This is the preprint of a target article in press at Psychological Inquiry. The preprint can be found at https://psyarxiv.com/zg84s, and my reply to 7 commentaries at https://psyarxiv.com/dt6ev.

Lack of theory building and testing impedes progress in the factor and network literature.

Eiko I. Fried

Department of Clinical Psychology, Leiden University

www.eiko-fried.com, eikofried@gmail.com

Abstract

The applied social science literature using factor and network models continues to grow rapidly. Most work reads like an exercise in model fitting, and falls short of theory building and testing in three ways. First, statistical and theoretical models are conflated, leading to invalid inferences such as the existence of psychological constructs based on factor models, or recommendations for clinical interventions based on network models. I demonstrate this inferential gap in a simulation: excellent model fit does little to corroborate a theory, regardless of quality or quantity of data. Second, researchers fail to explicate theories about psychological constructs, but use implicit causal beliefs to guide inferences. These latent theories have led to problematic best practices. Third, explicated theories are often weak theories: imprecise descriptions vulnerable to hidden assumptions and unknowns. Such theories do not offer precise predictions, and it is often unclear whether statistical effects actually corroborate weak theories or not. I demonstrate that these three challenges are common and harmful, and impede theory formation, failure, and reform. Matching theoretical and statistical models is necessary to bring data to bear on theories, and a renewed focus on theoretical psychology and formalizing theories offers a way forward.

1. Data rich and theory poor

"The present methodological and statistical solutions to the replication crisis will only help ensure solid stones; they don't help us build the house." (Muthukrishna & Henrich, 2019)

As this decade draws to an end, cumulative psychological science has seen important improvements, in part due to the open science movement. We have become better at identifying and preventing questionable research practices such as p-hacking and HARKing (hypothesizing after results are known) through tools like preregistration, registered reports, and sharing of data and code (O. Klein et al., 2018; Nosek et al., 2019). Identified challenges and proposed solutions have focused on increasing the reliability and replicability of psychological findings by improving methodological and statistical practices (Muthukrishna & Henrich, 2019).

In the best case, these new best practices lead to more reliable and replicable statistical effects, i.e. robust phenomena (Haig, 2005; Woodward, 2011). But phenomena in psychology, usually the relationship between two variables, or the difference of two groups, are effects that require explaining (explananda)—they are not theories that *do* the explaining (explanantia). Explanantia remain an elusive species in our field. As Cummins put it: "In psychology, we are overwhelmed with things to explain, and somewhat underwhelmed by things to explain them with" (Cummins, 2000). For this reason, many scholars have argued that psychology's attention to statistics and replicability has distracted from a problem that runs much deeper: a crisis of theory (Borsboom, 2013; Borsboom et al., 2020; Cummins, 2000; Guest & Martin, 2020; Haslbeck et al., 2019; Meehl, 1990b; Muthukrishna & Henrich, 2019; Phaf, 2020; Robinaugh, Haslbeck, et al., 2019; Smaldino, 2019; Szollosi et al., 2019; Vaidyanathan et al., 2015; Van Rooij & Baggio, 2020).

What are theories, and how do they relate to data, statistical models, and phenomena[1]? I understand theories as sets of axioms or assumptions that help explain, predict, and control phenomena. Phenomena are robust features of the world, usually obtained by fitting statistical models to data. Data contain the phenomena we are interested in, but also noise (measurement

[1] For the sake of simplicity, I use the term 'theory' broadly; for more detailed accounts differentiating between theoretical frameworks, theories, and hypotheses, see (Muthukrishna & Henrich, 2019).

error, experimenter bias, transcription errors), which is why theories explain phenomena, not data (Woodward, 2011). We can link theories to data by using statistical models that impose assumptions on the data, and choosing an appropriate statistical model that imposes assumptions consistent with the theory is instrumental in bringing data to bear on the theory (Morton, 2009; Spanos & Mayo, 2015; Vaidyanathan et al., 2015). Good psychological theories are useful but imperfect abstractions, and all false in the sense that they are incomplete (Meehl, 1990a). Nonetheless, they differ from each other in the degree to which they help us explain, predict, and control phenomena. Rutherford's model of the atom—electrons orbit a very small and heavy nucleus—was false: simulating data from it shows that the known universe would collapse in a split second. But it got many things right, such as separating electrons from a dense core, and was instrumental in bringing about crucial changes to atomic models in particle physics with a higher *degree* of verisimilitude (i.e. truthlikeness). These newer theories do a better job at explanation, prediction, and control.

1.1 Weak and strong theories

Theories in psychology are often weak theories, impeding theory formation, failure and reform. I define weak theories as narrative and imprecise accounts of hypotheses, vulnerable to hidden assumptions and other unknowns. They do not spell out the functional form in which two variables relate to each other, the conditions under which a hypothesized effect should occur, or the magnitude of a proposed effect. It therefore remains somewhat unclear what the theory actually explains or predicts, or how to use the theory for purposes of control (such as informing treatments in clinical psychology). The same verbal theory (e.g. "x relates to y") can often be formalized in numerous different ways, and these different formalizations can lead to drastically different predictions of what data we expect given the theory, as recently demonstrated by Robinaugh et al. (Robinaugh et al., 2020). Weak theories can usually be defended post-hoc by adding auxiliary assumptions. For example, psychologists have argued that original studies do not replicate due to 'hidden moderators': variables that may be different between original and replication study. Fully embracing this argument means that psychological theories need never be adjusted, because one can always posit hidden moderators (Gershman, 2019; Stroebe & Strack, 2014; Van Bavel et al., 2016). The less precise a theory, the more auxiliary scape goats can be blamed if the theory does not explain or predict well.

Strong theories[2], on the other hand, explicate a precise set of assumptions and axioms about a phenomenon non-ambiguously (Morton, 2009). One common way of doing so is by representing the theory as a formal model, using mathematical notation. Strong theories provide a clear explanation of a phenomenon, rather than just a description of data (Cummins, 2000; Woodward, 2011), and they do so independently of the theorists who designed them. Preregistration is increasing in popularity in part because we do not trust the predictions of theorists (that are often post-hoc) in the same way we trust predictions of theories (Borsboom, 2013). Strong theories enable us to test what would happen in situations that are not actually realized. For example, we know quite a bit about skyscrapers and earthquakes, and can test what would happen to a specific skyscraper under a specific earthquake scenario *in theory* (e.g. via a computational model), allowing the construction of better skyscrapers. Imagine we could do such a thing in clinical psychology: testing the effect of a treatment without actually conducting a clinical trial!

Most psychological theories are weak theories. Exceptions are found only in few disciplines, such as cognitive and mathematical psychology (Forstmann et al., 2011; Palminteri et al., 2017; Townsend, 2008). The psychological theory glass is, I conclude, at the very best half empty.

1.2 The present paper

Along with improvements to psychological science, the last decade has seen innumerable publications that feature factor models and network models. The present paper discusses how lack of attention to theory in both fields has led to problematic inferences and best practices. I chose these two fields not only because factor and network models are statistically closely related—they are also used broadly in the context of theoretical work that explains the same phenomena with competing causal explanations.

[2] Note that I use weak and strong theories as descriptive rather than evaluative terms, and conceptualize them as extremes on a continuum (e.g. how precisely and unambiguously did you spell out your theory).

I start by providing a brief introduction to factor and network models, and demonstrate that both were developed in the wake of theories about a broad set of psychological constructs, including personality traits, cognitive abilities, and mental disorders. At least for factor models, testing theories is not their primary use case today. The introduction is followed by describing ways the fields fall short of theory building and testing. First, I discuss the *conflation of statistical and theoretical models*, which threatens valid inferences in both disciplines. Psychologists regularly interpret statistical models *as* theoretical models. I showcase in a simulation study that a well-fitting statistical model does little to corroborate a theory, independent of the quality or quantity of data. Second, I discuss a common type of empirical contribution where authors do not spell out a theory, but use implicit beliefs or causal assumptions to guide inferences. I refer to this as the problem of *latent theories* in the remainder of the paper. Third, I discuss the problem of *weak theories* and its consequences in the literature. I conclude with a call for stronger theories, and some steps towards achieving this goal.

My argument is not that there are no strong theories in psychology, or that most researchers use factor or network models incorrectly. I am also not arguing that sound exploratory research that is reported as exploratory rather than confirmatory is not useful to establish robust phenomena in the first place. My claims instead are that the core issues identified here—latent theories, weak theories, and conflating theoretical and statistical models—are common and harmful, facilitate invalid inferences, and stand in the way of theory failure and reform.

2. Factor and network models: a primer

> “Heavy reliance on statistics is a poor route to scientific insight.” (Vaidyanathan et al., 2015)

Factor models require little introduction, and represent a sophisticated statistical framework (Brown, 2015; Hoyle, 2012; Kline, 2015). The most common factor models in psychology are the exploratory factor model (EFA) and the confirmatory factor model (CFA), which I will focus on in the remainder of the paper. These models are widely used because they are statistically well understood, and have many extensions that enable researchers to fit them to a variety of data. They can be estimated easily and quickly, and have a broad number of applications, from scale development over measurement invariance testing to conveniently reducing a larger number of observed items to a smaller number of latent variables.

Network psychometrics, on the other hand, is a relatively recent discipline that emerged around 2014 (Epskamp, 2017). Network psychometric models estimate conditional dependence relations among variables with the goal to guide causal inference[3]. Psychometric networks have been applied to data of numerous psychological constructs in the last few years, including mental disorders (Contreras et al., 2019; Fried et al., 2017; Robinaugh, Hoekstra, et al., 2019), personality (Beck & Jackson, 2019; Mõttus & Allerhand, 2017), attitudes (Dalege et al., 2015), cognitive abilities (Kan et al., 2019), empathy (Briganti et al., 2018), emotions (Lange et al., 2019), attachment (McWilliams & Fried, 2019), and resilience (Fritz et al., 2019). I see three reasons for this rapid increase of publications. First, scholars have long highlighted the importance of complexity to understand psychological phenomena, especially in clinical sciences where progress in understanding mental illness has been hampered by oversimplification and reductionism (Borsboom et al., 2019; Engel, 1977; Miller, 2010); network models may allow to map out this complexity in some more detail. Second, psychometricians have discovered network models as an important statistical topic (Beltz & Gates, 2017; Bringmann et al., 2013; Epskamp, 2017; Epskamp & Fried, 2018; Kruis & Maris, 2016; van Borkulo et al., 2014). Finally, free software and tutorial papers have enabled applied researchers to compute network models in little time and removed barriers for broad use.

2.1 Factor and network models are rooted in theory

Factor models represent the shared variance of a set of observed items in one or more latent variables. They can be traced back to Spearman (Spearman, 1904), who used correlational analysis to identify *general intelligence*—often called the *g* factor today—based on evidence that children

[3] Note that network psychometric models differ from other types of network models such as social networks in that they *estimate* relations among variables (Epskamp, Borsboom, et al., 2018; Wasserman & Faust, 1994).

who do well in an intelligence task in one domain also do well in other domains. Spearman interpreted this statistical construct as a psychological construct he termed *mental energy*, which causes a proportion of variance on observed test scores: a child does well on many cognitive subtests *because* she has high mental energy. Researchers after Spearman have proposed similar causal accounts for diverse psychological constructs, and used factor analytic tools to provide statistical estimates of such constructs. Caspi & Moffitt suggested that the *p* factor of psychopathology is a causal psychological construct that influences the degree to which people experience mental disorders broadly (Caspi & Moffitt, 2018). And McCrae & Costa posited that personality "traits as underlying tendencies cause and thus explain […] the consistent pattern of thoughts, feelings, and actions that one sees" (McCrae & Costa, 1995). These and other theories suggest that psychological constructs serve as *common causes* for observed psychological variables, and correlations among items are explained by one shared causal origin (Schmittmann et al., 2013) (Figure 1, left).

Network theory (Figure 1, right) suggests that correlations among items for the *g* factor (Kievit et al., 2017; Savi et al., 2019; van der Maas et al., 2006), *p* factor (Borsboom, 2017; van Bork et al., 2017), or personality traits (Cramer et al., 2012; Mõttus & Allerhand, 2017) stem from causal interactions between items rather than from one shared origin. Intelligence subtests are correlated because being higher in one domain leads to increases in the others, and symptoms of a given mental disorder are correlated because experiences of some lead to experiences of others through direct causal pathways (e.g. rumination → insomnia → fatigue).

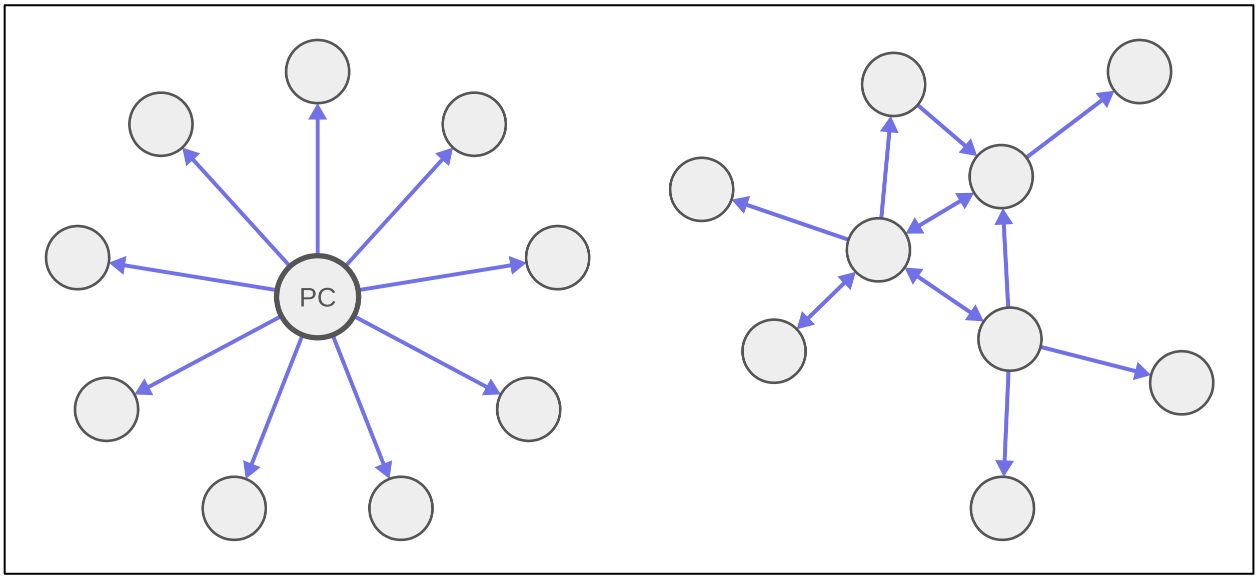

Figure 1. Two competing theories that explain the correlations among a set of items. Left: In the common cause theory, the psychological construct (PC) causes constructs we observe as correlated variables—they are correlated due to their shared origin. Right: According to network theory, constructs cause each other, leading to direct causal relations among items, and the PC is an emergent property resulting from these interactions.

3. The conflation of theoretical and statistical models

> “A huge proportion of the verbal claims made in empirical psychology articles turn out to have very little relationship with the statistical quantities they putatively draw their support from”. (Yarkoni, 2019)

The first problematic way in which the literature falls short of theory building and testing is the conflation of theoretical and statistical models. I believe the term ‘model’ has resulted in a considerable amount of obfuscation in psychology, and we need to clearly denote what model we mean. In the following sections, I first showcase the fundamental gap between statistical and theoretical models, followed by a discussion of common, invalid inferences in the factor and network literature based on conflating statistical with theoretical models.

3.1 Statistical equivalence and the inference gap

Network theory is based on the guiding principle that psychological constructs are emergent properties that arise out of the interactions of constituent elements. Statistical models, on the other hand, specify mathematical relationships among a set of variables. In the realm of network psychometrics, examples are the Gaussian Graphical Model (Epskamp & Fried, 2018), the Ising Model (van Borkulo et al., 2014), or Group Iterative Multiple Model Estimation (GIMME) (Beltz & Gates, 2017). Similarly, the common cause theory has been put forward to explain relations among items, and factor models are often used to test this theoretical model.

In data where columns are variables and rows are participants, I can usually estimate network and factor models without problems. Even if I know that the data generating mechanism is a network—for instance, because I simulated the data myself—I will not get arrested by the psychometrics police for fitting the 'wrong' model (i.e. a factor model) to the data, and vice versa. And there may even be scenarios under which this makes sense. For instance, there is preliminary evidence that when data are generated under a factor model, statistical tools from network psychometrics may be able to recover the number of latent variables more accurately than tools from structural equation modeling (Golino & Epskamp, 2017).

The main concern is not model fitting, but inferences authors draw from statistical parameters obtained. This is because in cross-sectional data, which make up the vast majority of applied factor and network papers in psychology, each factor model has a corresponding network model with (roughly) equal fit to the data. This statistical equivalence relation has been known for well over a decade (Epskamp, Fried, et al., 2018; Kruis & Maris, 2016; van Bork et al., 2019; van der Maas et al., 2006), and I want to showcase an example here.

Figure 2 left presents a causal graph under which I simulated data for 10,000 participants. Fitting a unidimensional CFA to the simulated data provides excellent (CFI=0.996, TLI=0.995, RMSEA=0.012), shown in Figure 2 right. Likewise, simulating under this CFA and fitting a network model to the data results in excellent fit (CFI=1.0, TLI=0.999, RMSEA=0.019; details see supplementary materials).

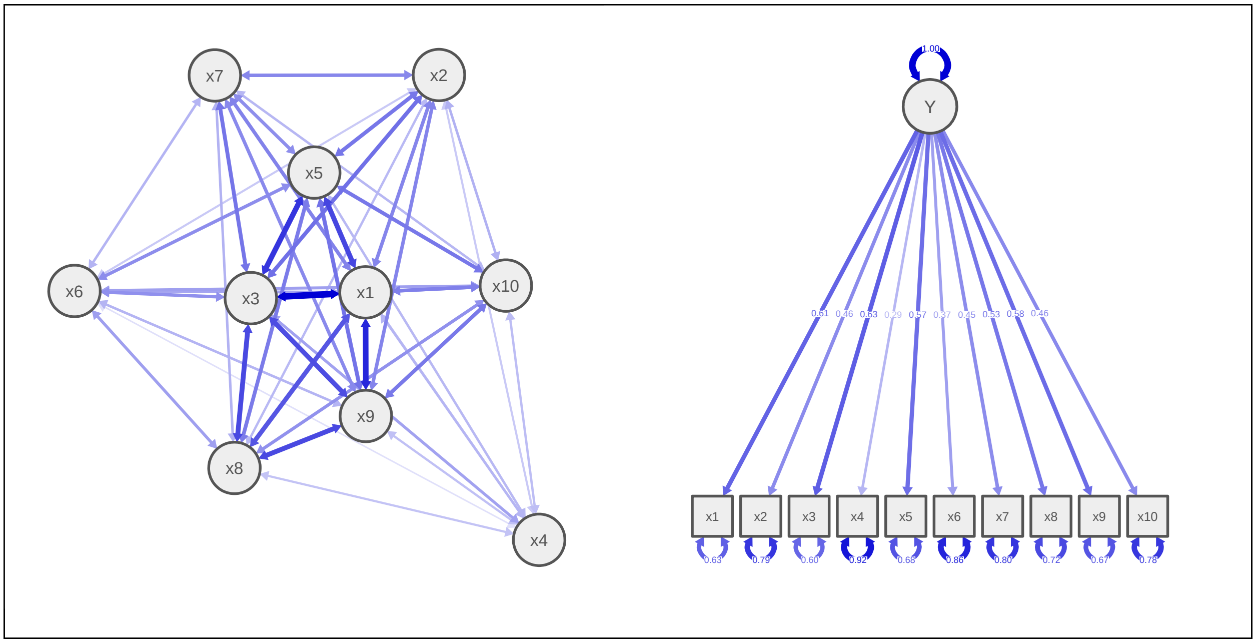

Figure 2. Simulation example of statistical equivalence, i.e. the fact that network models have an equivalent factor model with equal fit, and vice versa. I simulated n=10.000 observations from the causal system shown on the left side and the factor model on the right side. In both cases, fitting the alternative statistical model to the data leads to models with excellent (and near equivalent) fit.

The important message here is that models with excellent fit tell us little about the data generating mechanism due to statistical equivalence. A well-fitting factor model cannot be taken as evidence or proof that a psychological construct such as the *p* factor or *g* factor exists as a psychological construct (Kovacs & Conway, 2019; Vaidyanathan et al., 2015; van Bork et al., 2017). Similarly, a network model with good fit to the data cannot tell us where we need to intervene in a causal system—e.g. on x3 in Figure 2, which has the highest centrality (i.e. interconnectivity, operationalized as sum of all connecting edges)—because the data may have been generated under an alternative causal model such as the common cause model, in which case intervening on x3 would not change the other variables because they are independent given the common cause.

Three important notes before concluding this chapter. First, the problem of statistical equivalence also plays out within the model families listed above, not only across the families. For example, the bifactor model and correlated factors model are equivalent under a set of conditions (Greene et al., 2019), establishing the same inferential problems as discussed above. Second, temporal data can help eliminate some equivalent models, since it makes little sense to orient arrows backwards in time, but they do not eliminate the problem of statistical equivalence in principle (Raykov & Marcoulides, 2001). Third, many models are no longer equivalent under interventions (Fried & Cramer, 2017). Under a common cause model in the strictest sense (Figure 1 left), experimentally intervening on a depression symptom cannot lead to changes in other symptoms because symptoms are unrelated conditional on their shared origin. This is similar to tuberculosis, where treating a person's cough will not reduce their fever, but treating the shared origin of the symptoms via antibiotics will eliminate both symptoms. Network theory (Figure 1 right), on the other hand, posits that such interventions on relevant parts of the system *must* affect the rest of the system, and there are some plausible examples of such interventions (Blanken et al., 2019; Fried et al., 2015; Kievit et al., 2017). It is currently unknown where and how to precisely intervene on causal systems, e.g. on relations among variables, or on variables directly (Fried & Cramer, 2017). There is some initial evidence to support interventions on central symptoms (Elliott et al., 2019; Robinaugh et al., 2016; Rodebaugh et al., 2018); however, future behavior of complex systems is difficult to predict, and it will require theoretical, methodological, and experimental work to achieve actionable progress moving forward (Henry et al., 2020).

There are many reasons why we cannot escape theory: statistical equivalence is one of them.

3.2 Problematic inferences in the factor literature

Latent variables like the *g* factor and *p* factor sometimes refer to statistical constructs, at other times to psychological constructs. Authors have concluded that empirical results provide support for these general factors, but it is often unclear what type of claim is made (Kovacs & Conway, 2019; Vaidyanathan et al., 2015; Watts et al., 2019). Statistical support means that data can be summarized in a model with one latent variable that has decent fit to the data. Since data can *always* be summarized this way if the underlying correlation matrix features a positive manifold (Kovacs & Conway, 2019; van Bork et al., 2017), this tells us nothing about what *g* or *p* are as psychological constructs. Theoretical support, on the other hand, means that there is evidence that *g* or *p* exist as something other than statistical parameters, i.e. that authors have discovered processes or mechanisms—maybe similar to syphilis that was a latent disorder until bacteria in the brains of syphilitic patients were identified in the early 20th century, strongly corroborating the disease *as disease* (Noguchi & Moore, 1913).

Conflating these levels is common in the literature. A recent paper argued that "research into covariance structure has provided evidence of an overarching, general psychopathology factor [that] reflects an underlying liability to experience all forms of psychopathology" (Carragher et al., 2016). The statistical *p* factor is indeed well replicated, but what this factor reflects cannot be determined by observing a covariance matrix. As explained by Kovacs & Conway (Kovacs & Conway, 2019), such statements provide "an illusion of explanation: Positing a general factor gives the false impression that there is a psychological explanation, whereas the actual explanation is purely statistical". The above mentioned paper concludes that the "replication of a general factor underscores the importance and utility of transdiagnostic treatment" (Carragher et al., 2016). But the statistical model does not provide conclusive information about the data generating (i.e. causal) mechanism, and inferences about treatment based on such data therefore are speculative.

Recent work has also concluded that the literature supports the bifactor model in the realm of psychopathology, but it remains unclear how data can support a statistical model when multiple equivalent models exist with the same fit to the data. If the authors mean that data support a psychological construct, it remains unclear what this construct is (Greene et al., 2019; Watts et al., 2019). The incoming editor of the journal *Assessment* highlighted the interpretation of bifactor models as one of core issues for the journal moving forward, stating that "caution is encouraged about accepting a bifactor solution solely on the basis of improved fit statistics", and that it is important that "authors that examine a bifactor solution do so with a clear eye toward the conceptual meaning of the general and specific factors, as well as the practical implications of such a model" (Samuel, 2019).

A third example is the fact that intelligence or mental energy are often conceptualized as an intraindividual process, whereas the statistical evidence for the positive manifold and the *g* factor is largely based on interindividual statistical models (Borsboom et al., 2009; Jensen, 2002).

In sum, latent variables and psychological constructs are not the same kind of thing and should not be equated without spelling out all causal assumptions that are necessary for this mapping (Markus, 2008; Spanos & Mayo, 2015; Vaidyanathan et al., 2015). Watts et al. (2019) navigate this problem well: "We refer to the general factor of psychopathology as the *p* factor when discussing it as a substantive construct and as the general factor when discussing it as a methodological construct" (Watts et al., 2019).

3.3 Problematic inferences in the network literature

Researchers in the network literature have at times used the terms "network approach" or "network framework" ambiguously, i.e. without denoting if they mean theory or statistics—a crime I am guilty of as well. Clearly separating out these two models is necessary (but not sufficient) for valid inference.

For example, some statistical models impose assumptions on the data that do not align with theory, and it is unclear how results from such models can inform theory. Take directed acyclic graphs (DAGs) that are increasingly being applied to cross-sectional, clinical data (Robinaugh, Hoekstra, et al., 2019). DAGs are *acyclic* network models, which means that they cannot accommodate feedback loops: if you start at a given node and follow a directed edge, you can never get back to the original node (this differs from the cyclic graph in Figure 1 right). However, network theory is based at its core on the idea of self-sustaining interactions among problems, such as the vicious cycle between perceived threat and physiological arousal that can give rise to panic attacks (Borsboom, 2017; Robinaugh, Haslbeck, et al., 2019). It is unclear how the results of DAGs in such data can provide meaningful tests of predictions derived from network theory. Overall, I see four inferential challenges in the literature.

First, what appears to support network theory may only do so when ignoring statistical equivalence. For example, reductions in central symptoms predict stronger reductions in the other symptoms than reductions in peripheral (i.e. weakly connected) symptoms (Elliott et al., 2019; Robinaugh et al., 2016; Rodebaugh et al., 2018). This has been taken by some to support network theory—after all, reductions in central symptoms should deactivate numerous other symptoms and reduce overall severity more than reductions in peripheral symptoms. However, the data are equally consistent with a common cause explanation, given that the most central item in network model corresponds to the most reliable indicator of a latent variable, i.e. the item with the highest factor loading (Hallquist et al., 2019; Robinaugh, Hoekstra, et al., 2019).

Second, preliminary simulation work has demonstrated that common network psychometric models may be unable to recover some processes at the heart of network theory, such as feedback loops, higher-order interactions, and asymmetric relations (Haslbeck et al., 2019). This throws a considerable wrench into the statistical model → theoretical model inference works in applied network modeling, and highlights the inferential gap between theoretical and statistical models.

Third, network theoretical accounts of psychological systems (at least arguably) describe within-person processes—e.g. problems cause other problems across time in a given person—but most of the applied literature has estimated statistical models based on between-subjects data. Critics have pointed out this mismatch: identifying a relation between sleep problems and sad mood at the between-person level provides little information of this relation at the within-person level (Adolf & Fried, 2019; Bos et al., 2017; Fisher et al., 2018; Molenaar, 2004).

Finally, statistical parameters can be hypothesized to map onto causal processes, but they are not the same kind of thing, much in the same way psychological constructs and latent variables are not the same kind of thing. Causal conclusions require statistical models *and* causal assumptions, which should be spelled out clearly (Pearl, 2000). The idea that activation 'spreads' through a system is an inference that follows from current network psychometric models only under strict (and, some argue, unrealistic) assumptions (Borgatti, 2005; Bringmann et al., 2019). Similarly, a central symptom in an estimated network model may guide hypotheses regarding intervention, but requires strong causal assumptions for inferences about interventions (Bringmann et al., 2019; Fried et al., 2018; Henry et al., 2020; Robinaugh, Hoekstra, et al., 2019). And claims that central items "constitute the 'backbone' that sustains depressive symptoms in late-life" do not follow from data without assumptions (Belvederi Murri et al., 2018), e.g. that the estimated network maps onto the data generating model in a meaningful way. Or suppose we use a statistical time-series model and identify that A Granger-causes B (simplified: A precedes B and statistically relates to B; (Granger, 1969)). This has been taken to mean that intervening on A will lead to reductions in B. However, such inferences only hold under the assumptions that the estimated relation between A and B largely maps onto the data generating mechanism, which is unknown. Biased estimates could come from numerous sources, such as measuring A and B at the wrong time resolution; failing to include all important causal processes (A and B could be independent and both caused by C); measurement problems, which is likely in the field of ecological momentary assessment for which there is little work on scale validation; and processes such as response shift bias, which can occur when participants answer the same items repeatedly for a hundred times over the course of a few weeks.

3.4 Statistical models are not theoretical models

The literature faces a serious inferential problem. Theories are generally imprecise, which we will discuss in detail below, and it is often difficult to determine whether data corroborate a theory or not. As Lykken put it: "The process of planning, conducting, and analyzing any psychological experiment are complicated, frequently demanding decisions that are so weakly informed by any ancillary theory [...] as to seem essentially arbitrary" (Lykken, 1991). But if many of our decisions have little to do with our theories, how can we then use the results of statistical models to draw inferences about the theories we set out to study—how can we bring the data to bear on our theories? Despite these challenges, results of statistical models are often taken to corroborate theories. The most common (and invalid) inferential shortcut is to interpret the statistical models *as* the theoretical model.

Psychology is not alone in this struggle. A recent paper entitled "Can a neuroscientist understand a microprocessor" provided a demonstration of this inference gap by using modern tools of neuroscience to investigate a chip (Jonas & Kording, 2017). The authors were unable to recover the way information is processed in the chip: the statistical model had little to do with the data generating process, and taking the statistical model *for* the data generating model led to false conclusions.

Importantly, statistical equivalence and the inference gap hold *regardless* of the amount or quality of data. Statistical models can only tell us so much.

4. Latent theories and baking psychometric cakes

> "It is a paradox of scientific method that the branches of empirical science that have the least substantial theoretical developments often have the most sophisticated methods of evaluating evidence." (Suppes, 1962)

The majority of empirical publications in the network field, including some of my own work, have followed similar baking recipes, and resulted in fairly similar network cakes. Take an interesting and usually cross-sectional dataset[4]; bake the dataset in the network model oven; and sprinkle some network inference on top, such as investigations of centrality or connectivity.

It is reasonable to focus on descriptive, exploratory, hypothesis-generating work when a new field emerges, with the goal to establish robust phenomena worth explaining. For example, network structures of mental health data are rarely homogeneous (i.e. do not feature a 'boring' topology where everything is connected in the same way); can be described by both a positive manifold and a conditional positive manifold; and between-subjects conditional dependence relations appear consistent across time (Robinaugh, Hoekstra, et al., 2019). But no matter how much importance one ascribes to thorough descriptions of data, such empirical efforts generate explananda, not explanantia.

[4] A systematic review summarizing the network literature between 2008 and 2018 showed that 141 of 173 (81.5%) empirical clinical papers utilized cross-sectional data.

The network field aligns with the applied factor model literature that has baked similar psychometric cake for much longer. Psychologists have published thousands of papers utilizing factor models in cross-sectional data, which has also led to some robust phenomena. In clinical psychology, for example, the first eigenvalue tends to explain considerably more variance than subsequent ones; unidimensional factor models often have decent fit, although extracting additional factors tends to improve fit; and multidimensional factor solutions rarely follow simple structure. These and related phenomena remain to be explained. Similar to the network field, the factor literature has in large part not moved beyond the initial exploratory research stage, and the name of the most commonly used model—the *confirmatory* factor model—is a misnomer, given that CFAs are often adapted based on fit to the current data (Crede et al., 2019). Overall, empirical work usually leaves me wondering on what theoretical grounds authors opted to use a factor model in the first place, and how identified latent variables inform theory formation (i.e. what it *means* that authors identified three depression factors).

An anonymous referee criticized a network-analytic paper we had submitted, and the criticism is well taken: "The authors […] do not provide a compelling rationale for using this technique. Can the authors show a substantial increment or special leverage offered by this method? What theoretical or scientific question is at stake here?" Much of the applied network literature, including some of my own work, leaves these questions unanswered. Similarly, most applied factor analytic papers do not feature a compelling rationale for the use of factor models; the theoretical and scientific questions at stake remain unclear; and it is questionable in which way such research offers a substantial increment or special leverage. We should take these concerns very seriously.

4.1 Latent theories in the factor literature

Research aims, approaches, and conclusions should align. If our goal is to test or update a theory, we investigate whether the theory's predictions are realized in data; selecting appropriate statistical models that impose the right kind of assumptions on the data is important here. If our goal is data exploration and hypothesis generation—i.e. to establish robust phenomena—

statistical models may be chosen for different reasons, such as their value to map out data thoroughly. Psychologists use factor models for both purposes, but rarely spell out their goal transparently, impeding meaningful inferences[5]. We can think of this problem as one of *syntactical equivalence*, a situation where at least two different causal interpretations of the same statistical model exist (Markus, 2004); in our case, one is a specific causal interpretation, the other is agnostic.

If authors hold a realist and causal belief about a psychological construct—similar to Spearman's view of the *g* factor—choosing a CFA where variations in the latent variable lead to variations in the observed indicators is sensible. Describing this transparently encourages theoretical debates and attempts for alternative explanations of a phenomenon, which can lead to successful theory failures and facilitate progress. If, on the other hand, researchers are agnostic about the causal processes and use a CFA to summarize data, providing this rationale is equally important, because it allows readers to debate whether other statistical models may be more appropriate to summarize data (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000; Widaman, 1993), and whether the concept of measurement error that comes with using a CFA is sensible when summarizing data (Borsboom et al., 2003; Rhemtulla et al., 2018; van Bork et al., 2017). I have read well over 100 papers using factor models in depression data: most work fails to describe what it means for a theoretical account of depression that authors identified a number of factors, or how to interpret residual correlations or cross-loadings from a substantive perspective, or what theoretical motivation justifies extracting orthogonal latent variables via procedures such as varimax rotation. But neither do authors state that their goal is data exploration.

I describe this body of work featuring *latent theories* rather than *absent theories* (I take no issue with thorough exploration of data), because researchers often appear to hold causal beliefs that are not explicated. There are two reasons why I believe this to be the case. First, psychologists usually state that latent variables 'explain' the covariation of items, that they 'directly influence' the items, or that 'underlying' latent variables were identified. This implies a latent causal theory

[5] For an example of a statistical justification for the use of exploratory structural equation models, see (Fried et al., 2016).

akin to a common cause explanation: we would not claim that socioeconomic status (SES) 'explains' the covariation between neighborhood and income, 'directly influence' these items, or 'underlies' the items. Instead, SES is constructed based on the items, and features as a textbook example for a formative latent variable where variations in the items lead to variations in the latent variable (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000). Second, the common cause interpretation is the most sensible one given the reflective model, which features causal arrows from the latent variable to the observed indicators (Bollen & Lennox, 1991; Borsboom et al., 2003; Edwards & Bagozzi, 2000; Van Bork et al., 2017). The model was developed to test common cause theories, and disregards variance that is not shared among indicators as measurement error, which makes sense only under a causal-reflective perspective. Note that latent theories may be weak or strong theories—maybe researchers have an elaborate formal model on which they base their ideas on—but because the theories are not spelled out explicitly, this cannot be evaluated by readers.

I am not alone in interpreting the majority of the literature as implicitly causal. A recent evaluation of the *p* factor literature concluded that "tacitly implied in much of the existing research is that the *p* factor reflects a singular causal mechanism" (Watts et al., 2019). This causal account was explicated in recent work (Caspi & Moffitt, 2018): "correlations between symptoms (as well as disorders) arise from a *g*-like causal factor". Such transparent causal proposals matter for three reasons. First, they enable theoretical debates, disagreements, and falsification or modification attempts. Second, they guide future research: if the *p* factor is indeed a singular causal mechanism that explains comorbidity, the next logical steps may be to find *p*, investigate its neurobiological basis, and try to intervene on it. If, instead, *p* is a sensible and useful summary of the problems people tend to have—similar to SES, a useful formative construct that predicts mortality and morbidity (Marmot, 2015)—aiming to identify biological markers for *p* may make as much sense as trying to find the biological underpinnings for SES (Turkheimer, 2016). This is because formative constructs are pragmatic summaries of data: they can be useful, but remain convenient fictions. And third, if we fit a model to the data that does not correspond to the data generating model, we end up with biased estimates due to statistical misspecification, i.e. because we use a statistical model that imposes invalid assumptions on the data (Spanos & Mayo, 2015). If data are generated under a formative model, fitting a reflective model leads to biased parameter estimates (Rhemtulla et al., 2018)[6].

4.2 Latent theories in the network literature

It is difficult to find applied network papers that do not explicitly refer to network theory, i.e. the notion that observed correlations among items psychological science stem from mutual causal interactions. Latent (i.e. implied) theories rarely provide a problem. This transparency allows for evaluating whether statistical models are appropriate for the theoretical questions under investigation, a challenge discussed in detail in the section on the relation between statistical and theoretical models above. Network psychometric models used may often not lend themselves well to testing the theories put forward; this is the case when theories are within-person and statistical models between-person, or when statistical models are unable to recover theoretically proposed processes such as feedback loops (Haslbeck et al., 2019).

4.3 Latent theories lead to problematic best practices

Lack of attention to theory and a focus on sophisticated methodology have led to problematic best practices (Lilienfeld & Pinto, 2015; Vaidyanathan et al., 2015). I want to demonstrate this using an example from the field of measurement.

The common cause idea is deeply entrenched in the way researchers measure psychological constructs: We query people on a set of items that we think are good indicators for a construct, with the goal to obtain a unidimensional questionnaire, maximizing reliability (in the sense of internal consistency, i.e. all our items are good measures of the same construct) while maintaining validity (i.e. our items ask questions relevant for conceptual definition of the construct). Items that show low correlations with other items are considered "bad" items, and usually deleted. But bad

[6] While this may be unavoidable—statistical models may often be simplifications of phenomena under investigation—this is no excuse for causally mis-specifying the hypothesized relation between the psychological construct and observed items.

items from a common cause perspective are not necessarily bad items from other causal perspectives.

Consider the causal system in Figure 3A, where x1 plays a crucial causal role in activating the rest of the system; x1 could represent anxiety, which may lead to a cascade of worrying, fatigue, sadness, and sleep problems. Simulating 10.000 observations from this graph produces a dataset where our 5-item scale has a Cronbach's alpha of 0.71. This increases to 0.77 if item x1 is dropped from the scale, which would follow best practices in the field—but we would not want to drop x1 from the scale just to increase scale reliability, given that we know its causal importance; in fact, knowing the system, we would not want to drop any item from this scale. Reliability is not the right way to think about our scale. Similarly, x1 has the lowest factor loading of all indicators (0.1) when we fit a unidimensional factor model to this dataset simulated under the causal graph in Figure 3A. Common sense would dictate we should remove x1 (Figure 3B).

Dropping causally important variables based on internal consistency and factor loadings is not an unlikely scenario in psychology. For instance, psychometric work in a 10-item attachment scale identified 2 factors, anxiety and avoidance, but one item performed poorly, because it did not load clearly on either of the two factors (Fraley et al., 2011)[7]. I showcase an example of a 2-factor model estimated on data for this scale in Figure 3c, where x10 is the item with cross-loadings. This item was removed from the scale, in part because it improved the reliability of each subscale, and was therefore not assessed in subsequent studies. However, from a network perspective, items that load on two factors simultaneously make for the potentially most interesting items, because they may build causal bridges between two communities of items (Figure 3D showcases the network model estimated on this dataset, with x10 in the center of the graph) (McWilliams & Fried, 2019).

[7] This empirical demonstration is an example where the true causal model is unknown, and should not be taken as criticism of the cited work, whose lead author has been exceedingly helpful and shared data with us for the above investigation.

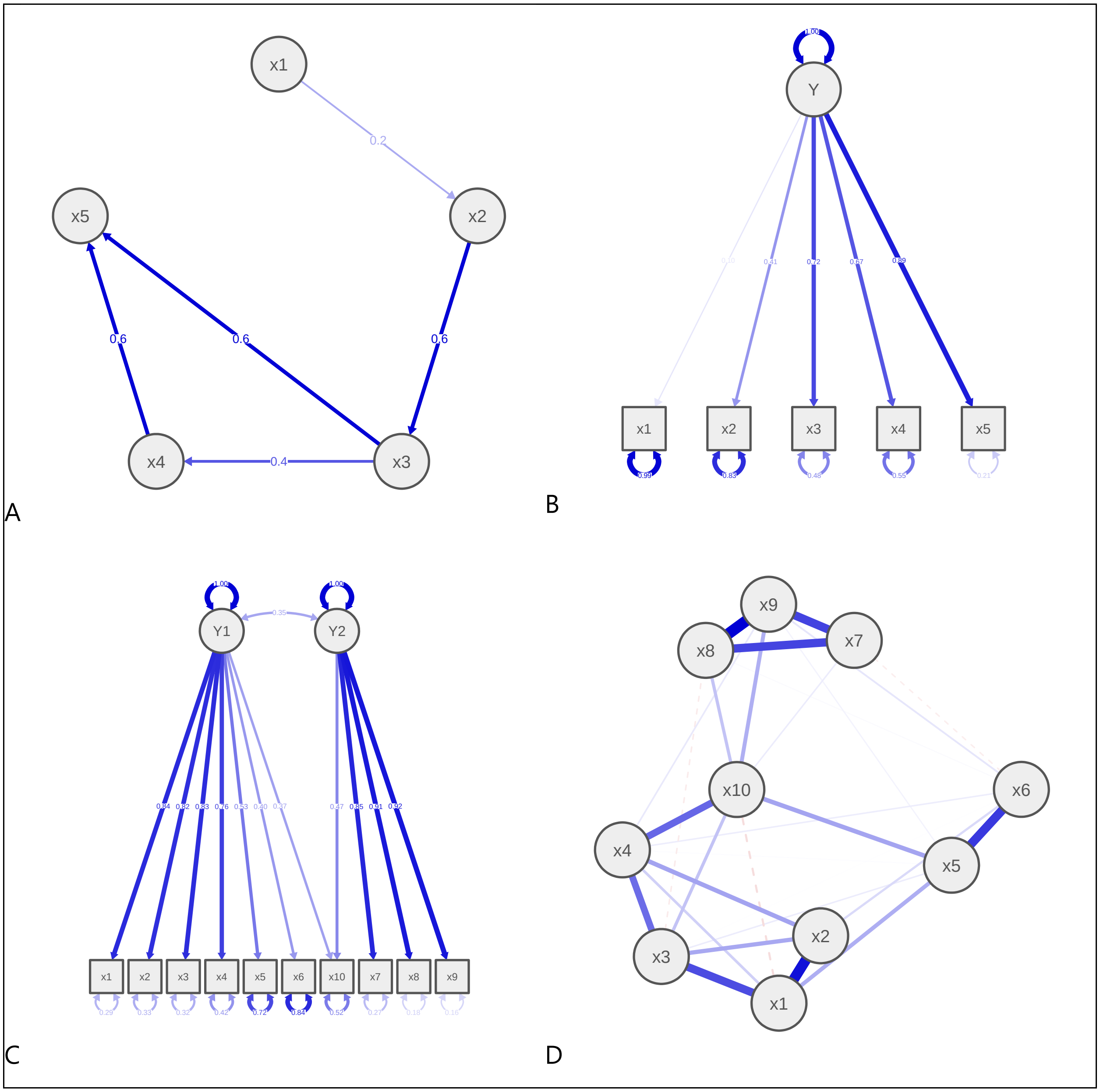

Figure 3. A: Example of a causal system from which I simulated n=10.000 observations. B: x1 has the lowest factor loading (0.1) when fitting a unidimensional CFA to this dataset, and dropping item x1 in the data would usually be recommended because it would increase the scale's internal consistency from 0.71 to 0.77. However, dropping x1 would be misguided, knowing the true causal structure. C: 2-factor model estimated in a 10-item attachment scale (n=310) in which x10 shows cross-loadings on both factors. The item was identified as problematic item from a latent variable perspective and dropped from the scale, increasing the reliability of the subscales. D: Network model of the same data, where x10 could be hypothesized to build a causal bridge between the other items (from: (McWilliams & Fried, 2019)).

While assessing internal consistency as a tool to guide scale construction—similar to the development of latent variable models—was based on the idea of psychological constructs as common causes, internal consistency has become a thing by itself, used without regarding for theoretical considerations. It is often utilized today as a synonym for scale quality in applied social sciences, rather than as a statistical tool that can be useful given a particular theory, based on a latent theory of psychological constructs. Cronbach & Meehl summarized this point well: "High internal consistency may lower validity. Only if the underlying theory of the trait being measured calls for high item intercorrelations do the correlations support construct validity" (Cronbach & Meehl, 1955).

Overall, latent theories pose a dangerous and underappreciated threat to valid inferences in our field, in part because they blur the lines between statistical and theoretical models[8]. The *p* factor literature provides an interesting example in this regard, where initiatives like the Hierarchical Taxonomy Of Psychopathology (HiTOP) are primarily concerned with *describing data*, but solely use one particular type of statistical model—the reflective latent variable model—that imposes serious assumptions on the data (Van Bork et al., 2017; Watts et al., 2019).

5. Why weak theories are poor theories

> "The present state of knowledge in psychology is very broad but very shallow. We know a little bit about a lot of things". (Lykken, 1991)

Theories in the applied factor and network literature are often weak theories. I defined weak theories in the introduction as theories that are narrative and imprecise descriptions that are vulnerable to hidden assumptions and unknowns. They provide neither precise explanations nor predictions, and it can be difficult to find out if data support weak theories or not. Precise predictions are related to what Meehl called 'risky' predictions: predictions more likely rejected rather than corroborated with more data (Meehl, 1990a). Suppose Theory 1 predicts traffic jams on a highway for each hour in upcoming week, and Theory 2 predicts that there will be at least

[8] I want to acknowledge a different position here as well: the *Tools to Theories* approach argues that the broad use of a particular type of statistical model can lead to new theoretical metaphors and concepts (Gigerenzer, 1991).

one traffic jam in the next week. Assuming both theories are false, Theory 1 is more likely rejected, Theory 2 more likely corroborated with more data. Psychological theories are similar to Theory 2, predicting a significant correlation between two variables or a significant group difference, rather than predicting the magnitude of the effect. Another aspect of weak theories is that they predict observations also predicted by other theories, in which case theories are fully underdetermined by data. That is, if two theories correctly predict identical traffic jams over the course of the next week, we cannot justify our choice for one theory over the other based on the data alone[9]. Meehl argued that theories get "money in the bank […] by *predicting* facts that, absent the theory, would be antecedently improbable" (Meehl, 1990a)(my highlight). Mendeleev used the periodic table in 1871 to correctly predict the existence and specific properties of three unknown elements that were discovered within the next 15 years: gallium, scandium, and germanium (Hitchcock & Sober, 2004)[10]. I believe that theories also get money in the bank by *explaining* facts that, absent theory, would be hard to understand. Weak theories will stay poor theories in both aspects.

The generic common cause and network theories discussed above can be considered weak theories. First, they are verbal accounts, and contain many unknowns. Second, common cause and network theories predict correlations among items, and larger samples will make it easier to pick these up these correlations as different from zero, and therefore produce data broadly consistent with the theories. At least in their most generic forms, neither theory makes precise predictions. Third, both theories predict identical cross-sectional data of moderately inter-correlated items (i.e. are underdetermined given data), and these correlations do not allow us to find out what generated the data due to statistical equivalence. The next sections discuss these problems in the factor and network literature, along with work that overcomes these challenges.

5.1 Weak theories in the factor literature

[9] The question whether all theories are underdetermined by data has been discussed in great detail in philosophy and is beyond the scope of this paper (Gershman, 2019). It is conceivable to me that in psychology, one theory may predict and explain aspects of a phenomenon better than a set of specific rival theories, facilitating abductive inference (i.e. inference to the best explanation).

[10] I am not arguing that explanation is necessary for prediction: we have been using tide charts for centuries to accurately predict high and low tide (and continue to do so), the insight that tides are affected by the moon did not increase predictive accuracy (Cummins, 2000). I am instead making the point that there are merits in having explanantia rather than explananda, one of which is that they can facilitate predictions.

Common cause theories have been put forward for numerous psychological constructs such as cognitive abilities (Spearman, 1904), personality traits (McCrae & Costa, 1995), and mental illness (Caspi & Moffitt, 2018). However, strong common cause theories in the applied literature that use factor models are largely absent, and it is therefore unclear to what degree factor models can provide tests for theoretical claims.

Recent work proposed different explanations for the *p* factor. These include the common cause theory where *p* is the shared origin of psychopathology broadly, but also various alternative ideas, e.g. that *p* represents the overall severity of symptoms; the overall liability for mental illness; or other constructs such as neuroticism, impulse control, intellectual function, or impairment (Caspi et al., 2013; Watts et al., 2019). These accounts meet all criteria for weak theories: they are verbal and imprecise, do not provide precise predictions, and do not predict data that is not predicted by alternative accounts (at least I am not aware of work spelling out differential predictions in detail). Many questions remain, such as how empirical evidence would have to look like to warrant theory reform. It is also unclear why the reflective latent variable model in particular that has held such a strong monopoly in this literature in the last decade is the most appropriate model to test these ideas. Statistical sophistication may distract from problems of theory. Box famously bemoaned what he called coolbookery and mathematistry, i.e. "to force all problems into the molds of one or two routine techniques, insufficient thought being given to the real objectives of the investigation or to the relevance of the assumptions implied by the imposed methods" (Box, 1976). Lilienfeld & Pinto argued more recently that the growing popularity of confirmatory factor analysis "may at times engender the illusion of methodological rigor in its absence", and that that "overreliance on these methods may inadvertently generate a misleading sense of comfort with the research status quo and a further reluctance to undertake risky tests of theoretical models" (Lilienfeld & Pinto, 2015). Relying on a limited set of statistical tools can lead to false conclusions (Vaidyanathan et al., 2015), and methodological plurality may help the field stand on firmer ground; this is especially the case because assumptions imposed on the data by reflective latent variable models do not necessarily follow from proposed theories about the *p* factor.

The *p* factor literature follows in the footsteps of work on the *g* factor, and intelligence researchers have had about a century head start in exploring the nature of *g*. In two famous collections from 1921 and 1986, numerous experts put forward their theories and perspectives about intelligence. Jensen summarized both collections, stating that "there are about as many different conceptions of 'intelligence' as the number of experts" (Jensen, 1987). The situation remains the same today, a century after the initial collection, and numerous theories persist that are underdetermined by data (Kovacs & Conway, 2019; Savi et al., 2019).

Stronger theories help resolve these problems. The multiplier model (Dickens & Flynn, 2001) and the dynamic wired intelligence model (Savi et al., 2019) are formal (i.e. mathematical or computational) models that aim to explain robust phenomena such as the positive manifold and the Matthew effect (the rich—or in this context, skilled—get richer, the poor get poorer); unambiguously spell out axioms and assumptions; enable precise predictions that follow from the theory, not the theorists; create novel predictions (the model by Savi et al., for instance, suggests that education has a much stronger role in the shaping intelligence than other models); and allow the simulation of data from the model that can be compared to observed data to see if the model behaves in unexpected ways (Robinaugh et al., 2020). While the scarcity of formal models in this literature leaves the psychological theory glass half empty at best, I hope that formal models that provide crucial opportunities for theory failure and modification will inspire work in other fields such as clinical and personality psychology.

5.2 Weak theories in the network literature

Network theories put forward in empirical papers are often weak as well. A common claim in the clinical literature has been that symptoms cause symptoms, and that syndromes occur because of mutual relations among symptoms. This is difficult to falsify: finding a few examples of symptoms causing symptoms would not strongly support the claim, and finding a few cases of symptoms not causing other symptoms would not falsify it (or require theory reform).

The idea that syndromes are due to causal relations among symptoms is usually contrasted to the common cause theory, which appears to imply symptoms are *solely* correlated due to causal interactions. This is not supported by evidence. Adverse life events and stressors occur commonly before depression onset (Hammen, 2005), and traumatic events often lead to the development of at least a subset of PTSD symptoms (APA, 2013). Attachment insecurity predicts a wide range of mental health problems (Mikulincer & Shaver, 2012), and so do personality traits such as neuroticism (Barlow et al., 2013). There appear to be important common causes that act as vulnerability for the onset of symptoms broadly. These common causes may be local (rather than global) common causes, in that they activate only a subset of symptoms (Fried & Cramer, 2017), which is somewhat inconsistent with the common cause literature at large that understands psychological constructs as *the core driving force* of relations among items. Nonetheless, it is striking that well-established common causes are not modeled or discussed in the large majority of the empirical network literature. While network theorists in principle allow for such common causes as part of the external field (i.e. influencing the symptom dynamics from outside of the system) (Borsboom, 2017; Fried & Cramer, 2017), the external field is absent from most applied work. Curiously, the same holds for the *p* factor literature where *p* is understood as common cause for symptoms, but data that statistical models are estimated on usually contain symptoms and diagnoses, not established common causes.

To escape vagaries of language and imprecise predictions, network theory should define what a symptom is, and provide a complete list of all symptoms, causal symptom relations, time-frames, underlying mechanisms, and conditions (e.g. moderators) under which these occur. We undertook such an effort recently when developing a formalized computational model of panic disorder (Robinaugh, Haslbeck, et al., 2019). This helped me realize how problematic verbal ambiguities are: I had never thought there might be so many different ways two variables can relate to each other until we needed to write their relation out as an equation, and that small changes to parameterization can dramatically impact what data the theory predicts (Robinaugh et al., 2020). The model aims to explain five phenomena such as individual differences in the propensity to experience attacks, key characteristics of attacks, and the efficacy of cognitive behavioral therapy;

quantifies specific causal relations among all variables; and quantifies the time frame and functional form of their relations. Like all formal models in the social science, the model is false in the sense that it is incomplete. It also fails to account for one of the five phenomena we set out to explain: that some people have panic attacks without getting panic disorder.

This work aligns with other contributions in the network literature that have led to riskier predictions. The formal hysteresis model predicts that connectivity among depression symptoms in an intraindividual network moderates the way transitions from a healthy to a clinical state work out: as continuous transition for networks with low connectivity, and abrupt transitions for those with high connectivity (Cramer et al., 2016). And the model of early warning signals predicts that the dynamics of such intraindividual networks will show signs of critical slowing down before they transition from a healthy to a clinical attractor state, defined as (among others) increasing autocorrelations among symptoms (van de Leemput et al., 2014; Wichers et al., 2016). Efforts are on the way to test whether these predictions that would not be expected under a common cause theory are realized in data.

5.3 Embracing complexity

Recall the notion of homoeostatic property clusters that inspired the development of network theory in psychology broadly, where features are proposed to co-occur in nature because the presence of one favors the presence of others (Boyd, 1991; Fried, 2015; Kendler et al., 2011). If we take this notion seriously, attempts of understanding the development and maintenance of personality traits, cognitive abilities, or mental disorders fall short if they do not embrace complexity more broadly. One important aspect that is absent from much of the applied network and factor analytic literature are contextual variables such as the environment. Personality traits dynamically interact with environmental processes, and might often be stable because people occupy environmental niches that serve as reinforcing feedback functions on their personality system (Hopwood, 2018; D. N. Klein et al., 2011; Mischel, 2004; Mõttus & Allerhand, 2017). Cognitive abilities appear to interact dynamically throughout development (Kievit et al., 2017; van der Maas et al., 2006), and such dynamics are likely routed through environmental multipliers: If a person does well in a specific domain, such as math, standing out in this domain might lead to environmental reinforcement and further improvement in this domain (Dickens, 2007; Dickens & Flynn, 2001). In clinical psychology, environmental variables such as stressors often precede the onset of mood and anxiety disorders, maintain psychopathology, and interfere with treatments (Hooley, 2007; Kendler et al., 1993; Paykel, 2003).

Moving forward also means to abandon generic common cause and network theories presented in Figure 1, and embrace the complexity of theoretical and statistical hybrid models that can feature both local common causes and mutual interactions (Epskamp et al., 2016; Fried & Cramer, 2017). After all, there are both obvious local common causes for psychopathology discussed above, as well as obvious causal relations among some problems some patients face, such as rumination → insomnia → fatigue. This calls for more refined theories, and the discussion whether generic network or common theories are the right theories has distracted from embracing this complexity.

Following participants over time can help with some of the challenges, because temporal data have higher diagnosticity (Kellen, 2019). While the generic common cause and network theories in Figure 1 predict the same cross-sectional data, recent publications have worked out differential predictions in temporal data. Network theories predict that a given variable at baseline (e.g. verbal intelligence or depression) should be related to *changes* in other variables over time (e.g. mathematical intelligence or anxiety), termed mutualistic coupling. Such predictions can be formalized via latent change score models and tested in data (Kievit et al., 2017, 2019; Peng & Kievit, 2020).

5.4 Constraints on theory

There is a fierce debate in psychology on what establishes a successful replication of a finding: is it sufficient to obtain a significant effect in the right direction, or should we be able to replicate the point estimate of a finding, such as the effect size of a clinical intervention? Related, some researchers have argued that non-replications are unconvincing, for a variety of reasons such as contextual variables (i.e. the influence of cultural, geographical, and historic context).

What is remarkable in this discussion is that theorists in psychology rarely spell out the boundary conditions under which they expect their effect to hold, but then use contextual sensitivity as explanation when research does not replicate (Gershman, 2019; Simons et al., 2017; Stroebe & Strack, 2014; Van Bavel et al., 2016). Yarkoni recently described this as a crisis of generalizability: researchers often assume that an effect generalizes broadly, when it was conducted under a very narrow set of conditions such as specific participants, measures, stimuli, experimental methods, and analytic strategies (Yarkoni, 2019).

One way forward is for authors to write a falsification paragraph in which they lay out precisely under what circumstances they would consider their theory falsified. Alternatively (and likely more fruitful), using a less black and white Popperian view, authors could describe when they would consider evidence sufficient to warrant adjustments to their theory. Depending on one's philosophy, these adjustments could serve the goal to achieve higher verisimilitude, or to improve explanatory or predictive power. A falsification paragraph that puts clear *constraints on theories*[11] would require stronger specifications of the core findings (what variables matter), auxiliary hypotheses (what variables can we vary that do no not matter), and mechanisms or processes that lead to the observed effect in an experiment (manipulations specifically of mechanism A, but not of other mechanisms, explain the observed effect). It would reveal that psychological theories are often so weak that they do not require modification, no matter the evidence—one can always suggest that there are hidden moderators (Gershman, 2019). The applied factor and network modeling literature should consider such efforts that would improve theory formation and modification.

6. Putting the theory back into psychology

[11] This is similar to recent work calling for a "constraints on generality" section that transparently describes problems with external validity.

> "The solution is clear, graduate students need more offerings of formal theory courses that are taught at the same level and are as up-to-date as the sophisticated methodological courses they routinely take." (Morton, 2009)

In this paper, I discussed three ways in which empirical contributions in the applied factor and network literature fall short of theory building and testing. Theories are usually completely absent, implied (i.e. latent), or weak. Due to this lack and imprecision of theory, statistical models are often interpreted *as theoretical models*, an invalid inference due to well-known problems such as statistical equivalence.

Overall, weak theories do not tell us what to what data to expect given the theory, which makes it difficult to know whether data corroborate a theory or not. They do not allow for testing clear predictions, but rather define a broad space of findings that are consistent with the proposed theory. While such discovery-oriented research is important to establish robust phenomena, it only generates explananda, not explanantia. Further, exploratory research in psychology is often presented as confirmatory, which can threaten valid inferences; after all, imprecise theories can easily be adapted post-hoc by changing assumptions that had not been spelled out, a slippery slope towards unfalsifiability that is discussed in detail in Gershman's piece "How to never be wrong" (Gershman, 2019). For example, data produced by the famous Michelson–Morley experiment in 1887 was inconsistent with the prevailing theory that light is propagated through the ether, which led FitzGerald and Lorentz to adapt properties of the ether post-hoc in a way that exactly fit the new data. This is little different than claims in psychology that unspecified hidden moderators explain non-replications of original findings.

One important pathway moving forward—a complementary pillar to discovery-orientated research—is theory-testing research (Muthukrishna & Henrich, 2019). Such work features explicit theories that provide more precise explanations and predictions, and allows for testing whether predictions are realized in data. Formalizing such theories in mathematical or computational models to escape vagaries of language is a critical step forward, and has been successful in many other disciplines (Epstein, 2008; Lakens & Debruine, 2020; Morton, 2009; Oberauer & Lewandowsky, 2019; Robinaugh, Haslbeck, et al., 2019; Smaldino, 2017, 2019; Szollosi et al., 2019). Formal models not only facilitate thinking clearly about our theories—they also allow to simulate data given a theory. Such data lead to an implied statistical model (e.g. how would correlations among items look like if the theory were true), which can then be compared against the statistical model estimated on the actual data. Divergences between real and implied statistical models can then lead to theory revisions (Haslbeck et al., 2019; Robinaugh et al., 2020). Another advantage of formal theories is that they bring theories into a shared social space: the fact that our model of panic disorder is online means that others can go about modifying, extending, and testing it without having to involve us in the process (Borsboom et al., 2020). Such efforts can be much more difficult with verbal theories, explaining 'adversarial collaborations' in psychology: collaborations where original authors and replicators first have to agree on a large set of boundary conditions and auxiliary assumptions.

As a whole, the applied psychological literature in general seems far away from formal models and strong theories. The first time I encountered the term *theoretical psychology* was in 2014, in a talk that Borsboom gave at the International Convention of Psychological Science. He pointed out that, in contrast to theoretical biology, theoretical physics, or theoretical economics, all of which are widely known pillars in their respective disciplines, there is no dedicated field in psychology concerned with theory formation (Borsboom, 2013). Looking back at my own studies, I learned how to estimate structural equation models in several statistical programs, but the topics theory construction, mapping between theoretical and statistical models, and challenges of inference (induction, deduction, or abduction) did not come up once in my curriculum. In fact, I had never heard about the term 'formal model' before my postdoctoral fellowship. This is not to say that there are no fields in psychology dedicated to building strong theories, such as cognitive and mathematical psychology (Forstmann et al., 2011; Palminteri et al., 2017; Townsend, 2008), or that there are no experts working on these issues (cf. division 24 of the American Psychological Society, 'Society for Theoretical and Philosophical Psychology'). But applied psychologists rarely receive training in theory building. Because I assume that my experiences are not an exception, I put together an introductory reading list of papers, book chapters, and blog posts on the topic of theory construction in psychology, and hope it will be useful for others (https://osf.io/mqsrc).

While the last decade was focused on improving our statistical practices, the next decade of psychological science should be one of improving our theoretical practices. I hope I have demonstrated that this will be necessary, but it is only fair to point out that this will be difficult as well: psychology is concerned with complex phenomena that are notoriously difficult to measure and understand, in part due to pronounced inter-individual differences (Lykken, 1991).

Somewhat surprisingly, since starting to work on this paper in early 2019, numerous contributions in different areas of psychology have identified this crisis of theory as a crucial challenge moving forward (Borsboom et al., 2020; Burger et al., 2019; Gershman, 2019; Guest & Martin, 2020; Haslbeck et al., 2019; Kellen, 2019; Lakens & Debruine, 2020; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Robinaugh, Haslbeck, et al., 2019; Savi et al., 2019; Smaldino, 2019; Szollosi et al., 2019; Van Rooij & Baggio, 2020), and both the *Journal of Abnormal Psychology* and *Perspective of Psychological Science* have opened calls for special issues on theory.

Maybe the theory glass in psychology is half-full, after all?

Acknowledgements

In am grateful to Ashley Greene, Michael Aristodemou, Alexander Shackman, Donald Robinaugh, David Condon, Erikson Kaszubowski, and Denny Borsboom for providing feedback on earlier versions of this draft. I am indebted to Iris van Rooij, Paul Smaldino, Donald Robinaugh, and Denny Borsboom, whose work have been an immense inspiration in structuring my thoughts on issues of theory. I also want to thank members of the HiTOP consortium for insightful discussions in the last years.

Funding statement

None.

Conflicts of interest statement

None.

Biographical note

Eiko Fried graduated in Psychology at the Free University of Berlin, in close collaboration with the University of Michigan. After 4 years of postdoctoral fellowships in Quantitative Psychology at the Universities of Leuven (Belgium) and Amsterdam (The Netherlands), he now works as an Assistant Professor in the Clinical Psychology Department at Leiden University (The Netherlands). Eiko's broad interests are the measurement, modeling, and ontology of mental disorders. Apart from that, he is interested in studying individual symptoms of mental disorders and their network configurations; theoretical and empirical work on complexity in mental health; and improving psychological science through open science practices. Eiko maintains two blogs and has written a number of tutorial papers that aim to help clinical researchers overcome statistical hurdles.

Bibliography

Adolf, J. K., & Fried, E. I. (2019). Ergodicity is sufficient but not necessary for group-to-individual generalizability. *PNAS*, *116*(14), 6540–6541. https://doi.org/10.1073/pnas.1818675116

APA. (2013). *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. American Psychiatric Association. http://books.google.com/books?id=S8p3lwEACAAJ&pgis=1

Barlow, D. H., Sauer-Zavala, S., Carl, J. R., Bullis, J. R., & Ellard, K. K. (2013). The Nature, Diagnosis, and Treatment of Neuroticism: Back to the Future. *Clinical Psychological Science*, *2*(3), 344–365. https://doi.org/10.1177/2167702613505532

Beck, E. D., & Jackson, J. J. (2019). *Network approaches to representing and understanding personality dynamics*.

Beltz, A. M., & Gates, K. M. (2017). Network Mapping with GIMME. *Multivariate Behavioral Research*, *52*(6), 789–804. https://doi.org/10.1080/00273171.2017.1373014

Belvederi Murri, M., Amore, M., Respino, M., & Alexopoulos, G. S. (2018). The symptom network structure of depressive symptoms in late-life: Results from a European population study. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-018-0232-0

Blanken, T. F., Van Der Zweerde, T., Van Straten, A., Van Someren, E. J. W. W., Borsboom, D., & Lancee, J. (2019). Introducing Network Intervention Analysis to Investigate Sequential, Symptom-Specific Treatment Effects: A Demonstration in Co-Occurring Insomnia and Depression. *Psychotherapy and Psychosomatics*, *88*(1), 55–57. https://doi.org/10.1159/000495045

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305–314.

Borgatti, S. P. (2005). *Centrality and network flow*. *27*(April 2002), 55–71. https://doi.org/10.1016/j.socnet.2004.11.008

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

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

Borsboom, D., Cramer, A. O. J., & Kalis, A. (2019). Brain disorders ? Not really: Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, *42*, 1–63. https://doi.org/10.1017/S0140525X17002266

Borsboom, D., Kievit, R., Cervone, D., & Hood, S. (2009). The Two Disciplines of Scientific Psychology, or: The Disunity of Psychology as a Working Hypothesis Denny. In J. Valsiner (Ed.), *Dynamic Process Methodology in the Social and Developmental Sciences*. Springer Science+Business Media LLC. https://doi.org/10.1007/978-0-387-95922-1

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., van der Maas, H. L. J., Dalege, J., Kievit, R., & Haig, B. (2020). Theory Construction Methodology: A practical framework for theory formation in psychology. *PsyArXiv, Preprint*. https://psyarxiv.com/w5tp8/

Bos, F. M., Snippe, E., de Vos, S., Hartmann, J. A., Simons, C. J. P., van der Kriege, L., de Jonge, P., & Wichers, M. (2017). Can We Jump from Cross-Sectional to Dynamic Interpretations of Networks? Implications for the Network Perspective in Psychiatry. *Psychotherapy and Psychosomatics*, *83*(3). https://doi.org/10.1159/000453583

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791. https://doi.org/10.2307/2286841

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, 127–148.

Briganti, G., Kempenaers, C., Braun, S., Fried, E. I., & Linkowski, P. (2018). Network analysis of empathy items from the Interpersonal Reactivity Index in 1973 young adults. *Psychiatry Research*, *265*, 87–92. https://doi.org/10.17605/OSF.IO/JV273

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, W., Schoch, D., Wichers, M., & Snippe, E. (2019). What do centrality measures measure in psychological networks. *Journal of Abnormal Psychology*, *128*(8), 892–90.

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

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Burger, J., van der Veen, D. C., Robinaugh, D., Quax, R., Riese, H., Schoevers, R., & Epskamp, S. (2019). *Bridging the Gap Between Complexity Science and Clinical Practice by Formalizing Idiographic Theories: A Computational Model of Functional Analysis*.

Carragher, N., Teesson, M., Sunderland, M., Newton, N. C., Krueger, R. F., Conrod, P. J., Barrett, E. L., Champion, K. E., Nair, N. K., & Slade, T. (2016). The structure of adolescent psychopathology: A symptom-level analysis. *Psychological Medicine*, *46*(5), 981–994. https://doi.org/10.1017/S0033291715002470

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2013). 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

Caspi, A., & Moffitt, T. E. (2018). All for One and One for All: Mental Disorders in One Dimension. *American Journal of Psychiatry*, appi.ajp.2018.1. https://doi.org/10.1176/appi.ajp.2018.17121383

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. https://doi.org/10.1159/000497425

Cramer, A. O. J., van Borkulo, C. D., Giltay, E. J., van der Maas, H. L. J., Kendler, K., 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

Cramer, A. O. J., Van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S., Kendler, K., & Borsboom, D. (2012). Author’s response: Measurable Like Temperature or Mereological Like Flocking? On the Nature of Personality Traits. *European Journal of Personality*, *26*, 451–459. https://doi.org/10.1002/per

Crede, M., Harms, P., Crede, M., & Harms, P. (2019). *Questionable research practices when using confirmatory factor analysis*. https://doi.org/10.1108/JMP-06-2018-0272

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4).

Cummins, R. (2000). “How does it work?” vs. “What are the laws?” Two conceptions of psychological explanation. *Explanation and Cognition*, 117–144. https://doi.org/10.1093/acprof:osobl/9780199548033.003.0016

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2015). Toward a formalized account of attitudes: The causal attitude network (CAN) model. *Psychological Review*, *123*(1), 2–22. https://doi.org/10.1037/a0039802

Dickens, W. T. (2007). What is g ? *The Brookings Institution*.

Dickens, W. T., & Flynn, J. R. (2001). Heritability Estimates Versus Large Environmental Effects : The IQ Paradox Resolved. *Psychological Review*, *108*(2), 346–369.

Edwards, J. R., & Bagozzi, R. P. (2000). On the Nature and Direction of Relationships Between Constructs and Measures. *Psychological Methods*, *5*(2). https://doi.org/10.1037//1082-989X.5.2

Elliott, H., Jones, P. J., & Schmidt, U. (2019). Central Symptoms Predict Posttreatment Outcomes and Clinical Impairment in Anorexia Nervosa: A Network Analysis. *Clinical Psychological Science*. https://doi.org/10.1177/2167702619865958

Engel, G. (1977). The need for a new medical model: a challenge for biomedicine. *Science*, *196*(4286), 129–136. https://doi.org/10.1126/science.847460

Epskamp, S. (2017). *Network Psychometrics*. University of Amsterdam.

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating Psychological Networks and their Accuracy: A Tutorial Paper. *Behavior Research Methods*, *50*(1), 195–212. https://doi.org/10.3758/s13428-017-0862-1

Epskamp, S., & Fried, E. I. (2018). A Tutorial on Regularized Partial Correlation Networks. *Psychological Methods*, *23*(4), 617–634. https://doi.org/10.1037/met0000167

Epskamp, S., Fried, E. I., Borkulo, C., Robinaugh, D. J., Marsman, M., Dalege, J., Rhemtulla, M., & Cramer, A. O. J. (2018). Investigating the Utility of Fixed-Margin Sampling in Network Psychometrics. *Multivariate Behavioral Research*, 1–15. https://doi.org/10.1080/00273171.2018.1489771

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2016). Generalized Network Psychometrics: Combining Network and Latent Variable Models. *Psychometrika*.

http://arxiv.org/abs/1605.09288

Epstein, J. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11*(4), 6. https://doi.org/10.1080/01969720490426803

Feynman, R. (1986). *Surely You're Joking, Mr. Feynman!* Bentam Books.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, 201711978. https://doi.org/10.1073/pnas.1711978115

Forstmann, B. U., Wagenmakers, E. J., Eichele, T., Brown, S., & Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: Opposites attract? *Trends in Cognitive Sciences*, *15*(6), 272–279. https://doi.org/10.1016/j.tics.2011.04.002

Fraley, R. C., Heffernan, M. E., Vicary, A. M., & Brumbaugh, C. C. (2011). The Experiences in Close Relationships-Relationship Structures Questionnaire: A Method for Assessing Attachment Orientations Across Relationships. *Psychological Assessment*, *23*(3), 615–625. https://doi.org/10.1037/a0022898

Fried, E. I. (2015). Problematic assumptions have slowed down depression research: why symptoms, not syndromes are the way forward. *Frontiers in Psychology*, *6*(306), 1–11. https://doi.org/10.3389/fpsyg.2015.00309

Fried, E. I., Bockting, C. L. H., Arjadi, R., Borsboom, D., Amshoff, M., Cramer, A. O. J., Epskamp, S., Tuerlinckx, F., Carr, D., & Stroebe, M. (2015). From Loss to Loneliness : The Relationship Between Bereavement and Depressive Symptoms. *Journal of Abnormal Psychology*, *124*(2), 256–265. https://doi.org/10.1037/abn0000028

Fried, E. I., & Cramer, A. O. J. (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

Fried, E. I., Eidhof, M. B., Palic, S., Costantini, G., Dijk, H. M. H., Bockting, C. L. H., Engelhard, I., Armour, C., Nielsen, A. B. S., & Karstoft, K. (2018). Replicability and Generalizability of Posttraumatic Stress Disorder (PTSD) Networks: A Cross-Cultural Multisite Study of PTSD Symptoms in Four Trauma Patient Samples. *Clinical Psychological Science*, *6*(3), 335–351. https://doi.org/10.1177/2167702617745092

Fried, E. I., van Borkulo, C. D., Cramer, A. O. J., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*, *52*(1), 1–10. https://doi.org/10.1007/s00127-016-1319-z

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring Depression Over Time . . . or not? Lack of Unidimensionality and Longitudinal Measurement Invariance in Four Common Rating Scales of Depression. *Psychological Assessment*, *28*(11), 1354–1367. https://doi.org/10.1037/pas0000275

Fritz, J., Stochl, J., Fried, E. I., Goodyer, I. M., Borkulo, C. D. Van, Wilkinson, P. O., & van Harmelen, A.-L. (2019). Unravelling the Complex Nature of Resilience Factors and their Changes between Early and Later Adolescence. *BMC Medicine*, 1–16.

Gershman, S. J. (2019). Psychonomic Bulletin and Review How to never be wrong. *Psychonomic Bulletin & Review*, *26*(1), 13–28. https://doi.org/10.3758/s13423-018-1488-8

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*(2), 254–267. https://doi.org/10.1037/0033-295X.98.2.254

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: a new approach for estimating the number of dimensions in psychological research. *Plos ONE*. https://doi.org/10.1371/journal.pone.0174035

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*(3), 424–438. https://doi.org/10.2307/1912791

Greene, A. L., Eaton, N. R., Li, K., Forbes, M. K., Krueger, R. F., Markon, K. E., Waldman, I. D., Cicero, D. C., Conway, C. C., Docherty, A. R., Fried, E. I., Ivanova, M. Y., Jonas, K. G., Latzman, R. D., Patrick, C. J., Reininghaus, U., Tackett, J. L., Wright, A. G. C., & Kotov, R. (2019). Are Fit Indices Used to Test Psychopathology Structure Biased? A Simulation Study. *Journal of Abnormal Psychology*. https://doi.org/10.1037/abn0000434

Guest, O., & Martin, A. E. (2020). How computational modeling can force theory building in psychological science. *Preprint, 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

Hallquist, M. N., Wright, A. G. C., & Molenaar, P. C. M. (2019). Problems with Centrality Measures in Psychopathology Symptom Networks: Why Network Psychometrics Cannot Escape Psychometric Theory. *Multivariate Behavioral Research*, *0*(0), 1–25. https://doi.org/10.1080/00273171.2019.1640103

Hammen, C. (2005). Stress and depression. *Annual Review of Clinical Psychology*, *1*, 293–319. https://doi.org/10.1146/annurev.clinpsy.1.102803.143938

Haslbeck, J., Ryan, O., Robinaugh, D., & Waldorp, L. (2019). *Modeling Psychopathology: From Data Models to Formal Theories*. 1–29. https://psyarxiv.com/jgm7f/

Henry, T. R., Robinaugh, D. J., & Fried, E. (2020). *On the control of psychological networks*. 1–30.

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*, *55*(1), 1–34. https://doi.org/10.1093/bjps/55.1.1

Hooley, J. M. (2007). Expressed Emotion and Relapse of Psychopathology. *Annual Review of Clinical Psychology*, *3*(1), 329–352. https://doi.org/10.1146/annurev.clinpsy.2.022305.095236

Hopwood, C. J. (2018). Interpersonal Dynamics in Personality and Personality Disorders. *European Journal of Personality*, *524*(June), 499–524. https://doi.org/10.1002/per.2155

Hoyle, R. H. (2012). *Handbook of structural equation modeling*. Guilford Press.

Jensen, A. R. (1987). Psychometric g as a focus of concerted research effort. *Intelligence*, *11*(3), 193–198. https://doi.org/10.1016/0160-2896(87)90005-5

Jensen, A. R. (2002). The general factor of intelligence: How general is it? In R. J. Sternberg & E. L. Grigorenko (Eds.), *Psychometric g: Definition and sub- stantiation* (pp. 39–53). Erlbaum.

Jonas, E., & Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLoS Computational Biology*, *13*(1), 1–24. https://doi.org/10.1371/journal.pcbi.1005268

Kan, K., Maas, H. L. J. Van Der, & Levine, S. Z. (2019). Extending psychometric network analysis : Empirical evidence against g in favor of mutualism ? *Intelligence*, *73*(January 2018), 52–62. https://doi.org/10.1016/j.intell.2018.12.004

Kellen, D. (2019). A Model Hierarchy for Psychological Science. *Computational Brain & Behavior*. https://doi.org/10.1007/s42113-019-00037-y

Kendler, K., Kessler, R. C., Neale, M. C., Heath, A., & Eaves, L. J. (1993). The Prediction of Major Depression in Women: Toward an Integrated Etiologic Model. *American Journal of*

Psychiatry, *150*(8), 1139–1148.

Kendler, K., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders? *Psychological Medicine*, *41*(06), 1143–1150. https://doi.org/10.1017/S0033291710001844

Kievit, R. A., Hofman, A. D., & Nation, K. (2019). Mutualistic Coupling Between Vocabulary and Reasoning in Young Children: A Replication and Extension of the Study by Kievit et al. (2017). *Psychological Science*, *30*(8), 1245–1252. https://doi.org/10.1177/0956797619841265

Kievit, R. A., Lindenberger, U., Goodyer, I. M., Jones, P. B., Fonagy, P., Bullmore, E. T., & Dolan, R. J. (2017). Mutualistic Coupling Between Vocabulary and Reasoning Supports Cognitive Development During Late Adolescence and Early Adulthood. *Psychological Science*, *28*(10), 1419–1431. https://doi.org/10.1177/0956797617710785

Klein, D. N., Kotov, R., & Bufferd, S. J. (2011). Personality and Depression: Explanatory Models and Review of the Evidence. *Annual Review of Clinical Psychology*, *7*(1), 269–295. https://doi.org/10.1146/annurev-clinpsy-032210-104540

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., IJzerman, H., Nilsonne, G., Vanpaemel, W., & Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*(1), 1–15. https://doi.org/10.1525/collabra.158

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kovacs, K., & Conway, A. R. A. (2019). What Is IQ? Life Beyond “General Intelligence.” *Current Directions in Psychological Science*, *28*(2), 189–194. https://doi.org/10.1177/0963721419827275

Kruis, J., & Maris, G. (2016). Three representations of the Ising model. *Scientific Reports*, *6*(October), 34175. https://doi.org/10.1038/srep34175

Lakens, D., & Debruine, L. M. (2020). *Improving Transparency, Falsifiability, and Rigour by Making Hypothesis Tests Machine Readable*. 1–13. https://psyarxiv.com/5xcda/

Lange, J., Dalege, J., Borsboom, D., van Kleef, G. A., & Fischer, A. (2019). Psychometrics models of emotions. *Perspectives on Psychological Science*. https://doi.org/10.1017/CBO9781107415324.004

Lilienfeld, S. O., & Pinto, M. D. (2015). Risky Tests of Etiological Models in Psychopathology Research: The Need for Meta-Methodology. *Psychological Inquiry*, *26*(3), 253–258.

https://doi.org/10.1080/1047840X.2015.1039920

Lykken, D. T. (1991). What's Wrong with Psychology Anyway? In D. Cicchetti & W. Grove (Eds.), *Thinking clearly about psychology* (Vol. 1, pp. 3–39).

Markus, K. (2004). *Varieties of Causal Modeling: How Optimal Research Design Varies by Explanatory Strategy* (pp. 175–196). https://doi.org/10.1007/978-1-4020-1958-6_10

Markus, K. (2008). Hypothesis formulation, model interpretation, and model equivalence: Implications of a mereological causal interpretation of structural equation models. In *Multivariate Behavioral Research* (Vol. 43, Issue 2). https://doi.org/10.1080/00273170802034802

Marmot, M. (2015). *The Health Gap: The Challenge of an Unequal World*.

McCrae, R. R., & Costa, P. T. (1995). Trait explanations in personality psychology. *European Journal of Personality*, *9*(4), 231–252. https://doi.org/10.1002/per.2410090402

McWilliams, L. A., & Fried, E. I. (2019). Reconceptualizing adult attachment relationships: A network perspective. *Personal Relationships*, *26*, 21–41. https://doi.org/10.1111/pere.12263

Meehl, P. E. (1990a). 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

Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195–244.

Mikulincer, M., & Shaver, P. R. (2012). An attachment perspective on psychopathology. *World Psychiatry*, *11*(8), 11–15.

Miller, G. A. (2010). Mistreating psychology in the decades of the brain. *Perspectives on Psychological Science*, *5*(6), 716–743. https://doi.org/10.1177/1745691610388774

Mischel, W. (2004). Toward an Integrative Science of the Person. *Annual Review of Psychology*, *55*(1), 1–22. https://doi.org/10.1146/annurev.psych.55.042902.130709

Molenaar, P. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *June 2014*, 37–41. https://doi.org/10.1207/s15366359mea0204

Morton, R. B. (2009). Formal Modeling and Empirical Analysis in Political Science. In *Methoden* *der vergleichenden Politik- und Sozialwissenschaft* (pp. 27–35). https://doi.org/10.1007/978-3-531-91826-6_2

Mõttus, R., & Allerhand, M. (2017). Why do traits come together? The underlying trait and network approaches. In V. Zeigler-Hill & T. Shackelford (Eds.), *SAGE handbook of personality and individual differences: Volume 1. The science of personality and individual differences* (pp. 1–22). SAGE.

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

Noguchi, H., & Moore, J. W. (1913). A demonstration of treponema pallidum in the brain in cases of general paralysis. *Journal of Experimental Medicine*, *17*(2), 232–238. https://doi.org/10.1084/jem.17.2.232

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration Is Hard, And Worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

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

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Paykel, E. S. (2003). Life events and affective disorders. *Acta Psychiatrica Scandinavica*, *108*(418), 61–66. http://www.ncbi.nlm.nih.gov/pubmed/12956817

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Peng, P., & Kievit, R. A. (2020). The development of academic achievement and cognitive abilities: A bidirectional perspective. *Child Development PerspectivesP*.

Phaf, R. H. (2020). Publish less, read more. *Theory & Psychology*, 095935431989825. https://doi.org/10.1177/0959354319898250

Raykov, T., & Marcoulides, G. A. (2001). Can There Be Infinitely Many Models Equivalent to a Given Covariance Structure Model? *Structural Equation Modeling*, *8*(1), 142–149. https://doi.org/10.1207/S15328007SEM0801_8

Rhemtulla, M., van Bork, R., & Borsboom, D. (2018). *Worse than measurement error*.

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. J. (2020). Invisible Hands and Fine Calipers: A Call to Use Formal Theory as a Toolkit for Theory Construction. *Perspectives on Psychological Science*, 1–11.

Robinaugh, D. J., Haslbeck, J. M. B., Waldorp, L. J., Kossakowski, J. J., Fried, E. I., Millner, A. J., Mcnally, R. J., Nes, E. H. Van, Scheffer, M., Kendler, K. S., & Borsboom, D. (2019). *Advancing the Network Theory of Mental Disorders: A Computational Model of Panic Disorder*. *647209*, 1–76. https://psyarxiv.com/km37w/

Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R., & Borsboom, D. (2019). The network approach to psychopathology: a review of the literature 2008–2018 and an agenda for future research. *Psychological Medicine*. https://doi.org/10.1080/01559982.2019.1584953

Robinaugh, D. J., Millner, A. J., & McNally, R. J. (2016). Identifying highly influential nodes in the complicated grief network. *Journal of Abnormal Psychology*, *125*(6), 747–757. https://doi.org/10.1037/abn0000181

Rodebaugh, T. L., Tonge, N. A., Piccirillo, M. L., Fried, E. I., Horenstein, A., Morrison, A. S., Goldin, P., Gross, J. J., Lim, M. H., Fernandez, K., Blanco, C., Schneier, F. R., Bogdan, R., Thompson, R. J., & Heimberg, R. H. (2018). Does Centrality in a Cross-Sectional Network Suggest Intervention Targets for Social Anxiety Disorder? *Journal of Consulting and Clinical Psychology.*, *86*, 831–844. https://doi.org/10.1037/ccp0000336

Samuel, D. B. (2019). Incoming Editorial. *Assessment*, *26*(2), 147–150. https://doi.org/10.1177/1073191118822760

Savi, A. O., Marsman, M., van der Maas, H. L. J., & Maris, G. K. J. (2019). The Wiring of Intelligence. *Perspectives on Psychological Science*, *14*(6), 1034–1061. https://doi.org/10.1177/1745691619866447

Schmittmann, V. D., Cramer, A. O. J., 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. http://linkinghub.elsevier.com/retrieve/pii/S0732118X1100016X

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed

Addition to All Empirical Papers. *Perspectives on Psychological Science*, 174569161770863. https://doi.org/10.1177/1745691617708630

Smaldino, P. (2017). Models are stupid, and we need more of them. *Computational Social Psychology*, 311–331. https://doi.org/10.4324/9781315173726

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

Spanos, A., & Mayo, D. G. (2015). Error statistical modeling and inference: Where methodology meets ontology. *Synthese*, *192*(11), 3533–3555. https://doi.org/10.1007/s11229-015-0744-y

Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292.

Stroebe, W., & Strack, F. (2014). The Alleged Crisis and the Illusion of Exact Replication. *Perspectives on Psychological Science*, *9*(1), 59–71. https://doi.org/10.1177/1745691613514450

Suppes, P. (1962). Models of Data. *Logic, Methodology, and Philosophy of Science: Proceedings of the 1960 International Congress*, *57*, 252–261.

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., Van Rooij, I., Van Zandt, T., & Donkin, C. (2019). Is Preregistration Worthwhile? *Trends in Cognitive Sciences*, 2–3. https://doi.org/10.31234/osf.io/x36pz

Townsend, J. T. (2008). Mathematical Psychology: Prospects For The 21st Century: A Guest Editorial. *Journal of Mathematical Psychology*, *52*(5), 269–280. https://doi.org/10.1016/j.jmp.2008.05.001

Turkheimer, E. (2016). Weak Genetic Explanation 20 Years Later: Reply to Plomin et al. (2016). *Perspectives on Psychological Science*, *11*(1), 24–28. https://doi.org/10.1177/1745691615617442

Vaidyanathan, U., Vrieze, S. I., & Iacono, W. G. (2015). The power of theory, research design, and transdisciplinary integration in moving psychopathology forward. *Psychological Inquiry*, *26*(3), 209–230. https://doi.org/10.1080/1047840X.2015.1015367

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences of the United*

States of America, *113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113

van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & van der Maas, H. L. J. (2017). What is P? Some risks of general factor modeling. *Theory & Psychology*, *27*(6), 759–773. https://doi.org/10.1177/0959354317737185

van Bork, R., Rhemtulla, M., Waldorp, L. J., Kruis, J., Rezvanifar, S., & Borsboom, D. (2019). Latent Variable Models and Networks: Statistical Equivalence and Testability. *Multivariate Behavioral Research*, 1–24. https://doi.org/10.1080/00273171.2019.1672515

Van Bork, R., Wijsen, L. D., & Rhemtulla, M. (2017). Toward a Causal Interpretation of the Common Factor Model. *Disputatio*, *IX*(47). https://doi.org/10.1515/disp-2017-0019

van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., & Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific Reports*, *4*(5918), 1–10. https://doi.org/10.1038/srep05918

van de Leemput, I. A., Wichers, M., Cramer, A. O. J., Borsboom, D., Tuerlinckx, F., Kuppens, P., van Nes, E. H., Viechtbauer, W., Giltay, E. J., Aggen, S. H., Derom, C., Jacobs, N., Kendler, K., van der Maas, H. L. J., Neale, M. C., Peeters, F., Thiery, E., Zachar, P., & Scheffer, M. (2014). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(1), 87–92. https://doi.org/10.1073/pnas.1312114110

van der Maas, H. L. J., Dolan, C. V, Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842

Van Rooij, I., & Baggio, G. (2020). Theory before the test: how to build high-verisimilitude explanatory theories in psychological sciences. *PsyArXiv, Preprint*.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications. In *Structural analysis in the social sciences* (Vol. 8). https://doi.org/10.1093/infdis/jir193

Watts, A. L., Poore, H. E., & Waldman, I. D. (2019). Riskier Tests of the Validity of the Bifactor Model of Psychopathology. *Clinical Psychological Science*. https://doi.org/10.1177/2167702616673363

Wichers, M., Groot, P. C., Psychosystems, ESM Group, & EWS Group. (2016). Critical Slowing Down as a Personalized Early Warning Signal for Depression. *Psychotherapy and Psychosomatics*, *85*, 114–116. https://doi.org/10.1159/000441458

Widaman, K. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, *28*(3), 263–311. http://www.tandfonline.com/doi/abs/10.1207/s15327906mbr2803_1

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

Yarkoni, T. (2019). The generalizability crisis. *PsyArXiv*, 1–26.