Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature

Eiko I. Fried


To link to this article: https://doi.org/10.1080/1047840X.2020.1853461

© 2020 The Author(s). Published with license by Taylor & Francis Group, LLC.
Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature

Eiko I. Fried

CONTACT
Eiko I. Fried eiko-fried.com; eikofried@gmail.com Department of Clinical Psychology, Leiden University, Leiden, Netherlands

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.

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 (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 overwhelmed 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, van der Maas, Dalege, Kievit, & Haig, 2020; Cummins, 2000; Guest & Martin, 2020; Haslbeck, Ryan, Robinaugh, & Waldorp, 2019; Meehl, 1990b; Muthukrishna & Henrich, 2019; Phaf, 2020; Robinaugh, Haslbeck, et al., 2019; Smaldino, 2019; Szollosi et al., 2019; Vaidyanathan, Vrieze, & Iacono, 2015; Van Rooij & Baggio, 2020).

What are theories, and how do they relate to data, statistical models, and phenomena? 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 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.

**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). Even the most simple verbal theory (e.g., “x relates to y”) can 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, Haslbeck, Ryan, Fried, and Waldorp (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, Mende-Siedlecki, Brady, & Reinero, 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, 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, Wagenmakers, Eichele, Brown, & Serences, 2011; Palminteri, Wyart, & Koechlin, 2017; Townsend, 2008). The psychological theory glass is, I conclude, at the very best half empty.

**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.

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 for which 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 toward 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.

---

2Note 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).
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. Psychometric networks have been applied to data of numerous psychological constructs in the last few years, including mental disorders (Contreras, Nieto, Valiente, Espinosa, & Vazquez, 2019; Fried et al., 2017; Robinaugh, Hoekstra, Toner, & Borsboom, 2019), personality (Beck & Jackson, 2019; Möttus & Allerhand, 2017), attitudes (Dalege et al., 2016), cognitive abilities (Kan, Maas, Van Der, & Levine, 2019), empathy (Briganti, Kempenaers, Braun, Fried, & Linkowski, 2018), emotions (Lange, Dalege, Borsboom, van Kleef, & Fischer, 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 (Fried & Robinaugh, 2020; Borsboom, Cramer, & Kalis, 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.

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 (1904), who used correlational analysis to identify general intelligence—often called the g factor today—based on evidence that children 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, Marsman, van der Maas, & Maris, 2019; van der Maas et al., 2006), p factor (Borsboom, 2017; van Bork, Epskamp, Rhemtulla, Borsboom, & van der Maas, 2017), or personality

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.
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 \(\rightarrow\) insomnia \(\rightarrow\) fatigue).

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.

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).

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 such data 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, Jones, & Schmidt, 2020; Robinaugh, Millner, & McNally, 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, Robinaugh, & Fried, 2020).

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

**Problematiobic 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, Poore, & Waldman, 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 shown by 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 remain 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 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, Kievit, Cervone, & Hood, 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).

**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., 2020; Robinaugh et al., 2016; Rodebaugh et al., 2018). This has been taken by some to support network theory—as if 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, Wright, & Molenaar, 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 (Bos et al., 2017; Fisher, Medaglia, & Jeronisimus, 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, Amore, Respino, & Alexopoulos, 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; cf. 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.

**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.

**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 dataset⁴; 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.

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 & Harms, 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.

**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 inferences⁵. 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

---

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

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

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 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, et al., 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., 2020).

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. For example, 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).

Latent theories are weak or strong theories 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 explicative in recent work (Caspí & 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
are considered “bad” items, and usually deleted. But bad items from a common cause perspective are not necessarily bad items from other causal perspectives.

Consider the causal system in Figure 3(A), 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, as shown in Figure 3(B), 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 3(A). It is standard practice to remove x1 in such cases.

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, Heffernan, Vicary, & Brumbaugh, 2011). I showcase an example of a 2-factor model estimated on data for this scale in Figure 3(C), 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 3(D) showcases the network model estimated on this data set, with x10 in the center of the graph (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 regard 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. 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).

### 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 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 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. 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). 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

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.

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).

---

9The 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).

8I 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 explanatory rather than explanatory, one of which is that they can facilitate predictions.
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

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., 2014; 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 mathe-matistry, 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 "over reliance 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). They 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 models that can be compared to observed data to see if the models behave 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.

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, Sauer-Zavala, Carl, Bullis, & Ellard, 2014). 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 & Groot, 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.

**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, Zachar, & Craver, 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 (Fried & Robinaugh, 2020). 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; Klein, Kotov, & Bufferd, 2011; Mischel, 2004; Möttus & 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, Kessler, Neale, Heath, & Eaves, 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, Rhemtulla, & Borsboom, 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 $\rightarrow$ insomnia $\rightarrow$ 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; Kievit, Hofman, Nation, et al., 2019; Peng & Kievit, 2020).

**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, Shoda, & Lindsay, 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 would require stronger specifications of the core findings (what variables matter), auxiliary hypotheses (what variables can we vary that do 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, which appear commonplace in other disciplines. Sociologists, for example, use so-called scope statements (Walker & Cohen, 1985) to make the scope and scope restrictions of their theories explicit.

Putting the Theory Back into Psychology

“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 data to expect given the theory, which makes it difficult to know whether data corroborate it 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 toward unfalsifiability that is discussed in detail in Gershman’s piece “How to never be wrong” (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-oriented 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 & Debruijne, 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)

11This is similar to recent work calling for a “constraints on generality” section that transparently describes problems with external validity.
did not come up once in my curriculum. In fact, I had never heard about the term “formal model” before my post-doctoral 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; Gershan, 2019; Guest & Martin, 2020; Haslbeck et al., 2019; Kellen, 2019; Lakens & DeBruyne, 2020; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Robinaugh, Haslbeck, et al., 2019; Savi et al., 2019; Smallino, 2019; Szollosi et al., 2019; Van Rooij & Baggio, 2020), and several highly visible applied journals have opened calls for special issues on theory.

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

Acknowledgements

I 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 Smallino, 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.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Eiko I. Fried

http://orcid.org/0000-0001-7469-594X

References


