/10.1037/rev0000067

Revol, J., Ariens, S., Lafit, G., Adolf, J. K., & Ceulemans, E. (in press). Episode-contingent experience-sampling designs for accurate estimates of autoregressive dynamics [Accepted for publication in Psychological Methods].

Reynolds, C. W. (1987). Flocks, herds, and schools: A distributed behavioral model. *Computer Graphics*, *21*, 25–34.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., LahozMonfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*, 913–929. https://doi.org/10.1111/ecog.02881

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367. https://doi.org/10.1037/0033-295X.107.2.358

Robinaugh, D. J., Haslbeck, J. M. B., Waldorp, L. J., Kossakowski, J. J., Fried, E. I., Millner, A. J., McNally, R. J., R., O., de Ron, J., van der Maas, H. L. J., van Nes, E. H., Scheffer, M., Kendler, K. S., & Borsboom, D. (2024). Advancing the network theory of mental disorders: A computational model of panic disorder. *Psychological Review*, *131*, 1482–1508. https://doi.org/10.1037/rev0000515

Robinson, M. M., Williams, J. R., Wixted, J. T., & Brady, T. F. (2025). Zooming in on what counts as core and auxiliary: A case study on recognition models of visual working memory. *Psychonomic Bulletin & Review*, *32*(2), 547–569. https://doi.org/10.3758/s13423-024-02562-9

Russel, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*, 493–502. https://doi.org/10.1037/0022-3514.57.3.493

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*, 12252–12257. https://doi.org/10.1073/pnas.1407535111

Schroeder, D. V. (2021). *An introduction to thermal physics*. Oxford University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Sheehan, W., Bell, T. E., Kennett, C., & Smith, R. W. (2021). Neptune: From grand discovery to a world revealed. https://doi.org/10.1007/978-3-030-54218-4

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational social psychology* (pp. 311–331). Routledge.

Smaldino, P. E. (2020). How to translate a verbal theory into a formal model. *Social Psychology*, *51*, 207–218. https://doi.org/10.1027/1864-9335/a000425

Spanos, A. (2024). How the post-data severity converts testing results into evidence for or against pertinent inferential claims. *Entropy*, *26*(1), 95. https://doi.org/10.3390/e26010095

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680. https://doi.org/10.1126/science.103.2684.677

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B (Methodological)*, *38*, 111–147.

Strogatz, S. H. (2018). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering* (2nd ed.). CRC Press.

Suppes, P. (1962). Models of data. In *Logic, methodology and the philosophy of science: Proceedings of the 1960 international congress* (pp. 252–261, Vol. 44). Stanford University Press.

Susko, E., & Roger, A. J. (2020). On the use of information criteria for model selection in phylogenetics (N. Saitou, Ed.). *Molecular Biology and Evolution*, *37*(2), 549–562. https://doi.org/10.1093/molbev/msz228

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16*, 437–450.

van Dongen, N., van Bork, R., Finnemann, A., Haslbeck, J. M. B., van der Maas, H. L. J., Robinaugh, D. J., de Ron, J., Sprenger, J., & Borsboom, D. (2025). Productive explanation: A framework for evaluating explanations in psychological science. *Psychological Review*, *132*, 311–329. https://doi.org/10.1037/rev0000479

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*, 682–697. https://doi.org/10.1177/1745691620970604

Vanhasbroeck, N., Loossens, T., Anarat, N., Ariens, S., Vanpaemel, W., Moors, A., & Tuerlinckx, F. (2022). Stimulus-driven affective change: Evaluating computational models of affect dynamics in conjunction with input. *Affective Science*, *3*, 559–576. https://doi.org/10.1007/s42761-022-00118-5

Vanhasbroeck, N., Loossens, T., & Tuerlinckx, F. (2024). Two peas in a pod: Discounting models as a special case of the VARMAX. *Journal of Mathematical Psychology*, *120-121*, 102856. https://doi.org/10.1016/j.jmp.2024.102856

Vanpaemel, W. (2020). Strong theory testing using the prior predictive and the data prior. *Psychological Review*, *127*, 136–145. https://doi.org/10.1037/rev0000167

Viscardi, C., Lachi, A., & Baccini, M. (2025). Discrete-time compartmental models with partially observed data: A comparison among frequentist and Bayesian approaches for addressing likelihood intractability. *Epidemiologic Methods*, *14*(1), 20240032. https://doi.org/10.1515/em-2024-0032

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28–50. https://doi.org/10.1016/j.jmp.2003.11.004

Wehrens, R., Putter, H., & Buydens, L. M. C. (2000). The boostrap: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, *54*, 35–52.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think (U. S. Tran, Ed.). *PLOS ONE*, *11*(3), e0152719. https://doi.org/10.1371/journal.pone.0152719

Woodward, J. (1989). Data and phenomena. *Synthese*, *79*, 393–472.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yu, K., Tuerlinckx, F., Vanpaemel, W., & Zaman, J. (2023). Humans display interindividual differences in the latent mechanisms underlying fear generalization behaviour. *Communications Psychology*, *1*(1), 5. https://doi.org/10.1038/s44271-023-00005-0

Yu, K., Vanpaemel, W., Tuerlinckx, F., & Zaman, J. (2025). The probabilistic and dynamic nature of perception in human generalization behavior. *iScience*, *28*(4), 112228. https://doi.org/10.1016/j.isci.2025.112228

Acknowledgements: The authors would like to thank Prof. Denny Borsboom and Kyra Evers for critical and constructive discussions on the contents of this paper. We furthermore thank the reviewers for their valuable comments.

Conflict of interest: The authors report no conflict of interest.

Contribution: All authors contributed equally.

Funding: N.V. is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 Excellent Science program (Grant agreement No. 101053880). K.Y. is supported by a grant from the Fund for Scientific Research - Flanders (FWO; G079520N) and in part by the Research Fund of KU Leuven (C14/23/062). S.A. is supported by a grant from FWO (1278525N).