AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

© 2021 American Psychological Association
ISSN: 1082-989X

https://doi.org/10.1037/met0000303

Modeling Psychopathology: From Data Models to Formal Theories

Jonas M. B. Haslbeck[1*], Oisín Ryan[2*], Donald J. Robinaugh[3*], Lourens J. Waldorp[1], and Denny Borsboom[1]

[1] Department of Psychology, University of Amsterdam
[2] Department of Methodology and Statistics, Utrecht University
[3] Massachusetts General Hospital, Harvard Medical School

Abstract

Over the past decade, there has been a surge of empirical research investigating mental disorders as complex systems. In this article, we investigate how to best make use of this growing body of empirical research and move the field toward its fundamental aims of explaining, predicting, and controlling psychopathology. We first review the contemporary philosophy of science literature on scientific theories and argue that fully achieving the aims of explanation, prediction, and control requires that we construct formal theories of mental disorders: theories expressed in the language of mathematics or a computational programming language. We then investigate three routes by which one can use empirical findings (i.e., data models) to construct formal theories: (a) using data models themselves as formal theories, (b) using data models to infer formal theories, and (c) comparing empirical data models to theory-implied data models in order to evaluate and refine an existing formal theory. We argue that the third approach is the most promising path forward. We conclude by introducing the abductive formal theory construction (AFTC) framework, informed by both our review of philosophy of science and our methodological investigation. We argue that this approach provides a clear and promising way forward for using empirical research to inform the generation, development, and testing of formal theories both in the domain of psychopathology and in the broader field of psychological science.

Translational Abstract

Over the last decade, there has been a surge of empirical research investigating mental disorders as networks of interacting symptoms. This rapidly growing empirical literature has raised a critical question: How can we best make use of these empirical findings to achieve our aim of explaining, predicting, and controlling mental disorders? In this article, we argue that achieving these aims requires the construction of formal theories and we investigate how empirical research can best inform the construction of well-developed formal theories. We begin by reviewing the philosophy of science literature to clarify the nature of formal theories, data models, and the relationship between them. We identify three plausible ways in which empirical data models can be used to develop formal theories. In the first, data models are treated as formal theories. In the second, data models are used to make direct inferences about the real world and, thereby, inform the development of a formal theory. In the third, the empirical data model is compared to a theory-implied data model, and any differences between them is used to inform subsequent theory development. Using simulations from a computational model of panic disorder, we investigate which of these three routes best informs the development of formal theories of psychopathology and conclude that the third approach is most promising. We then build on this evaluation by proposing the abductive formal theory construction (AFTC) framework: a three-stage framework rooted in abductive inference and the comparison between theory-implied and empirical data models. We argue that this approach provides a challenging, yet promising way forward for using empirical research to construct formal theories.

Keywords: theory development, formal theories, network approach, complex dynamical systems, computational modeling

Jonas M. B. Haslbeck https://orcid.org/0000-0001-9096-7837
Oisín Ryan https://orcid.org/0000-0003-3698-6396
Lourens J. Waldorp https://orcid.org/0000-0002-5941-4625

*Jonas M. B. Haslbeck, Oisín Ryan, and Donald J. Robinaugh contributed equally to this work and should be considered joint first authors.

We thank Tessa Blanken, Fabian Dablander, Nathan Evans, Madelyn Frumkin, Ellen Hamaker, Richard McNally, Alex Millner, and Dirk Wulff for their feedback on earlier versions of this article. Jonas M. B. Haslbeck has been supported by the European Research Council Consolidator Grant 647209. Oisín Ryan has been supported by the Netherlands Organization for Scientific Research Onderzoekstalent Grant 406-15-128. Donald J. Robinaugh has been supported by a National Institute of Mental Health Career Development Award (1K23MH113805-01A1).

Correspondence concerning this article should be addressed to Jonas M. B. Haslbeck, Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT, Amsterdam, the Netherlands. Email: jonashaslbeck@gmail.com

Mental disorders are complex phenomena: highly heterogeneous and massively multifactorial (e.g., Kendler, 2019). In recent years, researchers have called for approaches to psychiatric research that embrace this complexity, evaluating how mental disorders operate as complex systems (Gardner & Kleinman, 2019; Hayes & Andrews, 2020). The "network approach" to psychopathology addresses these calls, conceptualizing mental disorders as systems of interacting components, with emphasis on causal relations among the symptoms of a disorder (e.g., Borsboom, 2017; Borsboom & Cramer, 2013; Schmittmann et al., 2013). From this perspective, symptoms are not caused by an underlying disorder, rather the symptoms themselves and the causal relations among them constitute the disorder.

The notion that causal relationships among symptoms may figure prominently in the etiology of mental disorders has stimulated a rapidly growing body of empirical research (for reviews see e.g., Contreras et al., 2019; Robinaugh et al., 2019). Most of this work employs statistical models that allow researchers to study the multivariate dependencies among symptoms. This quickly expanding empirical literature has provided rich information about the statistical relationships among those symptoms. However, it has also raised a significant concern: it remains unclear precisely how best to make use of this growing number of empirical findings to produce advances in our understanding of how mental disorders operate as complex systems.

This problem is not unique to the network approach to psychopathology. Psychiatry and applied psychology produce a massive number of empirical findings every year, yet genuine progress toward our fundamental aims of explaining, predicting, and controlling mental disorders has remained stubbornly out of reach. In psychology more broadly, there is a growing concern that psychological theory is in a state of crisis (Oberauer & Lewandowsky, 2019; Muthukrishna & Henrich, 2019; Smaldino, 2019). Theories are rarely developed in a way that would indicate a genuine accumulation of knowledge, suggesting that we are failing to leverage the steady stream of empirical findings from psychological science into genuine understanding of psychological phenomena (Meehl, 1978).

Recently, we and others have argued that formalizing psychological theories as mathematical or computational models can help address the theory crisis in psychology (Borsboom et al., 2020; Fried, 2020; Guest & Martin, 2020; Robinaugh et al., 2021; Smaldino, 2017; van Rooij & Baggio, 2020). Formal theories have been fruitfully used in some areas of psychology, such as mathematical psychology (Estes, 1975), cognitive psychology (Ritter et al., 2019), and computational psychiatry (Friston et al., 2017; Huys et al., 2016), but remain relatively rare outside these domains. More expansive use of formal theories, it is hoped, will equip theorists with tools for more rigorously generating and evaluating theories, laying the groundwork for accumulative advancement of psychological knowledge (Robinaugh et al., 2021). However, much remains unknown about how best to construct formal theories in domains such as clinical psychology, and even less is known about how best to use empirical data models to inform theory construction.

In this article, we aim to address this gap in the literature by examining how data models commonly used within the network approach literature can best inform the construction of formal theories. We begin by discussing the nature of formal theories, the nature of data models, and their relation to one another. We identify three ways in which data models can be used to inform the construction of formal theories: (a) treating data models themselves as formal theories, (b) drawing direct inferences from data models to generate a formal theory, and (c) comparing theory-implied data models and empirical data models in order to evaluate and refine an existing formal theory. In which follows, we investigate each of these approaches in the context of the network approach to psychopathology, using an example in which the true underlying system is known and evaluating which approach best informs the development of a theory of that system. Our analysis suggests that the third approach comparing theory-implied and empirical data models, though rare in psychology, is the most promising path forward. In the penultimate part of the paper, we propose the abductive formal theory construction (AFTC) framework: a staged methodology for theory construction built around the approach of comparing theory-implied and empirical data models. Using this framework, we detail how best to use empirical data models at each stage of theory construction, including the generation, development, and testing of psychological theories.

Data Models and Formal Theories

In this section we will examine the nature of scientific theories and how they support explanation, prediction, and control. We will begin by introducing four key concepts that we will use throughout the remainder of the article: theory, target system, data, and data model (see Figure 1). We will illustrate each of these concepts using the example of panic disorder.

Theories, Phenomena, and Target Systems

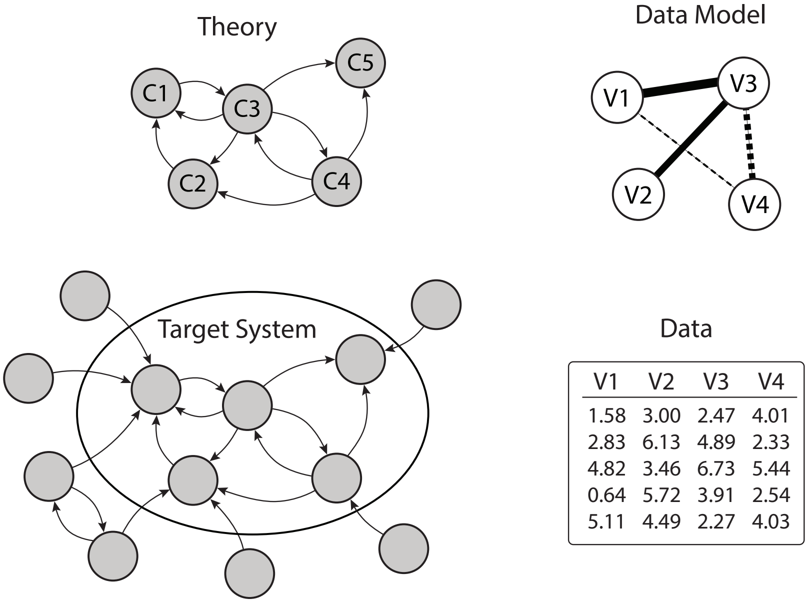
Theories seek to explain *phenomena*: stable, recurrent, and general features of the world (Bogen & Woodward, 1988; Haig, 2008, 2014) such as the melting point of lead or the orbit of planets. In psychiatry, the most common phenomena to be explained are symptoms and syndromes. For example, researchers seek to explain the tendency for some individuals to experience panic attacks and the tendency for recurrent panic attacks to be accompanied by persistent worry about those attacks and avoidance of situations in which they may occur (Spitzer et al., 1980), thereby cohering as the syndrome known as panic disorder.

Theories aim to explain phenomena by representing *target systems*: the particular parts of the real world and the relationships among them that give rise to the phenomena of interest (cf. Elliott-Graves, 2014). Theories can thus be understood as models that represent the target system (Suárez & Pero, 2019).[1] In psychiatric research, the target system comprises any components of the real world that give rise to these symptoms and syndromes, and may include genetic, neurobiological, physiological, emotional,

[1] The precise nature of scientific theories is a subject of ongoing debate among philosophers of science and the relationship between theories and models is muddled by inconsistent and often conflicting use of these terms across time, disciplines, and scientists (for a brief history of models and their relation to theory, see Bailer-Jones, 2009). In this article, we will adopt the perspective that theories are models (Suárez & Pero, 2019). However, the core arguments presented in this article do not require this precise conceptualization of theories and would similarly hold for pragmatic accounts that regard models as an intermediary between theory and the real world (e.g., Bailer-Jones, 2009; Cartwright, 1983).

Figure 1

Key Concepts Theory, Target System, Data, and Data Model



V1	V2	V3	V4
1.58	3.00	2.47	4.01
2.83	6.13	4.89	2.33
4.82	3.46	6.73	5.44
0.64	5.72	3.91	2.54
5.11	4.49	2.27	4.03

Note. The figure illustrates the concepts theory, target system, data, and data model. The target system is the system consisting of interacting components that gives rise to phenomena. Phenomena are robust features of the world captured by data models. Theories represent the structure of the target system, proposing a set of components C and the relations among them and positing that they give rise to the phenomena. Data for variables V are obtained by taking measurements of the components of the target system.

cognitive, behavioral, social, and cultural components. Theories of psychopathology aim to represent these target systems, positing a specific set of components and relationships among them that give rise to the phenomena of interest. For example, researchers have generated numerous theories of panic disorder, specifying a set of components that they think interact to give rise to panic attacks and panic disorder. Among these, perhaps the most influential is Clark's cognitive model of panic attacks, which posits that "if [stimuli] are perceived as a threat, a state of mild apprehension results. This state is accompanied by a wide range of bodily sensations. If these anxiety-produced sensations are interpreted in a catastrophic fashion, a further increase in apprehension occurs. This produces a further increase in bodily sensations and so on round in a vicious circle which culminates in a panic attack" (Clark, 1986). This cognitive theory of panic attacks specifies components (e.g., bodily sensations and a state of apprehension) and the relations among them (e.g., the "vicious cycle" of positive causal effects), positing that this is the target system that gives rise to panic attacks.

Because theories represent the target system, we can reason from theory in order to draw conclusions about the target system and the phenomena that arise from it. It is this capacity for *surrogative reasoning* (Swoyer, 1991) that allows theories to explain, predict, and control the world around us. For example, we can explain the rise and fall of predator and prey populations in the real world by appealing to the relationships between components specified in mathematical models representing these populations (H. I. Freedman, 1980; Nguyen & Frigg, 2017). We can predict what will occur when two atoms collide by deriving the expected outcome from models of particle physics (Higgs, 1964). And we can determine how to intervene to prevent panic attacks by appealing to the relationships posited in the cognitive model of panic attacks, determining that an intervention modifying a patient's "catastrophic misinterpretations" should prevent the "vicious cycle" between arousal and perceived threat, thereby circumventing panic attacks (Clark, 1986). It is this ability to support surrogative reasoning that makes theories such powerful tools.

The Importance of Formal Theories

Surrogative reasoning relies on a theory's structure: its components and the relations among them (Pero, 2015). This structure can be expressed in natural language (i.e., verbal *theory*) or a formal language, such as mathematics or computation (i.e., formal *theory*). For example, a verbal theory would state that the rate of change in an object's temperature is proportional to the difference between its temperature and the temperature of its environment. A formal theory would instead express this relationship as a mathematical equation, such as $\frac{dT}{dt} = -k(T - E)$, where $\frac{dT}{dt}$ is the rate of change in temperature, T is the object's temperature, and E is the temperature of the environment; or in a computational programming language, such as: for (t in 1:end) $\{T[t+1] = T[t] - k \times (T[t] - E)\}$.

Expressing a theory in a mathematical or computational programming language gives formal theories many advantages over verbal theories (Epstein, 2008; Lewandowsky & Farrell, 2010; Smaldino, 2017; Smith & Conrey, 2007; Robinaugh et al., 2021).

For our purposes here, there is one advantage of particular importance: Formalization enables precise deduction of the behavior implied by the theory. Verbal theories can, of course, also be used to deduce theory-implied behavior. However, due to the vagaries of language, verbal theories are typically imprecise, precluding exact predictions. For example, the verbal theory of temperature cooling described in the previous paragraph allows for some general sense of how the object's temperature will evolve over time, but cannot be used to make specific predictions about how it will change or what the temperature will be at any given point in time. Indeed, because of the imprecision of verbal theories, there are often multiple ways in which those theories could be interpreted and implemented, each with a potentially divergent prediction about how the target system will evolve over time (Robinaugh et al., 2021). Consider the interpersonal theory of suicide, which posits that suicide arises from the simultaneous experience of perceived burdensomeness and thwarted belongingness (Van Orden et al., 2010). This theory fails to specify many aspects of this causal structure, such as the strength of these effects or the duration for which they must overlap before suicidal behavior arises (Hjelmeland & Loa Knizek, 2018). As a result, there are many possible implementations of this verbal theory, each of which could potentially lead to a different prediction about when suicidal behavior should be expected to arise (Millner et al., 2020). This imprecision thus substantially limits the theories ability to support surrogative reasoning and the degree to which we can empirically test the theory.

In contrast, formal theories are precise in their implementation as the mathematical notation or code in a computer programming language forces one to be specific about the structure of the theory (e.g., specifying the precise effect of one component on another).[2] The precision of formal theories allows us to deduce precisely how the target system will behave. This deduction can either be obtained analytically (e.g., from the mathematical equation) or computationally (e.g., through simulations from a computational model). For example, whereas the verbal theory of cooling only permitted a general sense of how the temperature will evolve over time, we can use the formal theory of cooling to predict the exact temperature of our object at any point in the future. Similarly, a formal implementation of the interpersonal theory of suicide would make highly specific predictions that could inform the prediction of suicide attempts (Millner et al., 2020). In other words, formal theories substantially strengthen surrogative reasoning, the very characteristic of scientific theories upon which we wish to capitalize.

The cognitive model of panic attacks described above is a verbal theory and is limited by the imprecision characteristic of most verbal theories. For example, in two recent articles, Fukano and Gunji (2012) and Robinaugh et al. (2020) independently proposed two distinct formal implementations of this theory: taking the verbal theory and expressing it in two sets of differential equations. Notably, these distinct implementations of the same verbal theory make divergent predictions about when panic attacks should occur, illustrating the limitations of failing to precisely specify the theory (for further detail, see Robinaugh et al., 2019; Robinaugh et al., 2021).

In this article, we will make extensive use of the formal theory proposed by Robinaugh and colleagues. A complete description of the generation of this theory can be found in the original article (Robinaugh et al., 2019). For our purposes here, it is sufficient to note that the aim in developing this model was to take extant verbal theories, especially cognitive behavioral theories, and express them in the language of mathematics. For example, Clark's verbal theory posits that a perception of threat can lead to arousal-related bodily sensations. However, the actual form and strength of this effect remain unspecified. In the mathematical model, we used a differential equation to precisely define this effect: $\frac{dA}{dt} = \alpha(\nu T - A)$. In this equation, there is a linear effect of perceived threat (T) on the rate of change of arousal (A), with the strength of this effect specified by the parameter ν. The product of ν and T is the value arousal is pulled toward: If νT is smaller than the current level of arousal, $\frac{dA}{dt}$ will be negative and arousal will *decrease* toward νT; if νT is greater than arousal, $\frac{dA}{dt}$ is positive and arousal *increases* toward νT. That is, arousal is pulled toward νT, which is a linear function of T. Each model component was defined as a differential equation in this way (see middle panel in Figure 2), providing a formal theory of panic disorder.

By specifying the structure of the theory in the language of mathematics, we are able to solve the system numerically, thereby deducing the theory's predictions about how the target system will behave. We were able to demonstrate, for example, that when the effect of arousal on perceived threat is sufficiently strong, the positive feedback between these components is sufficient to send the system into runaway positive feedback, producing the characteristic surge of arousal, perceived threat, and escape behavior that we refer to as a panic attack (see right panel in Figure 2). That is, we were able to *show*, rather than merely assert, that the theory can explain the phenomenon of panic attacks. As this example illustrates, formalizing theory substantially strengthens our ability to deduce theory-implied target system behavior. A full realization of a theory's usefulness thus all but requires that the theory be formalized. For that reason, our aim in psychiatric research should not merely be the construction of theories, but the construction of formal theories.

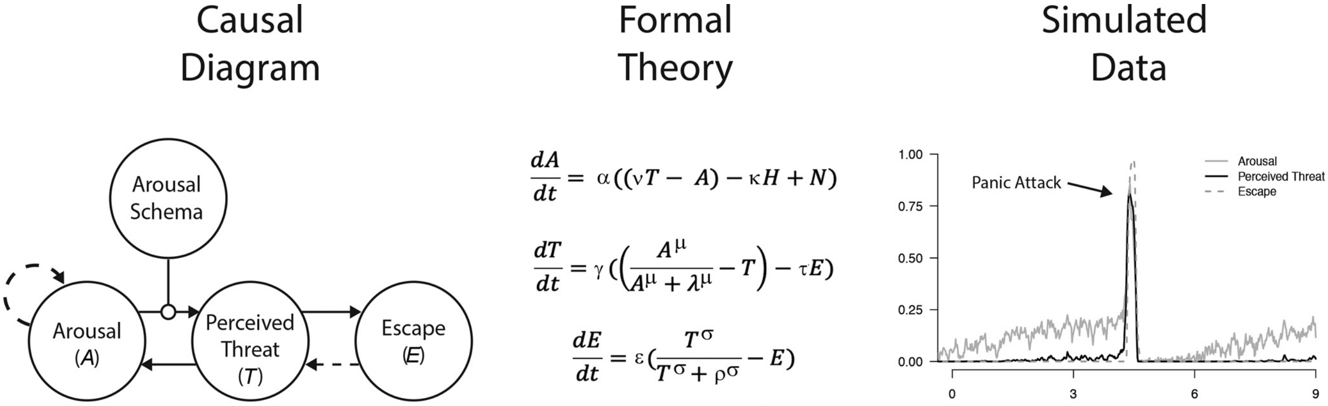
Data and Data Models

Our brief overview of the philosophy of science literature on theory suggests that if our aim is the explanation, prediction, and control of mental disorders, what we are after are well-developed formal theories: mathematical or computational models that represent the target system that gives rise to phenomena of interest. The key question then becomes: How can we best construct such a formal theory? The answer to this question will, of course, involve the collection and analysis of *data*. However, theories typically do not aim to explain data directly. Data are sensitive to the context in which they are acquired and subject to myriad causal influences that are not of core interest (Woodward, 2011). For example, panic disorder researchers collect data from diagnostic interviews, self-report symptom inventories, assessments of physiological arousal during panic attacks, time-series data, and a host of other methods. Data gathered using these methods will be influenced not only by

[2] It is, of course, possible to express verbal theories with the same level of precision as is provided by a mathematical equation (e.g., there are very few equations in the Principia, yet the laws Newton describes are not lacking in precision). Nonetheless, the specificity required by mathematics or computational programming makes them more amenable to expressing theories precisely and has the considerable practical advantage of supporting the derivation of predictions from the theory.

Figure 2

Causal Diagram, Formal Theory, and Simulated Data of the Formal Theory of Panic Disorder by Robinaugh et al. (2019)



$$\frac{dA}{dt} = \alpha((\nu T - A) - \kappa H + N)$$

$$\frac{dT}{dt} = \gamma\left(\left(\frac{A^{\mu}}{A^{\mu} + \lambda^{\mu}} - T\right) - \tau E\right)$$

$$\frac{dE}{dt} = \varepsilon\left(\frac{T^{\sigma}}{T^{\sigma} + \rho^{\sigma}} - E\right)$$

Note. The left panel displays the key components of the theory proposed by Robinaugh et al. (2019) at play during panic attacks: arousal, perceived threat, escape behavior, and arousal schema. The arrows indicate the direct causal relationships which are posited to operate between these components in the formal theory. The middle panel displays the formal theory that specifies the precise nature of the relations among these components. The top equation defines the rate of change in arousal $\frac{dA}{dt}$ where A is arousal, T is perceived threat, H is homeostatic feedback, and N is a noise variable representing fluctuations in arousal due to both internal and external stimuli. The rate parameter α specifies the intrinsic rate at which arousal can change, and the slope parameters ν and κ determines the strength of the effect of perceived threat and homeostatic feedback on arousal, respectively. The middle equation defines the rate of change in perceived threat $\frac{dT}{dt}$, which depends on arousal schema through the parameter λ. The state variable E denotes escape behavior. The rate parameter γ specifies the rate at which perceived threat can change, the parameters λ and μ together specify the strength of arousal on perceived threat, and the parameter μ specifies the strength of the effect of escape behavior on perceived threat. The final equation specifies the rate of change in escape behavior $\frac{dE}{dt}$ as a function of perceived threat, which is determined by the rate parameter ϵ and two parameters specifying the strength of perceived threat's effect on escape behavior: ρ and σ. The final panel on the right depicts the simulated behavior defined by these equations and, thus, implied by the theory.

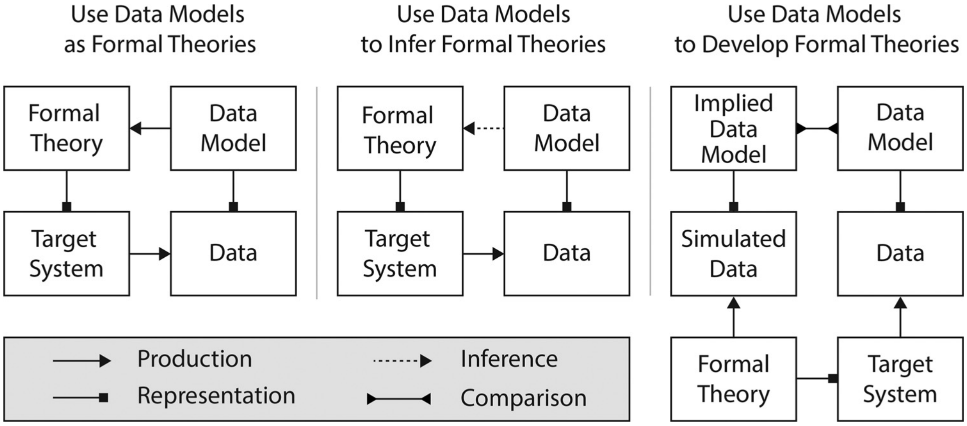
the experience of panic attacks, but also by recall biases, response biases, sensor errors, and simple human error. Accordingly, theories do not aim to account for specific "raw" data. Rather, theories explain phenomena identified through robust patterns in the data that cannot be attributed to the particular manner in which the data were collected (e.g., researcher biases, measurement error, methodological artifacts). To identify these empirical regularities in data, researchers use *data models*: representations of the data (Kellen, 2019; Suppes, 1962). Data models can take many forms. These can range from the most basic descriptive tools, such as a mean score, a correlation, or a fitted curve, to more complex statistical tools which are common in different areas of psychology and beyond, such as structural equation models (SEM), item response theory (IRT) models, time-series models, hierarchical models, network models, mixture models, loglinear models, and so forth. Essentially, we can consider a data model to be any descriptive statistic or statistical model that in some way summarizes the data.

Three Routes From Data Models to Formal Theories

Data models are ubiquitous in psychological research, most commonly appearing within the context of null-hypothesis significance tests. The question of how best to use data models to *test* formal theories is a critical one (Meehl, 1978, 1990), and one to which we will return later in this article. However, we will first focus on the question of how to use data models to *generate* and *develop* the kinds of formal theories that are ready to be subjected to rigorous testing. We see three possible routes researchers may take to move from data models to formal theories.

The first route arrives at theories directly by treating data models as formal theories. In this case, the transition from data model to formal theory is largely an act of interpretation. Instead of interpreting a data model as a representation of the data, we interpret it as a representation of the target system (see Figure 3, left panel). Specifically, the variables of the data model are treated as the components of our theory, and the statistical relationships among variables are treated as the structural relationships among the theory components. From this perspective, research is carried out by conducting an empirical study, estimating a data model, and treating the data model as a theory. If viable, this approach would be extremely powerful. A well-developed formal theory would be just one well-designed study away. Although we suspect that few researchers would explicitly endorse this approach, in practice, even seasoned researchers may fall victim to the tendency to interpret their statistical models as representations of the target system. For example, much recent debate in the psychopathology literature has focused on the interpretation of latent variable models. One can take the position that latent variables are simply data models: summaries of the covariances between items in the data and nothing more. Alternatively, one can interpret them as theoretical models, with latent variables representing some common underlying cause that *explains* the phenomenon of item covariances (Borsboom et al., 2003). Although many would likely endorse the former characterization, evidence of the latter can often be found in the description of factor analytic findings (e.g., when describing the identification of "underlying" factors that "account" for item covariance). Accordingly, we suspect that this route to theory may be more common than a mere show of hands would suggest (including in the network approach).

Figure 3
Three Routes From Data Models to Formal Theories



Note. The figure provides an overview of three routes to developing formal theories using data models. In the left panel, data models are treated as formal theories. In the middle panel, data models are used to draw inferences about the target system and, thereby, to generate formal theories of that system. In the right panel, data models used to develop formal theories by deducing implied data models and comparing them with empirical data models.

The second route arrives at formal theories by drawing direct inferences from data models (Figure 3, middle panel). That is, the data model is not directly treated as a theory, but rather as a kind of direct inference tool. From this perspective, research is carried out by conducting an empirical study, estimating a data model, and using the data model to infer characteristics of the target system, thereby informing the development of a theory. For example, one could observe a conditional dependence relationship between two variables and infer the presence of a causal relationship between the corresponding components in the target system. Multiple linear regression techniques—which statistically control for many covariates not of primary interest—are often used in this way, though interpretations of parameters themselves as being causal in nature are often studiously avoided (Grosz et al., 2020; Rohrer, 2018). Although this strategy is perhaps the most difficult to study or even define, since it relies on often undefined inference rules for particular data models, we suspect it is the most common approach to informing theory generation in many areas of psychology.

The third route arrives at formal theories through comparison between theory-implied data models (i.e., the data model predicted by our theory) and empirical data models. Some version of this approach is common in areas of psychology with well-developed traditions of formal theory (e.g., mathematical psychology and cognitive psychology), but is rarely applied outside these areas. In this route, research is carried out by first generating an initial formal theory. From this initial formal theory, we simulate data which can then be used to obtain a theory-implied data model. We can then compare the implied data model with the empirical data model, and adapt the formal theory if there are meaningful discrepancies between the two. This route thus relies upon the "immense deductive fertility" of formal theories (Meehl, 1978, p. 825) to make precise predictions about what data models we should expect to observe in our empirical data. By comparing these theory-implied data models to data models derived from empirical data, we can inform how the theory should be revised to be brought in line with empirical data. In other words, in this route, formal theory is not only the ultimate goal of the research process, it also plays a central role in theory development.

The three routes outlined here capture distinct ways in which researchers may use data models to inform formal theories. However, a key question remains: Which of these three strategies is most appropriate? Which will best help us achieve our aim of constructing well-developed formal theories that are sufficiently good representations of the target system that they support explanation, prediction, and control? The answer to this question is likely to be context specific, depending on the target system, the level(s) on which we aim to have such a theory, and the data and data models which are available to us. Accordingly, in the next section, we will focus on answering a more tractable question: Which route to formal theory is likely to be most fruitful within the broad theoretical framework of conceptualizing mental disorders as complex systems?

Evaluating Three Routes From Data Models to Formal Theories

We will evaluate the three routes from data models to formal theories with a focus on three data models that have become popular within the network approach to psychopathology: the Ising model, the Gaussian graphical model (GGM), and the vector autoregressive (VAR) model. We use these data models due to their popularity in the network approach literature and because they are broadly representative of—and share close connections to—the linear models typically used by applied researchers.

Route 1: Using Data Models as Formal Theories

The first route from data model to formal theory suggests that data models can themselves serve as formal theories. For this to be the case, the properties of those data models must be able to represent the properties we expect in the target system. Accordingly, to evaluate the first route (Figure 3, left panel), we must first outline the properties we expect in our target systems when working from the complex systems perspective and then evaluate whether these properties are captured by our data models of interest.

Properties of Mental Disorder Target Systems

Adopting the perspective that mental disorders arise, in part, from a complex network of interacting symptoms, there are a number of properties we would expect to be present in the target system. First, *feedback loops* among components are likely present. Researchers have frequently posited "vicious cycles," where the initial activation of one component (e.g., arousal) elicits activation of other components (e.g., perceived threat) which, in turn, further amplifies the activation of the original component. Second, causal effects between components are likely to be asymmetrical. That is, the effect of Component A on Component B may differ from the effect of Component B on Component A. For example, it is unlikely that concentration has the same effect on sleep as sleep has on concentration or that compulsions have the same effect on obsessions that obsessions have on compulsions. Third, interactions among components are likely to occur at *different time scales*. For example, the effect of intrusive memories on physiological reactivity in posttraumatic stress disorder is likely to occur on a time scale of seconds to minutes, whereas an effect of energy on depressed mood may play out over the course of hours to days, and the effect of appetite on weight gain may occur on a time scale of days to weeks. Fourth, it is likely that there are *higher order interactions* among components. For example, the presence of sleep difficulties may strengthen the effect of feelings of worthlessness on depressed mood or the effect of intrusive trauma memories on physiological reactivity. If data models are to serve as formal theories of the target system, they must be able to represent these types of causal structures.

We would further suggest that most, perhaps all, mental disorder target systems are likely to have *multiple stable states*. That is, multiple states into which the system can settle and remain in the absence of external perturbation. In the simplest case, the system can be characterized by the presence of two stables states: an unhealthy state (e.g., a state of elevated symptom activation, such as a depressive episode), and a healthy state (e.g., a state without elevated symptom activation). In other cases, there may be multiple stable states (e.g., healthy, depressed, and manic states in bipolar disorder). The presence of multiple stable states is, in turn, likely to be accompanied by other behavior often observed in mental disorders, including spontaneous recovery and sudden shifts into or out of a state of psychopathology, further suggesting that a model of any given mental disorder will almost certainly need to able to produce alternative stable states.

Comparing Target System Properties With Data Model Properties

The first model we will consider is the VAR model. The VAR model for multivariate continuous time-series data linearly relates each variable at time point t to all other variables and itself at previous time points (Hamilton, 1995), typically the time point immediately prior $t - 1$ (i.e., a first order VAR, or VAR(1), model; e.g., Bringmann et al., 2013; Fisher et al., 2017; Groen et al., 2020; Pe et al., 2015; Snippe et al., 2017). The estimated lagged effects of the VAR models indicate conditional dependence relationships among variables over time. The dynamics of the model is such that the variables are perturbed by random input (typically Gaussian noise) and the variables return to their means, which represent the single stable state of the system.

As depicted in Figure 4, the VAR model is able to represent some key characteristics likely to be present in mental disorder target systems. Most notably, it allows for feedback loops. Variables can affect themselves both directly (e.g., $X_t \rightarrow X_{t+1}$), or via their effects on other variables in the system (e.g., $X_t \rightarrow Y_{t+1} \rightarrow X_{t+2}$). The VAR model also allows for asymmetric relationships, because the effect $X_t \rightarrow Y_{t+1}$ does not have to be the same effect as $Y_t \rightarrow X_{t+1}$ in sign or magnitude. However, because the lag-size (i.e., the distance between time points) is fixed and consistent across all relationships, the VAR model does not allow for dynamics which unfold at different time scales. Moreover, because the VAR model only includes relations between pairs of variables, it is unable to represent higher-order interactions involving more than two variables. Finally, the VAR model has a single stable state defined by its mean vector and thus cannot represent multiple stable states of a system, such as a healthy state and unhealthy state.

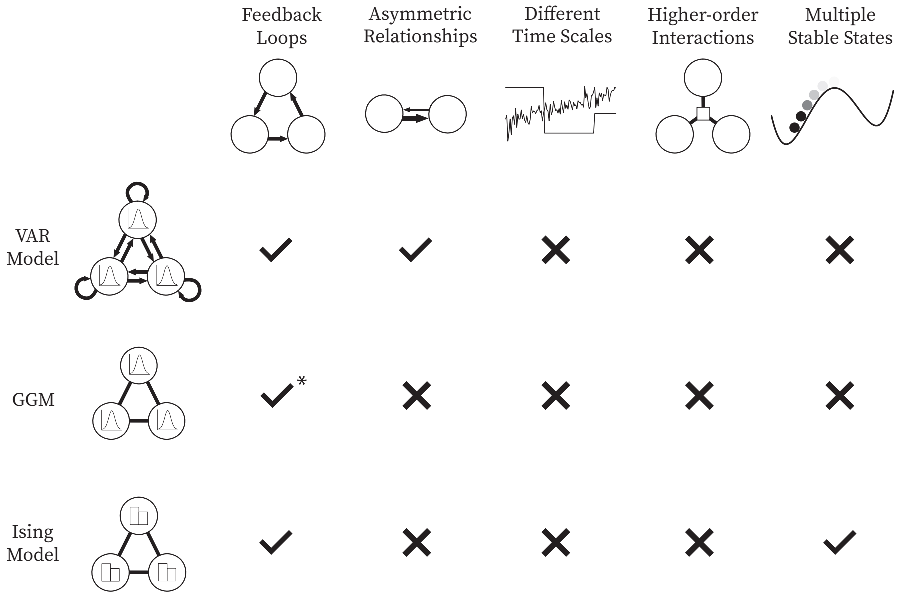
The second model we will consider is the GGM, which linearly relates pairs of variables in either cross-sectional (Haslbeck & Fried, 2017) or time-series data (Epskamp et al., 2018). In the case of time-series data the GGM models the relationships between variables at the same time point. Because it does not model any dependency across time, it is typically not considered a dynamic model and, thus, could not be used to represent the behavior of a mental disorder target system as it evolves over time. In principle the GGM could be augmented by a dynamic rule similar to one commonly used with the Ising model (i.e., "Glauber dynamics"; see below). However, in that case, the GGM would become a model similar to, but more limited than, the VAR model described above (e.g., it would be limited to symmetric relationships). Accordingly, the GGM is similarly unable to represent key features we expect to observe in a mental disorder target system.

The final model we will consider is the Ising model. The Ising model again represents pairwise conditional dependence relations between variables (Ising, 1925); however, it is a model for multivariate binary data. Although the original Ising model does not model dependencies over time, it can be turned into a dynamic model by augmenting it with Glauber dynamics (Glauber, 1963).[3] Like the VAR model, the Ising model is able to represent feedback loops. Moreover, due to its nonlinear form it is able to exhibit multiple stable states (and the behavior that accompanies such stable states, such as hysteresis and sudden shifts in levels of symptom activation, see e.g., Cramer et al., 2016; Dalege et al., 2016; Haslbeck et al., 2020; Lunansky et al., 2020). It is perhaps not surprising then, that the Ising model is used as a theoretical model across many sciences (Stutz & Williams, 1999), and to our knowledge, is the only of the three data models examined here that has been

[3] This dynamics works as follows: After specifying an initial value for each variable, it randomly picks one variable X_i at $t = 1$ and takes a draw from the distribution of X_i conditioned on the values of all other variables. This value (either 0 or 1) is set to be the new value of X_i and then the same process is repeated, thereby allowing the model to evolve over time.

Figure 4

Ability of Popular Network Models to Capture Key Properties of Mental Disorders



	Feedback Loops	Asymmetric Relationships	Different Time Scales	Higher-order Interactions	Multiple Stable States
VAR Model	✓	✗	✗	✗	✗
GGM	✓*	✗	✗	✗	✗
Ising Model	✓	✗	✗	✗	✓

Note. The figure shows whether the five properties of mental disorders discussed above can be represented by the three most popular network data models, the VAR(1) model, the GGM, and the Ising model with Glauber dynamics. Note that there is a check mark at feedback loops for GGMs because one could in principle endow the GGM with a dynamic similar to the Ising model, which would essentially lead to a restricted VAR model but with symmetric relations. The asterisk is present because this endowment of dynamics is not typically done in practice.

used as a formal theory of a mental disorder target system (Cramer et al., 2016). Unfortunately, the Ising model falls short in its ability to represent the remaining characteristics likely to be present in mental disorders. The relationships in the Ising models are exclusively symmetric; with the standard Glauber dynamics, there is only a single time scale; and the Ising model includes exclusively pairwise relationships, precluding any representation of higher-order interactions.

Data Models as Formal Theories of Mental Disorders?

We showed that the VAR, GGM, and Ising models are unable to represent most key properties we would expect in the target systems giving rise to mental disorders, and therefore cannot serve as formal theories for those disorders. In the statistical inference literature the problem of not being able to represent the target system would be seen as a problem of model *misspecification*. In the present case, this would mean that the data models are misspecified with respect to the target system (i.e., the data generating system).

Of course, more complex models would be able to produce more of the characteristics likely to be present in mental disorders. For example, one could extend the VAR model with higher-order interactions (e.g., $X_t \times Y_t \rightarrow X_{t+1}$) or latent state variables (Hamaker et al., 2010; Tong & Lim, 1980), thereby allowing it to represent multiple stable states. However, estimating data models is subject to fundamental constraints. More complex models require more data which are often unavailable in psychiatric research. For example, around 90 observations (about 2.5 weeks of a typical Experience Sampling Method [ESM] study) are needed for a VAR model to outperform the much simpler AR model (Dablander et al., 2019). Models more complex than the VAR model would require even more data to be estimated reliably.

Another constraint likely to be present in many psychological studies is the sampling frequency (e.g., measurement every 2 hr), which may be too low to capture the structure of the target system of interest (Haslbeck & Ryan, 2021). In this situation, a data model still contains some information about the target system, but cannot capture the structure of the target system to the extent that it can serve as a formal theory. Even where large amounts of high frequency data are available, efforts to estimate more complex models may be constrained by the simple fact that it is often unclear how such models can be estimated. For example, one could extend the Ising model with a second time scale (e.g., Lunansky et al., 2020), but it would be unclear how to estimate such a model from data. Finally, even where more complex models can be estimated, those models are often uninterpretable. For example, nonparametric models (e.g., splines; Friedman et al., 2001, p. 139), which can capture extremely complex behavior, typically consist of thousands of parameters, none of which can be interpreted individually. Accordingly, it is unlikely that any data model estimated from the type of data typically available in psychiatric research will be both interpretable and capable of capturing the characteristics of psychopathology in such a way that would allow it to serve as a formal theory of a mental disorder.

Route 2: Using Data Models to Infer Formal Theories

An alternative route from data models to formal theories is to use data models to draw inferences about a target system, inferences that we can use to construct a formal theory. There is good reason to think that this approach could work. Because the data are generated by the target system, and data models summarize these data, the parameters of any data model certainly *somehow* reflect characteristics of the target system. This means that it should be possible, in principle, to infer something about the target system and its characteristics from data and data models. Although we have seen already that the GGM, Ising and VAR models cannot directly reproduce the key characteristics of the target system, their parameters could potentially still yield insights into the structure or patterns of relationships between components. In line with this intuition, it has frequently been suggested that the GGM, the Ising model, and the VAR models can serve as "hypothesis-generating tools" for the causal structure of the target system (e.g., Borsboom & Cramer, 2013; Epskamp et al., 2018; Epskamp et al., 2018; Fried & Cramer, 2017; Jones et al., 2018; van Rooijen et al., 2017).

Although this approach seems intuitive, in practice it is unclear how exactly this inference from data model to target system should work. For example, if we observe a strong negative linear cross-lagged effect of X_t on Y_{t+1} in a VAR model, what does that imply for the causal relationship between the corresponding components in the target system? A precise answer to this question would require a rule that connects parameters in particular data models to the structure of the target system. For some simple systems, such a rule is available, and this type of inference can broadly be characterized as a causal discovery problem (Spirtes et al., 2000; Peters et al., 2017). For example, if the target system can be represented as a directed acyclic graph (DAG), then under certain circumstances its structure can be discovered from conditional (in)dependence relations between its components: Conditional independence implies causal independence, and conditional dependence implies either direct causal dependence or a common effect (Pearl, 2009; Ryan et al., 2019). However, this kind of precise deduction is not possible for the types of nonlinear dynamic systems we expect in a psychiatric context (although Mooij et al., 2013 and Forré & Mooij, 2018 have established some links in this regard). In these contexts, any inference from data model to target system must rely on some simplified heuristic(s) in an attempt to approximate the link between the two. Critically, however, the extent to which such heuristics are informative remains unclear.

In this subsection we evaluate whether the three data models introduced above can be used to make inferences about mental disorder target systems. To do this, we treat the panic model discussed in above as the data-generating target system and compare the causal structure inferred from the data models to the true causal structure. To yield these inferences we use a very simple and intuitive set of heuristics: (a) if two variables are conditionally dependent in the data model, we will infer that the corresponding components in the target system are directly causally dependent; (b) if there is a positive linear relationship, we will infer that the causal relation between the corresponding components is positive (i.e., reinforcing); (c) if there is a negative linear relationship, we will infer that the causal relationship among components is negative (i.e., suppressing).

Inferring the Panic System From Network Data Models

To be able to evaluate the success of the simple heuristics described above, we must first represent the structure of the panic model (introduced earlier) in the form of a square matrix, that is, in the same form as the parameters of the VAR, GGM, and Ising models. Because the relationships between components are formalized through *differential equations*, a natural choice is to represent the panic model as a network of moment-to-moment dependencies, drawing an arrow $X \rightarrow Y$ if the rate of change of Y is directly dependent on the value of X (known as a *local dependence graph*; Didelez, 2007; Ryan & Hamaker, 2021). Figure 5a displays these moment-to-moment dependencies. Note that this structure cannot capture many aspects of the true model, such as the presence of two time scales or the moderating effect of arousal schema (AS; see above for details). It is, thus, already clear that the models cannot recover the *exact* causal structure of the panic model. Nonetheless, we can still investigate whether applying the simple heuristics to these three data models allows us to infer this less detailed pattern of direct causal dependencies.

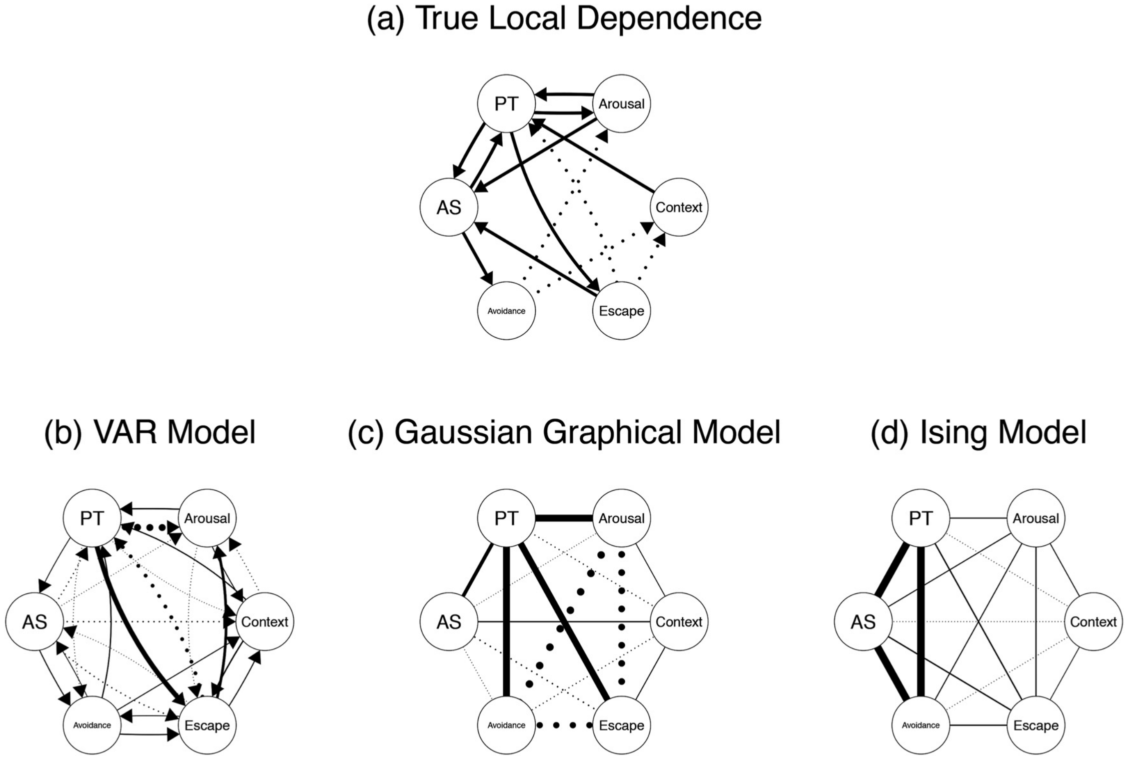
To evaluate how well these heuristics work, we compare this true causal structure to the causal structure inferred based on the three data models. To obtain the three data models, we first generate data from the target system (see Appendix A). Specifically, we use 4 weeks of minute-to-minute time-series data for 1,000 individuals. These individuals differ in their initial value of AS, with the distribution chosen so that the proportion of individuals for whom a panic attack is possible was equivalent to the lifetime history prevalence of panic attacks in the general population (R. R. Freedman et al., 1985). For the VAR model analysis, we create a single-subject experience-sampling-type dataset by choosing the individual who experiences the most (16) panic attacks in the 4-week period. To emulate ESM measurements for use with this model, we divide the 4-week period into 90-min intervals, taking the average of each component in that interval, yielding 448 measurements. For the GGM analysis, we emulate continuous cross-sectional measurements by taking the mean of each component for each individual over the four weeks. For the Ising model analysis, we emulate cross-sectional binary measurements by taking a median split of those same variables. The resulting VAR, GGM, and Ising model networks are displayed in Figure 5, panels b, c, and d, respectively.[4]

We will focus our evaluation on two important causal dependencies in the target system: the positive (i.e., reinforcing) moment-to-moment feedback loop between perceived threat (PT) and arousal, and the positive effect of AS (i.e., beliefs that arousal-related bodily sensations are dangerous, AS) on avoidance (i.e., efforts to avoid situations or stimuli that may elicit panic attacks). In the VAR model (panel b in Figure 5) we see a lagged positive relationship of arousal to PT, a strong *negative* lagged relationship from PT to arousal, and a weak positive effect of AS on avoidance. Applying the heuristics, we would infer a reinforcing relationship from arousal to PT, a suppressing relationship from PT to arousal, and a reinforcing effect of AS on avoidance. In the GGM (panel c in Figure 5) we see a positive conditional dependency between arousal and PT, but we also see a weak negative dependency between AS and avoidance. Applying the heuristics to the GGM, we would infer a reinforcing relationship between arousal and PT, and a suppressing relationship between AS and avoidance. Finally,

[4] Note that in the Ising model the parameter estimates are somewhat unstable due to near-deterministic relationships between some binarized variables.

Figure 5

Comparing Network Models With True Local Dependence Network



Note. Panel a shows the true model in terms of local dependencies between components (AS = arousal schema; PT = perceived threat); Panel b shows the VAR model estimated from ESM data sampled from the true model (VAR = vector autoregressive); Panel c shows the the Gaussian graphical model (GGM) estimated from the cross-sectional data of 1,000 individuals, generated from the true model; Panel d shows the Ising model estimated on the same data after being binarized with a median split. Solid edges indicate positive relationships, dotted indicate negative relationships. For Panels b to d, the widths of edges is proportional to the absolute value of the corresponding parameter. Note that in Panel b we do not depict the estimated auto-regressive parameters as the primary interest is in inferring relationships between variables.

in the Ising model (panel d in Figure 5), we see a strong positive dependency between AS and avoidance, and a very weak positive relationship between PT and arousal. This leads us to infer two reinforcing relationships, between arousal and PT, and AS and avoidance.

For the VAR model, the heuristics yield one correct and one incorrect inference. For the GGM, we make exactly the opposite inferences, with again one correct and one incorrect. The Ising model yields two correct inferences. However, inspecting the rest of the Ising model edges we can see a variety of incorrect inferences about other relationships, with independent components in the target system connected by strong edges in the Ising model, and the valences of various true dependencies flipped. At best, we can say that in each of the three network models, some dependencies do reflect the presence and/or direction of direct causal relationships, and some do not. Unfortunately, it is not possible to distinguish which inferences are trustworthy and which are not without knowing the target system, and in any real research context, the target system will be unknown. Consequently, these data models and simple heuristics cannot be used to reliably draw inferences about the target system.

The Mapping Between Data Model and Target System

Importantly, our inability to draw accurate inferences from these data models is not a shortcoming of the data models themselves. Each data model correctly captures some form of statistical dependency between the components in a particular domain (e.g., lagged 90-min windows). The scenario we emulated in this section is highly idealized in that we have directly and accurately observed all components of the target system—measurements are taken without error, and there is no potential for statistical dependencies to be produced by unobserved confounding variables. This means that the statistical dependencies in the data models can only be produced by causal dependencies in the target system. Thus, we know that there is *some* mapping from the causal dependencies in the target system to statistical dependencies in the data model. The fundamental barrier to inference is that the form of this mapping is unknown and considerably more complex than the simple heuristics we have used to draw inferences here. For example, consider the relationships between PT and arousal. The VAR model (panel b in Figure 5b), identifies a negative lagged relationship from PT to arousal in the data generated by the target system. Yet in the target system, this effect is positive. This "discrepancy" occurs because of a very specific dynamic between these components: After a panic attack (i.e., a brief surge of PT and arousal) there is a "recovery" period in which arousal dips below its mean level for a period of time. As a result, when we average observations over a 90-min window, a high average level of PT is followed by a low average level of arousal whenever a panic attack occurs. That same property of the system produces the observed findings for the GGM and Ising model through yet another mapping (for details, see Appendix B).

As this example illustrates, the mapping between target system and data model is intricate, and it is unlikely that any simple heuristics can be used successfully to work backward from the data model to the exact relationships in the target system. We can expect this problem to arise whenever we use relatively simple statistical models to directly infer characteristics or properties of a complex system (cf. the problem of underdetermination or indistinguishability; Eberhardt, 2013; Spirtes, 2010). Indeed, the same problem arises even for simpler dynamical systems when analyzed with more advanced statistical methods (e.g., Haslbeck & Ryan, 2021).

Of course, in principle, it must be possible to make valid inferences from data and data models to some properties of a target system using a more principled notion of how one maps to the other. For example, under a variety of assumptions, it has been shown that certain conditional dependency relationships can potentially be used to infer patterns of local causal dependencies in certain types of dynamic system (Bongers & Mooij, 2018; Forré & Mooij, 2018; Mooij et al., 2013). However, the applicability of these methods to the type of target system we expect to give rise to psychopathology is as yet unclear and even under the strict assumptions under which they have been examined, these methods still do not recover the full structure of the target system. Given the considerations reviewed here, the intricate mapping between target system and data model likely precludes reliable direct inferences about the kinds of target systems we are likely to see in mental health research. Accordingly, we cannot rely on this kind of inference to build formal theories. An alternative approach is needed.

Route 3: Using Data Models to Develop Formal Theories

In the previous section, we saw that the mapping between target system and data model is intricate and would be nearly impossible to discern when the target system is unknown. However, we also saw that when the target system is known, we can determine exactly which data models the target system will produce. Indeed, this is precisely what we did when we simulated data and fit data models to it in the previous section. In this section we consider a third route to formal theories, which makes use of this ability to determine which data models are implied by a given target system (or formal theory).

This third route works as follows. First and foremost, we must propose *some* initial formal theory which we take as a representation of the target system. The quality or accuracy of this theory may be good or bad, but, crucially, the theory must be formalized in such a way as to yield unambiguous predictions. Second, we can use this initial theory to deduce a theory-implied data model. This can be done by simulating data from the formal theory and fitting the data model of interest. Third, by comparing implied data models with their empirical counterparts, we can learn about where the theory falls short and adapt the formal theory to be more in line with empirical data. This approach is represented in schematic form in the right-hand panel of Figure 3. It can be seen as a form of inference. Not the direct deductive inference focused on in the previous subsection, but rather *abductive inference*: inference to the best explanation (Haig, 2005). We infer the best explanation for any discrepancies between empirical and theory-implied data models, and use those inferences to inform subsequent theory development.

Obtaining Theory-Implied and Empirical Data Models

In this section, we will treat the panic model introduced in earlier as our initial formal theory, which aims to represent the target system that gives rise to panic disorder (Figure 6, bottom row). The panic model can be used to simulate data and, in turn, to derive predictions made by the theory in the form of theory-implied data models (left-hand column of Figure 6). Although many data models can (and should) be used to compare theories and empirical data, here we will examine the implied cross-sectional Ising model of the three core panic disorder symptoms: (a) recurrent panic attacks (PA); (b) persistent concern (PC) following a panic attack; and (c) avoidance (Av) behavior following a panic attack (American Psychiatric Association, 2013). If our formal theory of panic disorder is an accurate representation of the target system that gives rise to panic disorder, the implied Ising model derived from this theory should be in agreement with a corresponding Ising model estimated from empirical data. Accordingly, any discrepancies between these models call for—and can inform—further development of the theory.

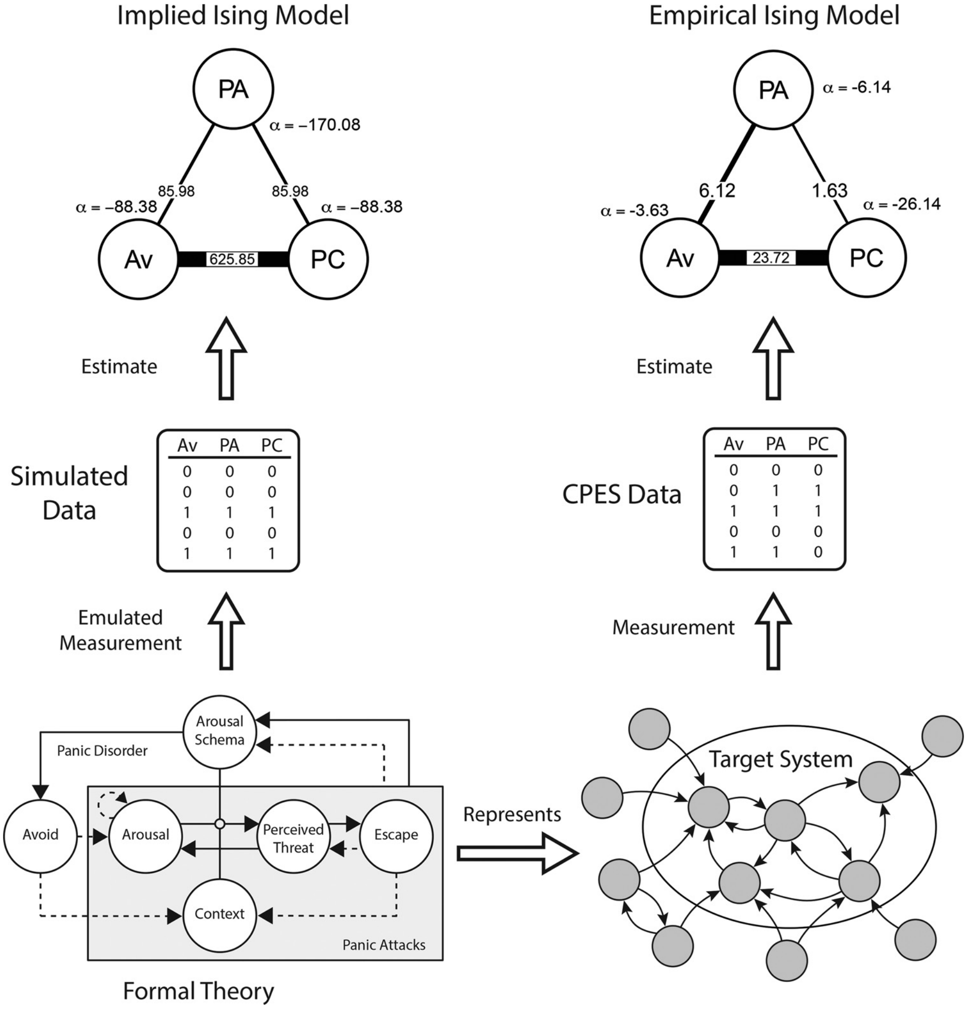
Notably, obtaining an implied data model requires not only a formal theory from which we can simulate data, but also a formalized process by which variables are "measured" from those data. The panic model generates intraindividual time-series data for multiple individuals (as described in Appendix A). We therefore need to define how cross-sectional symptom variables can be extracted from those time-series. Here, we specified that recurrent panic attacks (PA = 1) are present for an individual in our simulated data if there are more than three panic attacks in the one month observation period. PC was determined using the average levels of jointly experienced arousal and perceived threat (i.e., anxiety) outside the context of a panic attack. If an individual had a panic attack, *and* their average anxiety following a panic attack exceeds a threshold determined by "healthy" simulations (i.e., those without panic attacks), they are classified as having PC (PC = 1). We similarly defined Av as being present if an individual has a panic attack, *and* their average levels of avoidance behavior following that attack were higher than we would expect to see in the healthy sample. A more detailed account of how we generated these data can be found in Appendix C. This simulated cross-sectional data was then used to estimate the theory-implied Ising model (top left-hand corner, Figure 6).[5]

We obtained the corresponding empirical Ising model (right-hand column of Figure 6) using the publicly available Collaborative Psychiatric Epidemiology Surveys (CPES) 2001–2003 (Alegria et al., 2007). The CPES is a nationally representative survey of mental disorders and correlates in the United States, with a total

[5] Here again the Ising model estimates are somewhat unstable due to near-deterministic relationships between the variables.

Figure 6

Illustration of the Third Route From Data Models to Formal Theories



Note. We take the Panic model discussed in previously as our formal theory, representing the unknown target system that gives rise to panic disorder. To obtain an implied data model from this theory, we first formalize how the components of the theory produce the data of interest, emulating the measurement process. With this in place, we can simulate data from the model in the form of cross-sectional binary symptom variables. We obtain the theory-implied Ising model by estimating it from these simulated data (top-left corner). To estimate the empirical Ising model (top-right corner) we make use of empirical measurements of binary symptom variables from the CPES dataset. PA = recurrent panic attacks; PC = persistent concern; Av = avoidance; CPES = Collaborative Psychiatric Epidemiology Surveys.

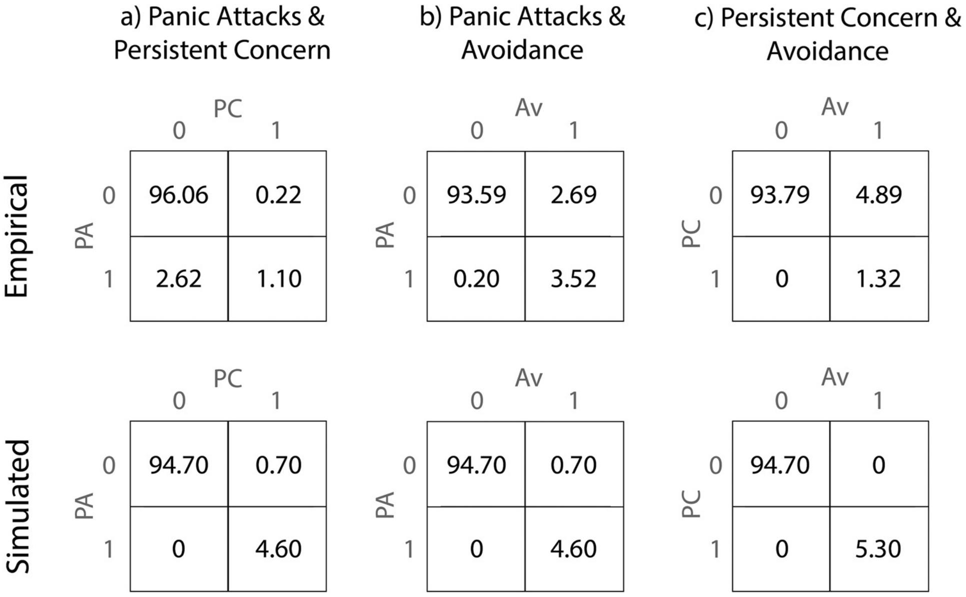
sample size of over twenty thousand participants (of which n = 11,367 are used in the current analysis; for details see Appendix C). The CPES combines more than 140 items relating to panic attacks and panic disorder, with a diagnostic manual describing how these items can be recoded into binary symptom variables reflecting recurrent PA, PC, and Av. Recurrent PA are present if the participant reported more than three lifetime panic attacks. PC is present if, following an attack, the participant experienced a month or more of persistent concern or worry. Avoidance is present if the participant reports either a month of avoidance behavior following an attack, or a general avoidance of activating situations in the past year. Note that these definitions correspond closely to the formalized measurement assumptions we made while generating our theory-implied data model.

Theory Development: Comparing Model-Implied and Empirical Data Models

As seen in Figure 6, there is a similar pattern of conditional dependencies in the implied and empirical data models. In both, all pairwise dependencies are positive, and all thresholds are negative. There is also a similar ordering of conditional dependencies in terms of their magnitude. Within each model, the conditional

Figure 7

Comparing Pairwise Frequencies of Symptoms in Empirical and Simulated Data



a) Panic Attacks & Persistent Concern

Empirical

PA \ PC	0	1
0	96.06	0.22
1	2.62	1.10

Simulated

PA \ PC	0	1
0	94.70	0.70
1	0	4.60

b) Panic Attacks & Avoidance

Empirical

PA \ Av	0	1
0	93.59	2.69
1	0.20	3.52

Simulated

PA \ Av	0	1
0	94.70	0.70
1	0	4.60

c) Persistent Concern & Avoidance

Empirical

PC \ Av	0	1
0	93.79	4.89
1	0	1.32

Simulated

PC \ Av	0	1
0	94.70	0
1	0	5.30

Note. Contingency tables showing percentages for each pair of symptom variables (one per column) for the empirical data (top row) and simulated data (bottom row). The CPES contingency tables are based on n_{CPES} = 11,367 observations. The simulated dataset contains n_{sim} = 1,000 observations. PA = recurrent panic attacks; PC = persistent concern; Av = avoidance.

relationships of recurrent PA with Av and recurrent PA with PC are of the same order of magnitude, and the conditional relationship between Av and PC is an order of magnitude greater. However, we also see some differences between the models. First, the absolute value of pairwise dependencies and thresholds are much greater in the implied Ising model (Figure 6a) than the empirical Ising model (Figure 6b). Second, we see that the relationships in the implied model are perfectly symmetric, with exactly the same thresholds for Av and PC, and precisely the same weights relating recurrent PA to both.

The bivariate contingency tables in Figure 7 provide further information about these intersymptom relationships. In both the implied and empirical data models only a small proportion of individuals experience recurrent PA (empirical 3.72%, simulated 4.6%). In the simulated dataset, the symptom relationships are almost deterministic: If one symptom is present, so too are all others, and vice versa for the absence of symptoms (apart from seven individuals who experience at least one, but less than three panic attacks in the time window). This is because there is a deterministic relationship between the components underlying these symptoms in the panic model: All participants who experience one panic attack have PC and Av behavior after those attacks. In contrast, there are nondeterministic relationships in the empirical data. For example, it is actually more common to have recurrent PA without PC than with PC (Column a of Figure 7). Similarly, more individuals experience Av without PC, than with PC (Column c of Figure 7). Conversely, there are no individuals who experience PC but not Av.

Having observed these differences between the theory-implied and empirical data models, our task is to consider the best explanation for the discrepancies. This explanation could rest at any step in the process from formal theory to the implied data model or from the target system to the empirical data model (i.e., any of the paths illustrated in Figure 6). It could be the case that any discrepancies we have observed here are due to inaccuracies in how we emulated the measurement process.[6] For example, perhaps PC and Av co-occur equally, but the former suffers from a greater degree of recall bias than the latter (for an example of differential symptom recall bias in depressed patients, see Ben-Zeev & Young, 2010). The discrepancies could have arisen due to the somewhat different time scales at which the simulated and empirical symptoms are defined. The simulated symptoms are defined over a 1-month period whereas the CPES items are defined over lifetime prevalence. Due to the deterministic nature of the panic model, we regard a 1-month period to be a good approximation for lifetime experience of panic symptoms in this case. Nonetheless, it is a discrepancy in measurement that could lead to discrepancies between the implied and empirical data models. It could also be that the discrepancy is due to estimation issues. However, due to the large sample sizes and simple models used, we suspect it is unlikely that sampling variance is a problem in this instance.

We consider the most likely explanation for the observed discrepancies to lie with the theory itself, thereby providing an opportunity to consider how the theory might be further developed to bring it in line with empirical data. For example, the implied Ising

[6] Inaccurate conceptualizations of how measurements represent the target system will be problematic for any approach to theory development or indeed any scientific endeavor, as evidenced by the growing attention on measurement in the psychopathology literature (e.g., Flake & Fried, 2019)

model overestimates the strength of intersymptom relationships relative to the empirical Ising model. This can largely be explained by the deterministic causal effects in the theory. In the simulated data, everybody who experiences Recurrent PA also develops PC and, in turn, Av. As seen in Figure 7, this is inconsistent with empirical data. To improve the model, we must include some mechanism by which individuals can experience a panic attack without developing the remaining symptoms of panic disorder. Where is such a mechanism most appropriate? In the empirical data, nearly all those with PC also exhibit Av. In contrast, only a few of those with recurrent PA also exhibit PC. Accordingly, the discrepancy seems most likely to arise from the effect of recurrent panic attacks on PC. Incorporating a mechanism that leads some to resist developing persistent concern following panic attacks (e.g., high perceived ability to cope with the effects of a panic attack) would allow the theory to better account for both the observation that some individuals experience recurrent panic attacks without developing the full panic disorder syndrome and the observation that those who do develop PC tend to develop the full syndrome.

As this example illustrates, discrepancies between the theory-implied and empirical data models can provide insights for how to further develop a theory. We have focused on just one set of discrepancies between our theory-implied and empirical data models. More insights may be gained by focusing on others and these insights can work together to triangulate on the most appropriate set of revisions. Further insights can almost certainly be gained by considering additional data models and different types of data. For example, experimental data or time series data on the relation between arousal and PT may allow us to refine the specification of the feedback between those two variables. In general this route offers a great deal of flexibility in theory development. Although the theory is likely to be complex, dynamic, and nonlinear, the form of the data models used to learn about that theory need not be. Instead, by starting with an initial theory, the researcher can use any data about the phenomena of interest to further develop that theory.

Abductive Formal Theory Construction

In previous section, we illustrated an approach to using empirical data to develop an existing formal theory. However, our description of this approach is not, by itself, a comprehensive approach to theory construction, because it does not address how the formal theory was generated nor how it should ultimately be tested. In this section, we propose a three-stage framework for formal theory construction built around using empirical data and data models for formal theory development (see Figure 8).

This framework uses the theory construction methodology proposed by Borsboom et al. (2020) as a foundation and seeks to extend that work by providing more concrete guidance for how to proceed with the generation, development, and testing of theories, with a focus on how data models are used at each of these stages of theory construction. In the *theory generation* stage, we use data models to establish the phenomenon to be explained, generate an initial verbal theory, and formalize that theory. In the *theory development* stage, the theory is revised and improved by repeatedly comparing theory-implied data models with data models from empirical data. Finally, in the *theory testing* stage, the theory is subjected to strong tests within a hypothetico-deductive framework, comparing precise theory-driven predictions and empirical data models with the aim of corroborating or refuting the theory. The approach to theory construction proposed here places considerable emphasis on the theory's ability to explain phenomena and emphasizes the importance of abductive inference in theory construction (Haig, 2005). We therefore refer to it as the abductive formal theory construction (AFTC) framework.

Stage 1: Generating Theory

Establish the Phenomena

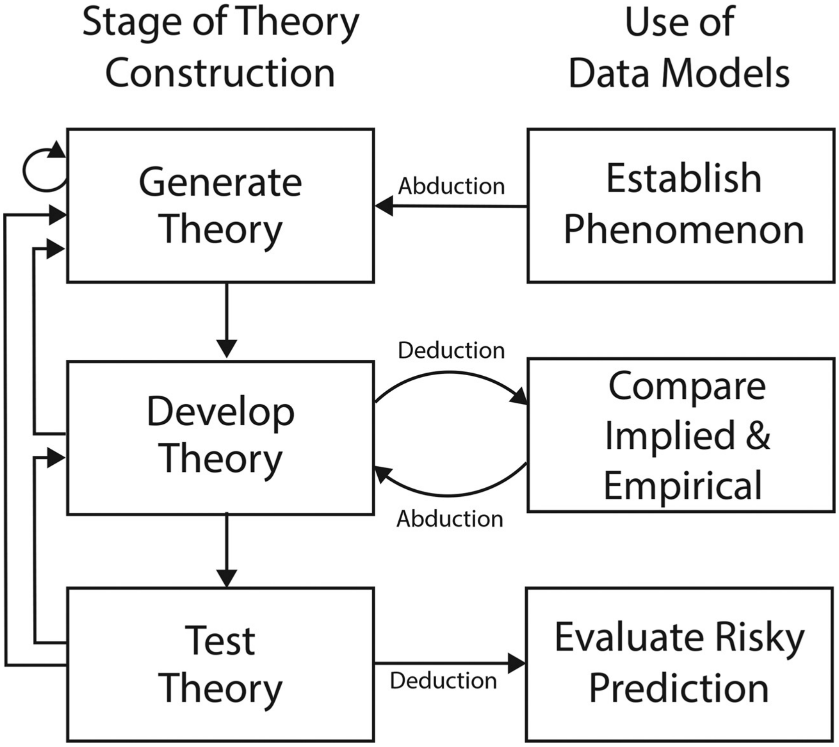
The goal of a formal theory is to explain phenomena. Accordingly, the first step of theory construction is to establish phenomena to be explained. Establishing phenomena is a core aim of science and a full treatment of how best to achieve this aim is beyond the scope of this paper (for a possible way to organize this process see Haig, 2005). However, it is worth highlighting that establishing robust phenomena is a prerequisite to theory construction. The most difficult phenomenon to explain are the ones that do not exist (Lykken, 1991), so researchers must take great care at this stage to ensure that the phenomenon for which they are trying to account are robust. We suspect that the most appropriate phenomena for initial theory development will often include things that researchers would not even think to subject to empirical analysis, taking them for granted as features of the real world. For example, in the case of panic disorder, the core phenomena to be explained are simply the observations that (a) some people experience panic attacks and (b) recurrent attacks often co-occur with persistent worry or concern about those attacks and avoidance of situations where such attacks may occur. These phenomena are so robust that they are typically not the focus of empirical research, yet they are the most important phenomena to be accounted for by a theory of panic disorder.

Generate an Initial Verbal Theory

Once the phenomena to be explained have been established, how do we go about generating an initial theory to explain them? A brief survey of well-known scientific theories reveals that this initial step into theory is often unstructured and highly creative. For example, in the 19th century August Kekulè dreamt of a snake seizing its own tail, leading him to the generate the theory of the benzene ring, a major breakthrough in chemistry (Read, 1995). In the early 20th century, Alfred Wegener noticed that the coastlines of continents fit together similar to puzzle pieces, and consequently developed the theory of continental drift (Wegener, 1966), which formed the basis for the modern theory of plate tectonics (Mauger et al., 1996). In the late 20th century, Howard Gardner explained that he developed his theory of multiple intelligences in the 1980s using "subjective factor analysis" (Walters & Gardner, 1986, p. 176). Although more codified approaches to theory generation exist (e.g., grounded theory; Strauss & Corbin, 1994), we are unaware of any evidence to suggest that any one approach to generating the seed of an initial theory is superior to any other.

Nonetheless, our review of theory in the earlier part of the paper (see "Theories, Phenomena, and Target Systems"), provides guidance for theory generation in at least two ways. The first is rooted in what is perhaps theory's most characteristic feature: its ability to explain phenomena. Initial efforts to generate a theory should

Figure 8
The Abductive Formal Theory Construction Framework (AFTC)



Note. Flowchart depicting the process of developing a formal theory with the abductive formal theory construction framework (AFTC). In the theory generation stage we first establish the phenomenon and then generate an initial verbal theory which is subsequently formalized . In the second stage the theory is developed by testing whether it is consistent with existing empirical findings that are not part of the core phenomenon. If the formal theory is not consistent with some findings, it is adapted accordingly. If these adaptations lead to a "degenerative" theory (Meehl, 1990), we return to the first stage; otherwise we continue to the final stage, in which we test the formal theory using risky predictions (Section 4.3). If many tests are successful, we tentatively accept the theory. If not, the theory must either be adapted (stage two) or a new theory must be generated (stage one).

begin with abductive inference, asking the simple question: What is the best explanation for the phenomena of interest (and, in turn, the data models used to establish those phenomena). The second is rooted in the observation that theories aim to explain phenomena by representing a target system. Accordingly, to generate an initial theory it will likely be helpful to specify the components thought to compose the target system. This process entails dividing the domain of interest into its constituent components (i.e., "partitioning") and selecting those components one thinks must be included in the theory (i.e., "abstraction"; cf. Elliott-Graves, 2014). For researchers adopting a "network perspective" (Borsboom, 2017), the target system is typically presumed to comprise cognitive, emotional, behavioral, or physiological components, especially those identified in diagnostic criteria for mental disorders. Having identified the relevant components we next specify the posited relations among them. Within the domain of the network approach, this second step will typically entail specifying causal relations among symptoms or momentary experiences (e.g., thoughts, emotions, and behavior). Having specified the theory components and the posited relationships among them, the researcher has generated an initial theory posited to account for the phenomena of interest.

Notably, in mental health research, we do not even necessarily need to rely on creative insight about the components and relations among them in order to generate an initial theory. There are already a plethora of verbal theories about mental disorders. If the initial verbal theory is well supported and specific, it will lend itself well to formalization and subsequent theory development and even poor verbal theories can be a useful starting point to developing a successful formal theory (Wimsatt, 1987; Smaldino, 2017).

Formalize the Initial Theory

Once a verbal theory has been specified, the next step is to formalize it. To do so, we first need to choose a formal framework. Dynamical systems are often modeled using differential equations, which describe how variables change over time (e.g., Strogatz, 2014). The panic model we have used as an example throughout this article, uses this formal framework. Another common framework is agent-based modeling (ABM), in which an autonomous agent interacts with its environment, which often includes additional agents (e.g., Grimm & Railsback, 2005). Both frameworks can be implemented in essentially any computer programming language and both are likely to be relevant to psychiatric and psychological research as a whole. The choice of a formalization framework will largely depend on the context (the types of components and types of relations we wish to describe) as well as the level of abstraction or granularity desired by the researcher. For instance, one reason differential equations are attractive is because they may be used to specify component relations on an infinitesimal time scale. In principle, by aggregating, these models can be used to describe behavior at any longer time scale. However, modeling phenomena directly at some longer time scale of principle interest (e.g., modeling symptom dynamics on a month-to-month rather than moment-to-moment level) may be both simpler to achieve and serve equally well in attaining a theoretical explanation of phenomena.

Having chosen a formal framework, the next step is to specify the relations between each component in the language of that framework. This process of formalization is an exercise in being specific. Mathematics and computational programming languages require theorists to specify the exact nature of the relationship between variables. Requiring this level of specificity is one advantage of computational modeling, as it has the effect of immediately clarifying what remains unknown about the target system of interest, thereby guiding future research. However, this also means that theorists will often be in the position of needing to explicate relationships when the precise nature of those relationships is uncertain. We think that, even in the face of this uncertainty, it is better to specify an exact relationship and be wrong than to leave the relationship ambiguously defined (as it often is in verbal theories). Nonetheless, the more theorists can draw on empirical research and other resources when specifying their theory, the firmer the foundation for subsequent theory development. There are several potential sources of information that can guide the formalization process.

First, empirical research can inform specification of components and the relations among them. For example, one could use the finding that sleep quality predicts next-day affect, but daytime affect does not predict next-night sleep (de Wild-Hartmann et al., 2013) to constrain the set of plausible relationships between those two components in the formal theory. There could also be empirical data on the rate of change of components, for example, Siegle et al. (2002) and Siegle et al. (2003) have shown that depressed individuals exhibit longer sustained physiological reactions to negative stimuli than healthy individuals, a finding which is echoed in self-report measures of negative affect (Houben et al., 2015). These findings suggest that the rate of decay of negative affect may be smaller in those with depression relative to those without.

Second, we can derive reasonable scales for components and relationships between components from basic psychological science. For example, classical results from psychophysics show that increasing the intensity of stimuli in almost all cases leads to a nonlinear response in perception (e.g., Fechner et al., 1966): When increasing the volume of music to a very high level, individuals cannot hear an additional increase. Similarly, formal theories of mental disorders may involve some forms of learning. To constrain the relations between components that constitute learning, one can leverage a wealth of research on basic learning, for example on classic or operant conditioning (Henton & Iversen, 2012).

Third, in many cases we can use definitions, basic logic, or common sense to choose formalizations. For example, by definition emotions should change at a time scale of minutes (Houben et al., 2015), while mood should only change at a time scale of hours or days (Larsen, 2000). And we can choose scales of some components using common sense, for example one cannot sleep less than zero or more than 24 hours per day.

Fourth, we could use an existing formal model of another target system, which we expect to have a similar structure as the target system giving rise to the phenomenon of interest. This approach is called “analogical modeling”. For example, Cramer et al. (2016) formulated a model for interactions between symptoms of major depression using the Ising model, which was originally formulated to model magnetism on an atomic level (Ising, 1925). Similarly, Fukano and Gunji (2012) formulated a model for interactions among core components of panic attacks using a Lotka-Volterra model originally formulated to represent predator-prey relationships (Brauer et al., 2012). However, in using this analogical approach, it will be critical to use models capable of representing the kinds of properties expected in the target system of interest.

Fifth, there are methods by which we can potentially estimate the parameters of a formal theory from empirical data.[7] These approaches require considerable development of the formal theory (e.g., the form of a differential equation), suitable data (typically intensive longitudinal data), and a clear measurement model relating observed variables to theory components. Accordingly, this approach already requires considerable progress in generating a formal theory and may be limited by practical considerations. Nonetheless, if successfully carried out, the direct estimation of parameter values would substantially strengthen the theory.

Sixth, one can build on models from cognitive and mathematical psychology. Mental disorders can often be viewed as dysfunctional states of otherwise functional systems of basic psychological processes, such as perception, learning, and memory; thereby allowing mental health researchers to draw on extant models of these processes. For example, Eldar et al. (2016) conceptualize mood as the difference between expectations and outcomes and show in a reinforcement learning model that it is adaptive in environments in which rewards are correlated. Using this model of functional (or adaptive) mood, they identify several meaningful ways one can change the model to produce prolonged depressed mood. Another example is the computational model for oppressive-compulsive disorder (OCD) by Fradkin et al. (2020), who explain various OCD phenomena by dysfunctional information processing within the Bayesian brain framework (Knill & Pouget, 2004).

Evaluating the Initial Theory

The aim of the theory generation stage is to produce a formal theory that is able to explain a phenomenon or set of core

[7] For example, if the theory is formalized as a system of differential equations, the parameters of such equations can in principle be estimated from time-series data using, amongst others, Kalman filter techniques and state-space approaches (e.g., Durbin & Koopman, 2012; Einicke, 2019; Kulikov & Kulikova, 2014). For implementations of these estimation methods see Ou et al. (2019), Carpenter et al. (2017), and King et al. (2015).

phenomena. As we have emphasized throughout this paper, formal theories precisely determine the behavior implied by the theory. Accordingly, explanation in this context means that the theory has demonstrated its ability to produce the behavior of interest (for a more detailed discussion of how formal theories support explanation, see Robinaugh et al., 2021). For example, a theory of panic attacks must be able to produce sudden surges of arousal and perceived threat; a theory of depression must be able to produce sustained periods of low mood; and a theory of borderline personality disorder must be able to produce affective instability. We would note that there are very few theories in psychiatry that have reached this stage of not merely positing, but demonstrating, that the theory can explain the phenomena of interest. Accordingly, completing this stage of theory construction would constitute a significant advance in theories of psychopathology. Once a theory has reached this stage, it is ready for the next stage of theory construction.

Stage 2: Developing Theory

The formal theory produced in the first stage of theory construction will have demonstrated its ability to explain the core phenomena of interest. However, the fact that the formal theory provides *some* explanation does not mean it is the *best* explanation. In other words, demonstrating an ability to explain the phenomenon of interest is a critical first step, but does not guarantee that the formal theory is the best representation of the target system. To move toward such a theory, we must increase both the *explanatory precision* and the *explanatory breadth* of the theory. We will use the term explanatory precision to refer to the specificity of the phenomena for which the theory can account. Whereas researchers in the theory generation stage will typically focus on explaining broad qualitative phenomena, in the development stage researchers will benefit from going further to evaluate whether the formal theory can account for more precise quantitative features of those phenomena. For example, beyond reproducing the broad qualitative features of a panic attack, a well-developed theory would also reproduce the typical duration of and peak arousal associated with a panic attack.

Explanatory breadth refers to the number of phenomena the theory can explain (Thagard, 1978). Whereas in the theory generation stage, researchers will typically be focused on a narrow set of phenomena of interest, to further advance the theory it is necessary to evaluate whether the theory can account for a broader range of phenomena. These phenomena may exist on different levels of aggregation and abstraction, and may necessitate the use of altogether different types of data sources. For instance, a well-developed theory of panic disorder should be able to both approximate the time evolution of panic attacks within a person as well as individual differences in vulnerability to panic disorder and the varying symptom profiles of panic disorder we observe at a population level. The more phenomena for which the theory can account, the greater confidence we can have that the theory is indeed an adequate representation of the target system.

We propose that efforts to increase explanatory precision and breadth should be carried out by repeating two steps. First, we deduce a data model from the formal theory and compare this theory-implied data model with a corresponding empirical data model. If the implied data model is in agreement with the empirical data model, we take this result to expand the theory's explanatory breadth, explanatory precision, or both. However, if a discrepancy is observed, we move to a second step. In this step, we use abductive inference and consider the best explanation for the observed discrepancy, thereby providing insight into how the theory can be adapted to better align with empirical data. By repeating this process with different data models and data sets, the theory is refined such that it becomes a better representation of the target system. We elaborate on these steps below.

Compare Theory-Implied and Empirical Data Models

To generate a theory-implied data model, we first deduce what the theory implies about the behavior of each of the theory's components. The nature of this deduction will depend on the formalism chosen for the theory. In the case of the differential equation modeling, we can either derive or simulate precisely how each theory component will evolve over time (e.g., see Figure 2, right panel). We can then use the simulated data along with a set of formalized auxiliary hypotheses regarding measurement to create a theory-implied dataset. For example, in an earlier part of the current paper we used the panic model to simulate how the components of the system will evolve over time and used these simulations to emulate the measurement that occurs in a cross-sectional epidemiological survey. We could use this same process to instead emulate the measurements likely to be obtained in an experience sampling study by sampling the time series every 90 min with precisely defined measurement functions that capture how those components relate to the self-report assessment of interest. This process of moving from the simulated values of each theory component to the theory-implied dataset requires researchers to specify precisely how we move from a theoretical entity to observed data. Accordingly, this process has the significant advantage of making our measurement assumptions explicit, transparent, and available for careful scrutiny (for more details see Robinaugh et al., 2021).

Once a theory-implied dataset is produced, we can fit our data model of choice to the dataset to produce theory-implied data models that can then be compared to empirical data models generated using the same statistical analysis. This approach of producing theory-implied data models is similar to predictive checks in Bayesian analysis, where data are generated from the fitted model, and checks are performed on summary statistics of those data (Gelman & Hill, 2006; Gelman et al., 2013). However, one key difference is that in Bayesian analysis parameters can typically be estimated directly from data, while the approach presented here is more general and can be applied in contexts where such direct estimation is not possible.

If the theory-implied data model and the empirical data model are in agreement, then theory can be said to provide an explanation for the phenomenon. If that phenomenon is a more specific characterization of a previously explained phenomenon, we have improved the explanatory precision of the theory. If that phenomenon is entirely new, we have improved the explanatory breadth of the theory. The more disparate the phenomenon from what has previously been examined, the greater the boost in our confidence that the theory is indeed representing the real world target system. If the theory-implied data model and the empirical data model are *not* in agreement, we move to the second step of our iterative process and consider whether to adapt the theory.

Abductive Inference and Theory Adaptation

If a discrepancy between a theory-implied and empirical data model is observed, the next step is to determine the best explanation for this discrepancy. The first possibility to consider is whether the discrepancy could have arisen due to the auxiliary hypotheses embodied in the generation of a theory-implied dataset. For example, in the comparison depicted in Figure 7, we observed a discrepancy in the prevalence of persistent concern between our theoretical model (5.30%) and our empirical model (1.32%). This discrepancy may have arisen from our assumptions about how subjects self-reported their level of worry or persistent concern. Because our measurement assumptions have been formalized, they are transparent and available for us to scrutinize. In this case, we may conclude that our measurement assumptions regarding self-reported PC require revision, thereby better equipping us to make use of empirical data on this symptom in future research.

If we do not see a plausible explanation for the discrepancy in our auxiliary hypotheses, we must turn our focus to our theory and use abductive reasoning to consider the best explanation for the observed discrepancies. This explanation may lie in the value of parameters, the functional form of certain relationships or even the broader structure of the theory (the components and the presence or absence of relationships between them). We can then use this abductive inference to make one or several of changes to the theory, producing a new *adapted formal theory*, which better accounts for the empirical data at hand. For example, consider again the discrepancy between the theory-implied and empirically observed prevalence of persistent concern or worry depicted in Figure 7. In our simulated data from our panic disorder theory, all "subjects" who exhibited Av also exhibited PC. However, in the empirical data, most individuals who exhibited Av *did not* exhibit PC. One explanation for this discrepancy is that we underestimated the success of avoidance as a strategy for reducing anxiety, thereby failing to account for individuals whose avoidance regulates arousal sufficiently well that they do not exhibit persistent concern or worry. As with our auxiliary hypotheses, because formalization has made all aspects of the theory precise and explicit, we can better evaluate what aspects of the theory may have given rise to the observed discrepancy, thereby providing clear guidance for how it may be improved. In this case, we may conclude that it is necessary to revise the parameter that defines the strength of the effect of avoidance on arousal or incorporate an effect directly on perceived threat. With a modestly revised theory, we can now continue the theory development process, comparing the revised theory-implied data models with additional models derived from empirical data. By repeating this process across many data models, the theory can be refined and begin to account for a broader and more precise set of phenomena.

Considerations for Theory Development

To this point, we have provided a broad sketch of how comparisons between theory-implied data models and empirical data models can be used, not for theory testing, but for theory development. The core of this approach is simple. Discrepancies between robust empirical data models and the data models implied by our theory give us both an opportunity to evaluate what the theory can explain and an opportunity to learn how the theory can be improved. However, while the core notion of this approach is simple, the specifics are difficult. We suspect there is no single recipe for the best approach to carrying out this stage of theory construction. In the remainder of this section we discuss a number of outstanding questions that researchers are likely to confront when carrying out this work. We view these questions as important avenues for future research that expand on the broad ideas presented here.

What Data and Data Models Are Most Appropriate for Theory Development? Psychological theories will be most successful if a multitude of different data types and data models are employed in developing them. For dynamical systems theories of mental disorders, intensive longitudinal data in large samples may play an especially important role as such data can provide relatively direct measurements of the components of the target system with reasonably high sampling frequency. Experimental data in which the target system is perturbed or manipulated may similarly be especially valuable, as it can help better inform conclusions about causal relations between different components of the target system. However, it is important to note that other data types will also be highly valuable. For example, we were able to use cross-sectional epidemiological data to provide insight into the shortcomings of our theory of panic disorder, despite the absence of longitudinal data or experimental control. In fact, the panic model implies a wide variety of observational consequences: that panic attacks should correlate with avoidance, but also that panic attacks will unfold on a shorter time scale than does learning to avoid panic-related situations, that perceived threat and arousal should move in a synchronized fashion as panic attacks build up, and that specific interventions on the system should produce specific theory-implied behavior. This allows for a much more detailed picture of which kinds of observations to expect if the theory is correct, and thus opens up a range of possibilities for informing theory development using a wide range of data models. This is a property we expect to be present in any theory which is formalized with a sufficient level of specificity. In other words, if we have a formal theory, then we do not need to choose between cross-sectional and intensive longitudinal research or between observational and experimental data. Rather, we can leverage the information provided by each of these data models to inform theory development using the approach we have proposed here.

Regardless of the type of data model, it is critical to the approach we have proposed that the data models used to inform theory be robust. Mental disorders are extremely heterogeneous and multifactorial, which means that large amounts of carefully sampled data are required both to establish the phenomena that should be explained by a formal theory, and to inform the development of the theory. If the phenomenon itself does not generalize in the intended population, it is futile to engage in theory construction, because it is unclear whether there is any phenomenon to be explained in the first place. Similarly, if the data models used for theory development are not reliable and generalizable, theory development becomes an exercise of fitting noise, and different research teams will disagree about theory development because they base their decisions on noisy data sets that point in different directions. Formal theory construction in psychiatry and clinical psychology will therefore be substantially strengthened by initiating large-scale data collection efforts aimed at establishing and characterizing mental health phenomena.

How Do I Know If I Have the Right Measurement Function? Discrepancies between empirical and theory-implied data models can arise from both the theory and from the measurement function that connects the formal theory with the data. This poses a significant barrier to our ability to use empirical data to inform theory development. This barrier is not unique to the approach we have proposed here. Discrepancies between theoretical predictions and the findings from empirical research can only ever be attributed to the conjunction of theory and auxiliary hypotheses (e.g., measurement functions), never to the theory alone (Meehl, 1978; Robinaugh et al., 2021). This problem is pervasive in psychology, where we are typically unsure about the precise relationship between the observed variables in our data and the components of our target system (Kellen et al., 2020). Indeed, the causal connection between attributes of the real world and observed data is notoriously hard to establish (Borsboom et al., 2004), especially at the level of higher order latent constructs (e.g., "internalizing", the "p-factor"; Caspi et al., 2014; Krueger, 1999). As a result, psychology lacks the elegant coordination between mathematical theory and observational data that often characterizes the physical sciences (Wigner, 1963).

Although our approach does not resolve this fundamental issue, it helps to address it by forcing researchers to be transparent about their measurement assumptions. Comparisons between theory-implied and empirical data models require that measurement functions be formalized, and thus, made explicit and available for critical evaluation. This allows researchers to evaluate whether the posited measurement function may have given rise to any observed discrepancies between the theory-implied and empirical data models. In principle, this means that data model comparisons can improve both our theories and our measurement models: If we have a formal theory that is already well-established based on different data types, it is less likely to be the source of data model discrepancies than the measurement function itself and the and so it the proposed measurement function may require revision. Conversely, if a given measurement function has been well-established, any observed discrepancies may be more readily attributed to the theory. This suggests that formal theories and measurement models must be developed in parallel, with the development of one strengthening the development of the other.

When is Theory Adaptation Warranted? We suspect it is unlikely that empirical and implied data models will ever be exactly the same in most areas of psychology. Indeed, we should not necessarily aspire for them to be equivalent, as formal theories are representations only of the target system and are not intended to represent all aspects of the real world that could bear on the data. We therefore need a way to decide whether a disagreement between the empirical and theory-implied data model is due to stochastic noise (e.g., sampling variance) or whether this indicates a flaw in how our theory represents the target system. Only the latter would warrant adapting the formal theory or associated auxiliary hypotheses. One possible means of informing this decision is to compute the likelihood of both the empirical and implied data model given the empirical data, and use statistical procedures to compare both likelihoods. This could entail, for example, a likelihood ratio test or the computation of a Bayes-factor. If we conclude that the likelihoods are not sufficiently different, revision of the theory may not be warranted and we can tentatively conclude that the theory can explain the empirical data model. On the other hand, if we conclude that the likelihoods are sufficiently different, it calls for an adaptation to be made to either our theory or our auxiliary hypotheses.

When Has an Adaptation Improved the Theory? Once we have adapted our formal theory, we would like to test whether those adaptations lead to robust improvements in the correspondence between the theory-implied and empirical data models. We can again make use of a likelihood ratio test or Bayes factor, except that here our comparison is not between empirical and implied data models, but rather between the implied data models of the original formal theory and the implied data models of the adapted formal theory. That is, we can compare the likelihood of the empirical data given those two implied data models.

Notably, our discussion of theory adaptations to this point has focused only on "local optimization," providing an adapted formal theory that fits the present data better than the original formal theory. However, that same adaptation could *worsen* the fit with other types of data. This possibility presents a major challenge to theory development, because it means that we can get stuck in a process of adapting the formal theory to increase the fit to one type of data, while decreasing the fit to a number of other data sets. To avoid this problem and properly evaluate whether an adaptation improves the theory more globally, we need strategies that take the fit to all available data into account when choosing whether to adapt a formal theory. This is a difficult task. However, it is likely that relevant methodology can be adapted from fields with a longer tradition of formal theory development which frequently deal with this problem.

When Should a Theory Be Developed and When Should It Be Abandoned? The theory development stage can have two possible outcomes. One is that the theory becomes increasingly difficult to develop, because each adaptation introduced to better fit a certain data set worsens the fit to a set of other data sets. For such "degenerative" theories (Meehl, 1990; Lakatos, 1976), it is most appropriate to return to the theory generation stage and choose a different starting point. Alternatively, iteratively completing the steps of deducing data models and adapting the formal theory may lead to a theory that not only explains the core phenomena, but also explains many related phenomena with high precision. In this case, the formal theory is ready for the final step of theory testing.

Stage 3: Testing Theory

In the framework we have proposed here, the heavy-lifting of theory construction is done in the generation and development stages. Theory generation is not merely the act of writing down a plausible explanation, it is precisely specifying a mathematical or computational model and *showing* that the model can explain the phenomena of interest. Theory development is not merely expounding on one's initial ideas, it is demonstrating an expansion of its explanatory breadth and precision. The great majority of theories in psychology remain in the stage of theory generation and exceedingly few have demonstrated the kind of explanatory breadth and precision that would suggest they no longer require further development. Accordingly, we anticipate nearly all theory construction efforts in psychology and psychiatry for the foreseeable future will be focused on generating and developing theories with considerable explanatory capacity. Nonetheless, we do think there is a final stage of theory construction worth noting: that of the traditional hypothesis test.

Importantly, this stage of theory construction does not call for null hypothesis significance tests, but rather strong tests of a theory; tests of risky predictions that render the theory vulnerable to refutation (Meehl, 1990). Risky predictions are those that would be unlikely were it not for the theory. To take an example from another domain of science, perhaps one of the most remarkable instances of a risky prediction occurred more than 150 years ago, when two astronomers observed that the orbit of Uranus deviated from the orbit anticipated by Newtonian physics and determined that these deviations could be explained by the presence of an as yet unseen planet (Bamford, 1996). Using Newtonian physics, they predicted the presence and precise location of this previously undiscovered celestial body and, startlingly, their prediction was correct. The planet Neptune was discovered. This prediction was spectacularly unlikely in the absence of Newton's theory, so, when the prediction bore out, it afforded the theory enormous credibility.

The example of Neptune's discovery is instructive not only because it illustrates the kinds of tests that afford corroboration, but also because it clarifies the kinds of theories that can support such risky tests. The existence of Neptune was predicted when the theory *failed* to account for the orbit of Uranus. Yet, despite this discrepancy between the theory and empirical data, astronomers did not abandon the theory. Rather they doubled down, asserting that if there was a discrepancy between the theory and our knowledge of the universe, the problem must lie with our knowledge of the universe. More precisely, they determined the problem was with their auxiliary hypothesis that there were no unaccounted for celestial bodies affecting the orbit of Uranus and they revised these hypotheses rather than the core of the theory. This allegiance to the theory was only possible because the theory had been so well-developed and shown to explain so much that it had already accrued substantial credibility long before this dramatic risky prediction. In the face of a discrepancy between the theory's prediction and the empirical data, astronomers inferred the best explanation for the discrepancy and used that insight in conjunction with the theory itself to make a highly specific risky prediction. Accordingly, this example suggests that to subject a theory to a risky prediction, the theory should have already gained considerable credibility and must be capable of making precise predictions.

Formal theories that have gone through the stages of theory generation and development we have outlined here will meet these criteria. In the theory development stage, the theory will have gained credibility by demonstrating its explanatory breadth and it will be capable of supporting strong hypothesis testing by allowing the researcher to derive predictions sufficiently precise that they would be unlikely in the absence of the theory. As in the theory development stage, the theory testing stage thus calls for us to deduce the precise data models implied by our theory and to compare those implied data models with empirical data models. For example, a very well-developed theory of panic disorder would be able to predict the peak value of perceived threat likely to result from a particular arousal-inducing manipulation (e.g., by breathing CO2 enriched air; Roberson-Nay et al., 2017). More importantly, it may predict that previously unconsidered treatments should be effective in the treatment of panic disorder or that a characteristic pattern of target system behavior signifies vulnerability to panic attacks, thereby identifying a marker of vulnerability that can support preventive interventions before the disorder arises (Robinaugh et al., 2019). Regardless of the prediction, as this stage of theory construction, when theory-implied data models are compared with empirical data models, the aim is not to develop the theory, but to corroborate or refute the theory.

The function of this stage of theory construction is thus less to build the theory than it is to show whether the theory is ready to stand on its own. It is an attempt to demonstrate to other researchers and applied practitioners that the theory is sufficiently well developed that it can make clear and accurate predictions about what we will see in the real world. Accordingly, research at this stage must be confirmatory in the strictest sense of the term (Wagenmakers et al., 2012). These studies should be preregistered with model simulations showing the formal theory, formalized measurement, and specific analyses that will be used in the study and, thus, the precise data model that the researcher expects to observe. A theory that passes such a strong test by predicting the observed empirical data model will be strongly corroborated, having demonstrated that practitioners can trust the theory to help them predict and control the world around them. For example, using the theory to identify those in need of care, make diagnostic decisions, and determine the most appropriate treatment. Theory testing is thus the final stage of theory construction that must be passed before a theory is ready to be taken out of the hands of researchers and used effectively in the real world, firmly and reliably supporting not only explanation, but also the prediction and control of psychological phenomena.

Discussion

In this article, we examined how data models can best inform the development of formal theories of psychopathology. We focused especially on the network approach to psychopathology and considered three possible routes by which the conditional dependence networks used in this literature may inform formal theories about how mental disorders operate as complex systems. We found that these data models were not themselves capable of representing the structure we presume will be needed for a formal theory of any mental disorder. Perhaps more surprisingly, we also found that we were unable to draw clear and reliable inferences from data models about the underlying system. Importantly, this analysis is not intended as a critique of the specific data models we examined here, nor is it a dismissal of their value. Quite the opposite. These data models provide rich and valuable information about the relationships among components of a system and we strongly suspect that our concerns about their ability to inform theory would hold for any data model that can feasibly be estimated from empirical data. Nonetheless, our analysis does strongly suggest that the network approach to psychopathology cannot succeed using these data models alone.

We found that the most promising use of empirical data models for theory development was to compare them with theory-implied data models. Building on this observation, we proposed the abductive formal theory construction framework (AFTC), a roadmap for theory construction that specifies how data models can best be used in the generation, development, and testing of formal theories. In this framework, formal theories play an active role in their own development, with an initial formalized theory refined over time through ongoing comparison between theory-implied and empirical data models. The foundation of this approach is the generation of an initial formal theory that initiates a cycle of theory development. Within this cycle of theory development, theory-implied data models and empirical data models are compared and abductive inference regarding any discrepancy between these models is used to guide ongoing revision to the theory. Only after an extended period of development is the theory well-equipped to undergo rigorous theory testing.

As theory construction in psychology progresses, we believe it will be important to integrate theories across levels of analysis. Formalizing theories as mathematical or computational models makes them explicit, transparent, and expressed in a language that is consistent across domains of science, thereby rendering them widely accessible to researchers across disciplines. This is especially noteworthy in the context of theories of psychopathology, because there are other disciplines that bear on or are directly concerned with mental health that already have strong traditions of mathematical and computational modeling. For example, there is a rich tradition of formal theory in the domains of mathematical and cognitive psychology that focuses on understanding the healthy functioning of thoughts, emotions, and behaviors. Integrating such theories with the kinds of formal theories we have emphasized in this article could help us to produce theories that can represent systems both in healthy and unhealthy states of functioning, and explain the transition between those states (e.g., in the development or treatment of psychopathology; Haslbeck, 2020). Similarly, the highly formalized field of computational psychiatry studies functional pathways in the brain, and how dysfunctional versions of them lead to symptoms of mental disorders (e.g., Friston et al., 2014; Huys et al., 2016; Stephan & Mathys, 2014; Wang & Krystal, 2014). Results from computational psychiatry can inspire mechanisms for models on the higher level of analysis we have emphasized here, and constrain such models by their neuroscientific plausibility.

Theories at these distinct levels of analysis could ultimately be integrated to produce more comprehensive theories that can account for phenomena across multiple levels of analysis. The promise of this type of multiscale modeling approach has recently been discussed in the field of biological psychiatry (Joshi et al., 2020; Lytton et al., 2017; Readhead et al., 2018). If we are to develop these comprehensive multiscale models of these disorders, it will be necessary to have well-developed formal theories at each level of analysis, including the behavioral, cognitive, and affective level that is the focus of this article. This level of analysis is highly relevant to clinical practice, both because it is the level at which the core objects of psychiatry (i.e., signs and symptoms) reside (Parnas et al., 2013) and because many first-line treatments for mental disorders operate at this level (e.g., cognitive behavioral therapy). Ultimately, we hope that formalizing theories across these related disciplines and across different levels of analysis will allow for more unified and collaborative development of theories of psychopathology, with theoretical gains in one area (e.g., a model of neurobiological function) facilitating gains in others (e.g., a model of panic disorder; Cicchetti & Dawson, 2002; Kopniksy et al., 2002).

We have argued that research on mental disorders will be best advanced by constructing formal theories of psychopathology and we have put forward a framework that posits how best to use data models at each stage of theory construction. This framework has the potential to help our field tackle the complexities of mental health, connect psychiatric research to work in basic psychological science, and to better integrate research on psychopathology across different levels of analysis. We do not intend this framework to be a prescription for how all research must proceed, but rather a guide that can provide researchers with ideas for how to advance their theories. We would also stress that any given researcher need not engage in all aspects of the work outlined in this framework. Indeed, we suspect some division of labor in psychology would strengthen the field, allowing some researchers to focus on rigorous detection and description of empirical phenomena while others focus on the generation and evaluation of rigorous formal theories that explain those phenomena. The approach we have proposed here provides a framework in which these diverse efforts, across different domains and different levels of analysis, can work together to advance our understanding of psychopathology. Indeed, we believe that it will only be through this ongoing collaboration and integration across researchers that we will be able to leverage the empirical literature to produce genuine advances in our ability to explain, predict, and control psychopathology.

References

Alegria, M., Jackson, J. S., Kessler, R. C., & Takeuchi, D. (2007). *Collaborative Psychiatric Epidemiology Surveys (CPES), 2001-2003 [United States]*. Inter-university Consortium for Political and Social Research.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM–5* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Bailer-Jones, D. M. (2009). *Scientific models in philosophy of science*. University of Pittsburgh Press.

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. https://doi.org/10.1016/0039-3681(95)00045-3

Ben-Zeev, D., & Young, M. A. (2010). Accuracy of hospitalized depressed patients' and healthy controls' retrospective symptom reports: An experience sampling study. *The Journal of Nervous and Mental Disease*, *198*(4), 280–285. https://doi.org/10.1097/NMD.0b013e3181d6141f

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

Bongers, S., & Mooij, J. M. (2018). *From random differential equations to structural causal models: The stochastic case*. arXiv. https://arxiv.org/abs/1803.08784

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. https://doi.org/10.1002/wps.20375

Borsboom, D., & Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Borsboom, D., van der Maas, H., Dalege, J., Kievit, R., & Haig, B. (2020). *Theory construction methodology: A practical framework for theory formation in psychology*. PsyArXiv. https://doi.org/10.31234/osf.io/w5tp8

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Brauer, F., Castillo-Chavez, C., & Castillo-Chavez, C. (2012). *Mathematical models in population biology and epidemiology* (Vol. 2). Springer.

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PloS ONE*, *8*(4), e60188. https://doi.org/10.1371/journal.pone.0060188

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Cartwright, N. (1983). *How the laws of physics lie*. Oxford University Press.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, L., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p factor: one general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*(2), 119–137. https://doi.org/10.1177/2167702613497473

Cicchetti, D., & Dawson, G. (2002). Multiple levels of analysis. *Development and Psychopathology*, *14*(3), 417–420. https://doi.org/10.1017/s0954579402003012

Clark, D. M. (1986). A cognitive approach to panic. *Behaviour Research and Therapy*, *24*(4), 461–470. https://doi.org/10.1016/0005-7967(86)90011-2

Contreras, A., Nieto, I., Valiente, C., Espinosa, R., & Vazquez, C. (2019). The study of psychopathology from the network analysis perspective: A systematic review. *Psychotherapy and Psychosomatics*, *88*(2), 71–83.

Cramer, A. O. J., van Borkulo, C. D., Giltay, E. J., van der Maas, H. L. J., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major depression as a complex dynamic system. *PloS ONE*, *11*(12), e0167490. https://doi.org/10.1371/journal.pone.0167490

Dablander, F., Ryan, O., & Haslbeck, J. M. B. (2019). Choosing between AR(1) and VAR(1) models in typical psychological applications. *PsyArXiv*. https://doi.org/10.1371/journal.pone.0240730

Dalege, J., Borsboom, D., Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. (2016). Toward a formalized account of attitudes: The causal attitude network (CAN) model. *Psychological Review*, *123*(1), 2–22. https://doi.org/10.1037/a0039802

de Wild-Hartmann, J. A., Wichers, M., van Bemmel, A. L., Derom, C., Thiery, E., Jacobs, N., van Os, J., & Simons, C. J. P. (2013). Day-to-day associations between subjective sleep and affect in regard to future depression in a female population-based sample. *The British Journal of Psychiatry*, *202*(6), 407–412. https://doi.org/10.1192/bjp.bp.112.123794

Didelez, V. (2007). Graphical models for composable finite Markov processes. *Scandinavian Journal of Statistics*, *34*(1), 169–185. https://doi.org/10.1111/j.1467-9469.2006.00528.x

Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press.

Eberhardt, F. (2013). Experimental indistinguishability of causal structures. *Philosophy of Science*, *80*(5), 684–696. https://doi.org/10.1086/673865

Einicke, G. A. (2019). *Smoothing, Filtering and Prediction: Estimating the past, present and future*. Prime Publishing.

Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as representation of momentum. *Trends in Cognitive Sciences*, *20*(1), 15–24. https://doi.org/10.1016/j.tics.2015.07.010

Elliott-Graves, A. (2014). *The role of target systems in scientific practice*. [Unpublished doctoral dissertation]. University of Pennsylvania.

Epskamp, S. (2015). IsingSampler: Sampling methods and distribution functions for the Ising model (R package version 0.2) [Computer software manual]. https://CRAN.R-project.org/package=IsingSampler

Epskamp, S., van Borkulo, C. D., van der Veen, D. C., Servaas, M. N., Isvoranu, A.-M., Riese, H., & Cramer, A. O. J. (2018). Personalized network modeling in psychopathology: The importance of contemporaneous and temporal connections. *Clinical Psychological Science*, *6*(3), 416–427. https://doi.org/10.1177/2167702617744325

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*(4), 453–480. https://doi.org/10.1080/00273171.2018.1454823

Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11*(4), 12.

Estes, W. (1975). Some targets for mathematical psychology. *Journal of Mathematical Psychology*, *12*(3), 263–282. https://doi.org/10.1016/0022-2496(75)90025-5

Fechner, G. T., Howes, D. H., & Boring, E. G. (1966). *Elements of psychophysics* (Vol. 1). Holt, Rinehart and Winston.

Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the idiographic dynamics of mood and anxiety via network analysis. *Journal of Abnormal Psychology*, *126*(8), 1044–1056. https://doi.org/10.1037/abn0000311

Flake, J. K., & Fried, E. I. (2019). Measurement schmeasurement: Questionable measurement practices and how to avoid them. PsyArXiv. https://doi.org/10.31234/osf.io/hs7wm

Forré, P., & Mooij, J. M. (2018). Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. *arXiv*. https://arxiv.org/abs/1807.03024

Fradkin, I., Adams, R., Parr, T., Roiser, J., & Huppert, J. (2020). Searching for an anchor in an unpredictable world: A computational model of obsessive compulsive disorder. *Psychological Review*, *127*(5), 672–699. https://doi.org/10.1037/rev0000188

Freedman, H. I. (1980). *Deterministic mathematical models in population ecology* (Vol. 57). Marcel Dekker Incorporated.

Freedman, R. R., Ianni, P., Ettedgui, E., & Puthezhath, N. (1985). Ambulatory monitoring of panic disorder. *Archives of General Psychiatry*, *42*(3), 244–248. https://doi.org/10.1001/archpsyc.1985.01790260038004

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288.

Fried, E. I., & Cramer, A. O. (2017). Moving forward: challenges and directions for psychopathological network theory and methodology. *Perspectives on Psychological Science*, *12*(6), 999–1020. https://doi.org/10.1177/1745691617705892

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). Springer Series in Statistics.

Friston, K. J., Redish, A. D., & Gordon, J. A. (2017). Computational nosology and precision psychiatry. *Computational Psychiatry*, *1*, 2–23. https://doi.org/10.1162/CPSY_a_00001

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148–158. https://doi.org/10.1016/S2215-0366(14)70275-5

Fukano, T., & Gunji, Y.-P. (2012). Mathematical models of panic disorder. *Nonlinear Dynamics, Psychology, and Life Sciences*, *16*(4), 457–470.

Gardner, C., & Kleinman, A. (2019). Medicine and the mind—the consequences of psychiatry's identity crisis. *New England Journal of Medicine*, *381*(18), 1697–1699. https://doi.org/10.1056/NEJMp1910603

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press.

Glauber, R. J. (1963). Time-dependent statistics of the Ising model. *Journal of Mathematical Physics*, *4*(2), 294–307. https://doi.org/10.1063/1.1703954

Grimm, V., & Railsback, S. F. (2005). *Individual-based modeling and e ecology*. Princeton University Press.

Groen, R. N., Ryan, O., Wigman, J. T. W., Riese, H., Penninx, B. W. J. H., Giltay, E. J., Wichers, M., & Hartman, C. A. (2020). Comorbidity between depression and anxiety: assessing the role of bridge mental states in dynamic psychological networks. *BMC Medicine*, *18*(1), 1–17. https://doi.org/10.1186/s12916-020-01738-z

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *PsyArXiv*. https://doi.org/10.31234/osf.io/8hr7n

Guest, O., & Martin, A. E. (2020). *How computational modeling can force theory building in psychological science*. PsyArXiv. https://doi.org/10.31234/osf.io/rybh9

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, *10*(4), 371–388. https://doi.org/10.1037/1082-989X.10.4.371

Haig, B. D. (2008). Precis of 'an abductive theory of scientific method'. *Journal of Clinical Psychology*, *64*(9), 1019–1022. https://doi.org/10.1002/jclp.20506

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences*. MIT Press.

Hamaker, E. L., Grasman, R. P., & Kamphuis, J. H. (2010). Regime-switching models to study psychological process. In P. C. M. Molenaar & K. M. Newell (Eds.), *Individual pathways of change: Statistical models for analyzing learning and development* (pp. 155–168). American Psychological Association.

Hamilton, J. D. (1995). *Time series analysis*. Princeton University Press.

Haslbeck, J. M. B. (2020). *Modeling psychopathology: From data models to formal theories* [Unpublished doctoral dissertation]. University of Amsterdam.

Haslbeck, J. M. B., & Fried, E. I. (2017). How predictable are symptoms in psychopathological networks? A reanalysis of 18 published datasets. *Psychological Medicine*, *47*(16), 2767–2776. https://doi.org/10.1017/S0033291717001258

Haslbeck, J. M. B., & Ryan, O. (2021). Recovering within-person dynamics from psychological time series. *Multivariate Behavioral Research*. Advance online publication. https://doi.org/10.1080/00273171.2021.1896353

Haslbeck, J. M. B., Epskamp, S., Marsman, M., & Waldorp, L. J. (2020). Interpreting the Ising model: The input matters. *Multivariate Behavioral Research*, *12*, 1–11.

Hayes, A. M., & Andrews, L. A. (2020). A complex systems approach to the study of change in psychotherapy. *BMC Medicine*, *18*(1), 1–13. https://doi.org/10.1186/s12916-020-01662-2

Henton, W. W., & Iversen, I. H. (2012). *Classical conditioning and operant conditioning: A response pattern analysis*. Springer Science & Business Media.

Higgs, P. W. (1964). Broken symmetries and the masses of gauge bosons. *Physical Review Letters*, *13*(16), 508–509. https://doi.org/10.1103/PhysRevLett.13.508

Hjelmeland, H., & Loa Knizek, B. (2018). The emperor's new clothes? A critical look at the interpersonal theory of suicide. *Death Studies*, *43*(3), 168–178. https://doi.org/10.1080/07481187.2018.1527796

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, *141*(4), 901–930. https://doi.org/10.1037/a0038822

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413. https://doi.org/10.1038/nn.4238

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus [Contribution to the theory of ferromagnetism]. *Zeitschrift Für Physik A Hadrons and Nuclei*, *31*(1), 253–258.

Jones, P. J., Mair, P., Riemann, B. C., Mugno, B. L., & McNally, R. J. (2018). A network perspective on comorbid depression in adolescents with obsessive-compulsive disorder. *Journal of Anxiety Disorders*, *53*, 1–8. https://doi.org/10.1016/j.janxdis.2017.09.008

Joshi, A., Wang, D.-H., Watterson, S., McClean, P. L., Behera, C. K., Sharp, T., & Wong-Lin, K. (2020). Opportunities for multiscale computational modelling of serotonergic drug effects in Alzheimer's disease. *Neuropharmacology*, *174*, 108118. https://doi.org/10.1016/j.neuropharm.2020.108118

Kellen, D. (2019). A model hierarchy for psychological science. *Computational Brain & Behavior*, *2*(3-4), 160–165. https://doi.org/10.1007/s42113-019-00037-y

Kellen, D., Davis-Stober, C., Dunn, J. C., & Kalish, M. (2020). *The problem of coordination and the pursuit of structural constraints in psychology*. PsyArXiv. https://psyarxiv.com/3eupv/

Kendler, K. S. (2019). From many to one to many—the search for causes of psychiatric illness. *JAMA Psychiatry*, *76*(10), 1085. https://doi.org/10.1001/jamapsychiatry.2019.1200

King, A. A., Nguyen, D., & Ionides, E. L. (2015). Statistical inference for partially observed Markov processes via the R package pomp. *arXiv*. https://arxiv.org/abs/1509.00503

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Kopniksy, K. L., Cowan, W. M., & Hyman, S. E. (2002). Levels of analysis in psychiatric research. *Development and Psychopathology*, *14*(3), 437–461.

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry*, *56*(10), 921–926. https://doi.org/10.1001/archpsyc.56.10.921

Kulikov, G. Y., & Kulikova, M. V. (2014). Accurate numerical implementation of the continuous-discrete extended Kalman filter. *IEEE Transactions on Automatic Control*, *59*(1), 273–279. https://doi.org/10.1109/TAC.2013.2272136

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

Larsen, R. J. (2000). Toward a science of mood regulation. *Psychological Inquiry*, *11*(3), 129–141. https://doi.org/10.1207/S15327965PLI1103_01

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. SAGE Publications.

Lunansky, G., van Borkulo, C. D., & Borsboom, D. (2020). Personality, resilience, and psychopathology: A model for the interaction between slow and fast network processes in the context of mental health. *European Journal of Personality*, *34*(6), 969–987. https://doi.org/10.1002/per.2263

Lykken, D. T. (1991). What's wrong with psychology anyway. *Thinking Clearly about Psychology*, *1*, 3–39.

Lytton, W. W., Arle, J., Bobashev, G., Ji, S., Klassen, T. L., Marmarelis, V. Z., Schwaber, J., Sherif, M. A., & Sanger, T. D. (2017). Multiscale modeling in the clinic: Diseases of the brain and nervous system. *Brain Informatics*, *4*(4), 219–230. https://doi.org/10.1007/s40708-017-0067-5

Mauger, R., Tarbuck, E. J., & Lutgens, F. K. (1996). *Earth: An introduction to physical geology*. Prentice Hall.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Millner, A. J., Robinaugh, D. J., & Nock, M. K. (2020). Advancing the understanding of suicide: the need for formal theory and rigorous descriptive research. *Trends in Cognitive Sciences*, *24*(9), 704–716. https://doi.org/10.1016/j.tics.2020.06.007

Mooij, J. M., Janzing, D., & Schölkopf, B. (2013). *From ordinary differential equations to structural causal models: The deterministic case*. ArXiv. https://arxiv.org/abs/1304.7920

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nguyen, J., & Frigg, R. (2017). Mathematics is not the only language in the book of nature. *Synthese*, *28*, 1–22.

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

Ou, L., Hunter, M. D., & Chow, S.-M. (2019). dynr: Dynamic modeling in R (R package version 0.1.14-9) [Computer software manual]. https://CRAN.R-project.org/package=dynr

Parnas, J., Sass, L. A., & Zahavi, D. (2013). Rediscovering psychopathology: The epistemology and phenomenology of the psychiatric object. *Schizophrenia Bulletin*, *39*(2), 270–277. https://doi.org/10.1093/schbul/sbs153

Pe, M., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., Mata, J., Jaeggi, S. M., Buschkuehl, M., Jonides, J., Kuppens, P., & Gotlib, I. H. (2015). Emotion-network density in major depressive disorder. *Clinical Psychological Science*, *3*(2), 292–300. https://doi.org/10.1177/2167702614540645

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pero, F. (2015). *Whither structuralism for scientific representation?* [Unpublished doctoral dissertation]. University of Florence.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. MIT Press.

Read, J. (1995). *From alchemy to chemistry*. Courier Corporation.

Readhead, B., Haure-Mirande, J.-V., Funk, C. C., Richards, M. A., Shannon, P., Haroutunian, V., Sano, M., Liang, W. S., Beckmann, N. D., Price, N. D., Reiman, E. M., Schadt, E. E., Ehrlich, M. E., Gandy, S., Dudley, J. T. (2018). Multiscale analysis of independent Alzheimer's cohorts finds disruption of molecular, genetic, and clinical networks by human herpesvirus. *Neuron*, *99*(1), 64–82. https://doi.org/10.1016/j.neuron.2018.05.023

Ritter, F. E., Tehranchi, F., & Oury, J. D. (2019). Act-r: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(3), e1488. https://doi.org/10.1002/wcs.1488

Roberson-Nay, R., Gorlin, E. I., Beadel, J. R., Cash, T., Vrana, S., & Teachman, B. A. (2017). Temporal stability of multiple response systems to 7.5% carbon dioxide challenge. *Biological Psychology*, *124*, 111–118. https://doi.org/10.1016/j.biopsycho.2017.01.014

Robinaugh, D. J., Haslbeck, J. M. B., Waldorp, L., Kossakowski, J. J., Fried, E. I., Millner, A., McNally, R. J., van Nes, E. H., Scheffer, M., Kendler, K. S., & Borsboom, D. (2019). Advancing the network theory of mental disorders: A computational model of panic disorder. *PsyArXiv*. https://doi.org/10.31234/osf.io/km37w

Robinaugh, D. J., Haslbeck, J., Ryan, O., Fried, E. I., & Waldorp, L. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/1745691620974697

Robinaugh, D. J., Hoekstra, R. H. A., & Borsboom, D. (2020). The network approach to psychopathology: A review of the literature 2008-2018 and an agenda for future research. *Psychological Medicine*, *50*(3), 353.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Ryan, O., & Hamaker, E. (2021). Time to intervene: A continuous-time approach to network analysis and centrality. *Psychometrika*. Advance online publication. https://doi.org/10.1007/s11336-021-09767-0

Ryan, O., Bringmann, L. F., & Schuurman, N. K. (2019, October 1). The challenge of generating causal hypotheses using network models. *PsyArXiv*. https://doi.org/10.31234/osf.io/ryg69

Schmittmann, V. D., Cramer, A. O., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, *31*(1), 43–53. https://doi.org/10.1016/j.newideapsych.2011.02.007

Siegle, G. J., Steinhauer, S. R., Carter, C. S., Ramel, W., & Thase, M. E. (2003). Do the seconds turn into hours? Relationships between sustained pupil dilation in response to emotional information and self-reported rumination. *Cognitive Therapy and Research*, *27*(3), 365–382. https://doi.org/10.1023/A:1023974602357

Siegle, G. J., Steinhauer, S. R., Thase, M. E., Stenger, V. A., & Carter, C. S. (2002). Can't shake that feeling: Event-related fMRI assessment of sustained amygdala activity in response to emotional information in depressed individuals. *Biological Psychiatry*, *51*(9), 693–707. https://doi.org/10.1016/S0006-3223(02)01314-8

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

Smaldino, P. E. (2019). Better methods can't make up for mediocre theory. *Nature*, *575*(7781), 9. https://doi.org/10.1038/d41586-019-03350-5

Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, *11*(1), 87–104. https://doi.org/10.1177/1088868306294789

Snippe, E., Viechtbauer, W., Geschwind, N., Klippel, A., De Jonge, P., & Wichers, M. (2017). The impact of treatments for depression on the dynamic network structure of mental states: Two randomized controlled trials. *Scientific Reports*, *7*, 46523. https://doi.org/10.1038/srep46523

Spirtes, P. (2010). Introduction to causal inference. *Journal of Machine Learning Research*, *11*(5), 1643–1662.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT Press.

Spitzer, R. L., Kroenke, K., & Williams, J. B. (1980). *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association.

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, *25*, 85–92. https://doi.org/10.1016/j.conb.2013.12.007

Strauss, A., & Corbin, J. (1994). Grounded theory methodology. *Handbook of Qualitative Research*, *17*, 273–285.

Strogatz, S. H. (2014). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. Westview Press.

Stutz, C., & Williams, B. (1999). Obituary: Ernst Ising. *Physics Today*, *52*(3), 106–108. https://doi.org/10.1063/1.882538

Suárez, M., & Pero, F. (2019). The representational semantic conception. *Philosophy of Science*, *86*(2), 344–365. https://doi.org/10.1086/702029

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

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87*(3), 449–508. https://doi.org/10.1007/BF00499820

Thagard, P. R. (1978). The best explanation: Criteria for theory choice. *The Journal of Philosophy*, *75*(2), 76–92. https://doi.org/10.2307/2025686

Tong, H., & Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(3), 245–268.

Van Orden, K. A., Witte, T. K., Cukrowicz, K. C., Braithwaite, S. R., Selby, E. A., & Joiner, T. E., Jr, (2010). The interpersonal theory of suicide. *Psychological Review*, *117*(2), 575–600. https://doi.org/10.1037/a0018697

van Rooij, I., & Baggio, G., (2020). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *PsyArXiv*. https://doi.org/10.1177/1745691620970604

van Rooijen, G., Isvoranu, A.-M., Meijer, C. J., van Borkulo, C. D., Ruhé, H. G., & de Haan, L. (2017). A symptom network structure of the psychosis spectrum. *Schizophrenia Research*, *189*, 75–83. https://doi.org/10.1016/j.schres.2017.02.018

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A., (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Walters, J. M., & Gardner, H., (1986). The theory of multiple intelligences: Some issues and answers. In R. J. Sternberg, J. E. Davidson, & R. K.

Wagner (Eds.), *Practical intelligence: Nature and origins of competence in the everyday world* (pp. 163–182). CUP Archive.

Wang, X.-J., & Krystal, J. (2014). Computational psychiatry. *Neuron*, *84*(3), 638–654. https://doi.org/10.1016/j.neuron.2014.10.018

Wegener, A. (1966). *The origin of continents and oceans*. Dover Publications.

Wigner, E. P. (1963). The problem of measurement. *American Journal of Physics*, *31*(1), 6–15. https://doi.org/10.1119/1.1969254

Wimsatt, W. C. (1987). *False models as means to truer theories*. Cambridge University Press.

Woodward, J. F. (2011). Data and phenomena: A restatement and defense. *Synthese*, *182*(1), 165–179. https://doi.org/10.1007/s11229-009-9618-5

Appendix A

Simulated Data From the Panic Model

In this appendix we describe in more detail how raw time series data is simulated from the panic model, the full specification of which is given by Robinaugh et al., 2019. Data is simulated using the statistical programming language *R*. We use the panic model to generate time-series data of 1,000 individuals, on a single minute time scale, for 12 weeks, using Euler's method with a step size of .001. This yields a total of n_t = 12,0960 repeated measurements per person. Each individual starts with a different initial value of arousal schema, drawn from a normal distribution with μ = .25 and σ = .0225. The parameters of this distribution were chosen to roughly generate a representative number of panic disorder sufferers (for more details see Robinaugh et al., 2019). Otherwise each individual obtains the same parameter values and the same starting values on all processes, with the stochastic noise terms drawn using a different random seed for each individual. The mapping from this raw data to the variables used in the network models presented in Figure 5 is described in the main text. Code to reproduce this data-generation scheme can be found in the reproducibility archive of this article.[8]

[8] https://osf.io/bnteg/

Appendix B

Panic Model and Statistical Dependencies

In this appendix we describe in more detail the patterns of statistical dependencies produced by the three data models fitted to data simulated from the panic model presented in Figure 5. While in the main text we discuss the statistical dependencies between arousal and perceived threat, and arousal schema and avoidance, here we focus only on the former. Key to the arousal-perceived threat dependencies is the positive feedback loop between arousal and perceived threat in the panic model. If arousal and perceived threat become sufficiently elevated, this "vicious cycle" leads to runaway positive feedback, with a pronounced spike in both arousal and perceived threat (i.e., a panic attack). This spike initiates a process of homeostatic feedback that brings arousal down and suppresses arousal below its baseline for a period of time after this panic attack, a period which we will refer to as a *recovery period*. The panic attack itself lasts about 30 min. However, the recovery period lasts for 2–3 hr (see Figure B1a).

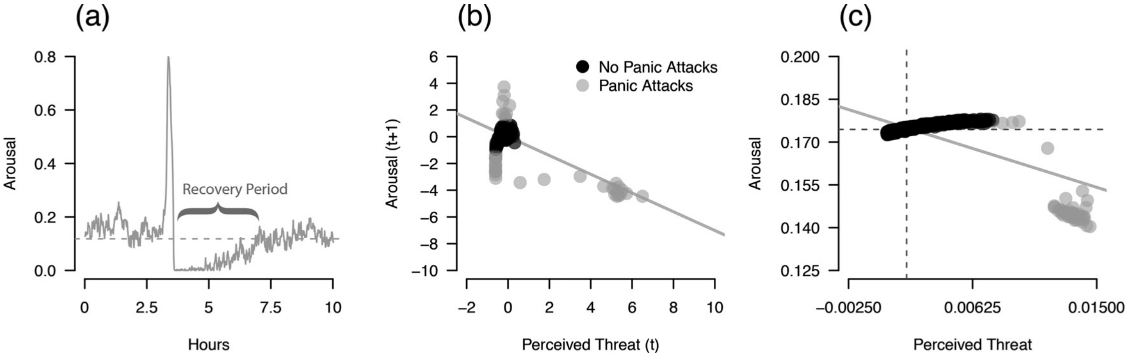
In the VAR model in Figure 5b in the main text, we observed a strong negative conditional relationship between perceived threat at *time t* and arousal at *time t* + 1, conditioning on all other variables at *time t*. The distribution of these lagged variables is shown in Figure B1b, with the gray line representing the also negative marginal relationship. This strong negative cross-lagged relationship is a direct consequence of the recovery period of arousal: High values of perceived threat are closely followed by a long period of low arousal values. This can be seen in Figure B1b, where observations over windows in which a panic attack and recovery period occur are shaded in gray. By averaging arousal values over a window of 90 min, the strong positive causal effects operating locally in time (i.e., over a very short time-interval) are not directly captured, but instead the VAR(1) model describes correctly describes the negative relationship between the *means* of each variable over this window.

In the GGM in Figure 5c in the main text, we saw a positive linear relationship between arousal and perceived threat in the estimated GGM. This dependency indicates that high mean levels of arousal are associated with high mean levels of perceived threat, *conditional* on all other variables. We stress the conditional nature of this relationship, because the *marginal* relationship between the two variables is in fact negative as can be seen in Figure B1c. This negative marginal relationship comes about by combining two groups of individuals that have different mean values on both variables. Individuals who experience panic attacks (gray points) have high average perceived threat, but low average arousal, due to the long recovery period of arousal after a panic attack. On the other hand, individuals who do *not* experience panic attacks have higher average values of arousal, and lower average values of perceived threat. When inspecting the two groups separately, we see that there is a positive linear relationship between mean arousal and perceived threat in the group without panic attacks. The group with panic attacks is too small to determine a relationship. Because escape and avoidance behavior only occur after panic attacks, conditioning on those two variables amounts to conditioning on whether an individual had panic attacks. This conditional relationship is then driven mostly by the positive relationship in the (much larger) group of individuals who have no panic attacks, indicated by the black dots in Figure B1c.

(*Appendices continue*)

Figure B1

Explanation of Negative Relationship Between Arousal and Perceived Threat in Data Models



Note. Panel a depicts arousal during a panic attack, showing the short sharp peak of arousal levels, followed by a longer recovery period of low arousal, before the system returns to the usual resting state. The dotted line indicates the mean level of arousal over the observation window (0 hr–10 hr). Panel b depicts the state-space plot of perceived threat and arousal at the next measurement occasion, as captured by the emulated ESM study and VAR model. Gray points indicate an observation window of 90 min in which either part of a panic attack or the following recovery period is captured. The solid gray line reflects the marginal lagged relationship. Panel c depicts the crosssectional marginal relationship between the mean of arousal and mean of perceived threat, as analyzed in the GGM model. Gray dots indicate individuals who suffer from panic attacks, and black dots represent "healthy" individuals. The solid gray line shows the negative marginal relationship. The dotted gray lines indicate the median of both variables, by which the binarized values used in the Ising model analysis are defined.

Finally, we can explain the weak positive relationship between arousal and perceived threat in the Ising model (Figure 5d in the main text): The levels of these variables are defined by a median split of their mean values, depicted as dotted lines in Figure B1c. Unlike in the GGM, there is a positive marginal relationship between these binarized variables, as the majority of individuals without panic attacks (denoted by the black points) end up in the low perceived threat and low arousal groups (lower left quadrant Figure B1c) or high perceived threat and high arousal groups (upper right quadrant). How then do we end up with a weakly positive conditional relationship between these two binary variables? Similarly to the GGM above, it turns out that conditioning on variables such as escape behavior and avoidance almost entirely separates individuals into either the low arousal and low perceived threat category (e.g., for low escape values) or the high arousal and high perceived threat category (for high escape value). This means that, once we have conditioned on other variables which have direct and indirect causal connections to arousal and perceived threat, there is very little additional information which arousal can add to predicting perceived threat levels (and vice versa). This produces the weak positive conditional relationship between arousal and perceived threat, as well as the stronger positive connections between avoidance and perceived threat.

Appendix C

Details Empirical Versus Simulated Ising Model

In this appendix we describe in more detail how the theory-implied and empirical Ising models presented in Figure 6 are obtained.

C.1. Simulated Data and Implied Ising model

To obtain the theory-implied Ising model we use the raw time-series data generated from the panic model and described in Appendix A.

To create the binary symptom variables, we transformed the raw time-series data of each individual as follows. First, we define anxiety at a given time point as the geometric mean of the arousal and perceived threat components at that point in time. Second, we define a panic attack as short, sharp peak of arousal and perceived threat. We code a panic attack to be present in the time-series data if anxiety takes on a value greater than .5. The duration of a panic attack is the length of time the anxiety variable stays above this threshold, and so we define a single panic attack as a sequence of consecutive time points in which anxiety stays over this threshold. This allows to define our first binary symptom variable, recurrent panic attacks:

1. Recurrent panic attacks (PA): PA is present if the individual experience more than three panic attacks over the observation window.

We define recurrent as more than three over the observation window for consistency with how this symptom is defined in the CPES dataset, detailed below.

(*Appendices continue*)

Next, we can define the symptom persistent concern (PC), again using the time-series of anxiety. This symptom is typically described as experiencing a heightened level of anxiety following a panic attack (American Psychiatric Association, 2013). To define this, for each individual who experiences a panic attack, we calculate the mean level of anxiety in a window of 1,000 min (16.67 hr) following the end of each panic attack. If another panic attack occurs in that window, we instead take the mean level of anxiety between the end of one panic attack and before the beginning of the next. This gives us a vector of mean anxiety levels per person, one for each panic attack experienced. Next, we must define what we consider to be a "heightened" level of anxiety. We do this by obtaining the distribution of mean anxiety levels for healthy individuals, that is, those members of our sample who never experience a panic attack. We consider mean anxiety levels following an attack to be "heightened" if they are greater than the 90th percentile of mean anxiety levels in the healthy population. This gives us our second binary symptom variable.

2. Persistent Concern (PC): PC is present if, following at least one panic attack, higher average levels of Anxiety are present than in the healthy population, as defined by the 90th percentile of average Anxiety in the healthy population.

Finally we take a similar approach to defining the symptom avoidance (Av), typically described as engaging in a heightened level of avoidance behavior following a panic attack. For this symptom, we use the time-series of the avoid component. For each individual who experiences a panic attack, we calculate the mean level of avoid in a window of 1,000 min (16.67 hr) following the end of each panic attack, or before the beginning of the next attack, whichever is shorter. Heightened avoidance behavior is defined relative to the 90th percentile of avoid levels in the healthy population. This gives us our third binary symptom variable.

3. Avoidance (Av): Av is present if, following at least one panic attack, higher average levels of avoid are present than in the healthy population, as defined by the 90th percentile of average avoid in the healthy population.

The Ising model of these three symptom variables is fit using the *EstimateIsing* function from the *IsingSampler* package (Epskamp, 2015), that is, using a nonregularized pseudolikelihood method.

C.2. Empirical Symptom Data

To test the empirical predictions of the panic model, we made use of the publicly available Collaborative Psychiatric Epidemiology Surveys (CPES) 2001–2003 (Alegria et al., 2007). The CPES is a nationally representative survey of mental disorders and correlates in the United States. The CPES is attractive to use for testing the panic model, first because of the large sample size (20,013 participants) ensuring reliable estimates of empirical dependencies, and second, because approximately 146 items in the survey assess either panic attack or panic disorder experiences, characteristics, and diagnoses, typically in terms of lifetime prevalence.

To define our three panic disorder symptoms, we first use the diagnostic manual of the CPES to define whether individuals have ever experienced a panic attack based on responses to 18 items. There are three criteria which must be met for the individual to be classed as having experienced at least one lifetime panic attack. These are shown in Table C1. In coding the presence or absence of a panic attack, individuals must positively report at least four out of the 31 characteristics of a panic attack, according to the second criteria in Table C1. Missing values were taken as a failure to report that characteristic.

With this definition of a panic attack in place, we define the three binary symptoms of panic disorder, following the definitions laid out in the diagnostic manual for panic disorder.

1. PA: PA is present if participant reports more than three lifetime occurrences of an unexpected, short, sharp attack of fear or panic (Item PD4 and criteria in Table C1), more than one of which is out of the blue (PD17a).

Table C1
CPES Diagnostic Criteria

Criterion	Description	Item number(s)
A	A discrete period of intense fear or discomfort	SC20 or SC20a
B (four or more)	Palpitations, pounding heart	PD1a
	Sweating	PD1e
	Trembling or shaking	PD1f
	Sensation of shortness of breath or smothering	PD1b
	Feeling of choking	PD1h
	Chest pain or discomfort	PD1i
	Nausea or abdominal distress	PD1c
	Feeling dizzy, unsteady, lightheaded or faint	PD1d or PD1m
	Derealization or depersonalization	PD1k or PD1l
	Fear of losing control or going crazy	PD1j
	Fear of dying	PD1n
	Paresthesia (numbing or tingling sensations)	PD1p
	Chills or hot flushes	PD1o
C	Symptoms developed abruptly and reached a peak within 10 min	PD3

Note. CPES = Collaborative Psychiatric Epidemiology Surveys. Table shows the three criteria necessary to code an individual as having one liftetime panic attack based on their responses specific to CPES survey items, with corresponding item numbers. Criteria are taken from the CPES diagnostic manual.

(*Appendices continue*)

2. PC: PC is present if reported that following an attack, a month or more of at least one of: (a) persistent concern about having another attack (PD13a); or (b) worry about the implications or consequences of having an attack (PD13b).

3. Av: Av is present if participant reports at least one of: (a) following an attack, changing everyday activities for a month or more (PD13c); (b) following an attack, avoiding situations due to fear of having an attack for a month or more (PD13d); or (c) in the past 12 months, avoiding situations that might cause physical sensation (PD42).

In coding this, if two out of three PA criteria were present, and the third was missing, we assigned a positive value to the PA item. The empirical Ising model was fit using the same procedure as the theory-implied Ising model.

Received April 28, 2020
Revision received February 1, 2021
Accepted February 25, 2021 ■