Folk Classification and Factor Rotations

Haeffel, Gerald J.; Jeronimus, Bertus F.; Kaiser, Bonnie N.; Weaver, Lesley Jo; Soyster, Peter D.; Fisher, Aaron J.; Vargas, Ivan; Goodson, Jason T.; Lu, Wei

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2022

[Link to publication in University of Groningen/UMCG research database](#)

Theoretical/Methodological/Review Article

Folk Classification and Factor Rotations: Whales, Sharks, and the Problems With the Hierarchical Taxonomy of Psychopathology (HiTOP)

Gerald J. Haeffel[1] [ID]**, Bertus F. Jeronimus[2], Bonnie N. Kaiser[3],**
Lesley Jo Weaver[4] [ID]**, Peter D. Soyster[5], Aaron J. Fisher[5]** [ID]**,**
Ivan Vargas[6] [ID]**, Jason T. Goodson[7], and Wei Lu[8]**

[1]Department of Psychology, University of Notre Dame; [2]Department of Psychology, University of Groningen; [3]Department of Anthropology and Global Health Program, University of California, San Diego; [4]Department of Global Studies, University of Oregon; [5]Department of Psychology, University of California, Berkeley; [6]Department of Psychology, University of Arkansas; [7]PTSD Clinical Team, VA Salt Lake City Health Care Systems, Salt Lake City, Utah; and [8]Carver School of Medicine, University of Iowa Hospitals and Clinics

Abstract

The Hierarchical Taxonomy of Psychopathology (HiTOP) uses factor analysis to group self-reported symptoms of mental illness (i.e., like goes with like). It is hailed as a significant improvement over other diagnostic taxonomies. However, the purported advantages and fundamental assumptions of HiTOP have received little, if any, scientific scrutiny. We critically evaluated five fundamental claims about HiTOP. We conclude that HiTOP does not demonstrate a high degree of verisimilitude and has the potential to hinder progress on understanding the etiology of psychopathology. It does not lend itself to theory building or taxonomic evolution, and it cannot account for multifinality, equifinality, or developmental and etiological processes. In its current form, HiTOP is not ready to use in clinical settings and may result in algorithmic bias against underrepresented groups. We recommend a bifurcation strategy moving forward in which the *Diagnostic and Statistical Manual of Mental Disorders* is used in clinical settings while researchers focus on developing a falsifiable theory-based classification system.

Structural approaches to the classification of psychopathology use factor analysis to cluster symptoms of mental illness into dimensional groupings. This quantitative approach is currently exemplified by the Hierarchical Taxonomy of Psychopathology (HiTOP; Kotov et al., 2017). There has been a steady stream of articles from the HiTOP consortium (e.g., Conway et al., 2019; DeYoung et al., 2020; Kotov et al., 2018, 2020; Krueger et al., 2018; Latzman et al., 2020; Ruggero et al., 2019; Widiger et al., 2019) touting the benefits of its system. They claim it can "carve nature at its joints" (Conway et al., 2019, p. 429), resolve problems of comorbidity and heterogeneity (Ruggero et al., 2019, p. 1071), revolutionize clinical practice (Hopwood et al., 2019, p. 15), and advance

psychiatric genetics and neuroscience research (Latzman et al., 2020; Waszczuk et al., 2020).

These extraordinary claims have received little, if any, scientific scrutiny. A critical evaluation of HiTOP and its purported advantages is needed. The purpose of this article is to fill this gap in the literature. First, we critically evaluated five fundamental claims about HiTOP. Second, we compared HiTOP with alternative taxonomies to evaluate the degree to which they lend themselves to taxonomic evolution (from description

Corresponding Author:
Gerald J. Haeffel, Department of Psychology, University of Notre Dame
E-mail: ghaeffel@nd.edu

to theory) and scientific progress (e.g., falsification). Finally, we made recommendations for future research.

Claim 1. Symptom Correlations "Carve Nature at Its Joints"

> Humans are prone to a "folk understanding bias"—the sensation that simplistic explanations lead us to believe we truly understand more complex phenomena.
>
> —Jolly and Chang (2019, p. 436)

Ostensibly, the HiTOP approach follows the same logic as the Linnaean system in biology, in which every organism is classified over seven hierarchical taxa on the basis of shared features (kingdom, phylum, class, order, family, genus, species). However, there is a critical difference between the HiTOP and Linnaean system. HiTOP dimensions are derived in a theoretical vacuum in which all characteristics (predominantly self-reported symptoms) are considered equally important. For example, the symptom of "avoidance" is weighted the same as "sleep difficulties" and "hearing voices." No symptom is considered more essential than any other symptom in this system. In contrast, the Linnaean system uses a theoretical perspective in which some characteristics are more important and, when present, take precedence over all other shared similarities because of their ontogenetic precedence.

In the Linnaean system, classification decisions are not based on total levels of "likeness" (i.e., their covariation), as in HiTOP, but rather on a subgroup of highly meaningful features as determined by evolutionary theory (i.e., phylogeny; e.g., Nickels & Nelson, 2005). To this end, the Linnaean system distinguishes between homology and analogy (Petto & Mead, 2009). Homologous structures are those that descended from a common evolutionary ancestor. For example, the forelegs of horses and dogs are homologous structures because they evolved from a common ancestral tetrapod. Thus, horses and dogs are considered more "alike" than animals that do not share this common ancestor. In contrast, analogous features are those that have a similar structure and function (because of convergent evolution) but did not evolve from a common ancestor. For example, birds, bats, moths, and sea snails (pteropoda) have wings to fly but do not share a common ancestor that evolved wings. And because this shared feature (wings) is not homologous, they are not grouped together (e.g., birds are classed as Aves, bats as Mammalia, moths as Insecta, and sea snails as Gastropoda). Likewise, echolocation evolved independently in birds (e.g., swiftlets), noctuid moths, bats, cetaceans (e.g.,

dolphins), shrews, tenrec, and humans, which are each grouped in different phyla and clades and use this skill in radically different environments (e.g., seas, skies, caves, and cities). Differentiating homologous features from analogous features is critical to the Linnaean system because it is the basis for understanding the evolution and the origin of species (Dawkins & Wong, 2016).

In contrast to the Linnaean system and the newer genetically informed cladistics systems, HiTOP resembles a folk classification system (Nickels & Nelson, 2005; Petto & Mead, 2009). HiTOP puts like with like without considering etiological or underlying developmental processes. This is a problem because "like things" may be grouped together inaccurately on the basis of superficial characteristics (analogous features), and "unlike things" might be classified separately despite sharing a common etiology (homologous features). To illustrate this point, consider what biological classification might look like if it were created using the same strategy as HiTOP (see Fig. 1)—that is, classifying animals on the basis of shared features regardless of evolutionary ancestry. This process would likely lead to an overarching factor of "animal" (the A-factor), which might then break down into a bifactor model of "land" and "water" animals. An examination of the subgroups of animals organized within these two levels starts to reveal the problems with HiTOP. For example, whales and sharks would be incorrectly classified together given the high correlations among their shared features (e.g., ocean dwellers, fins for locomotion, fish and crustacean eaters, similar life spans, can adapt to multiple aquatic habitats, both largest of their family). This is because in HiTOP, features such as being warm-blooded and having hair do not carry special importance.

Moreover, bats would likely be incorrectly classified with other flying animals such as birds, moths, and butterflies. Red pandas would likely be classified with raccoons despite phylogenetic analysis confirming that they belong in their own evolutionary family. Elephants would be grouped with other large, thick-skinned herbivores such as hippos and rhinos even though their closest evolutionary relatives are hyraxes (which look like prairie dogs) and manatees. And the Tasmanian tiger would be grouped with canids (dogs, wolves, foxes) despite being a marsupial. These are just a few of a myriad of examples that illustrate a fundamental flaw in the structural approach to classification—theoretical and etiological factors are ignored. Using an empirically based strategy to sort (i.e., correlate) a large set of features does not necessarily lead to "more accurate" (Kotov et al., 2017, p. 469) or valid diagnoses even when the model has an excellent statistical fit.

This calls into question HiTOP's most fundamental assumption: that individuals who report similar patterns
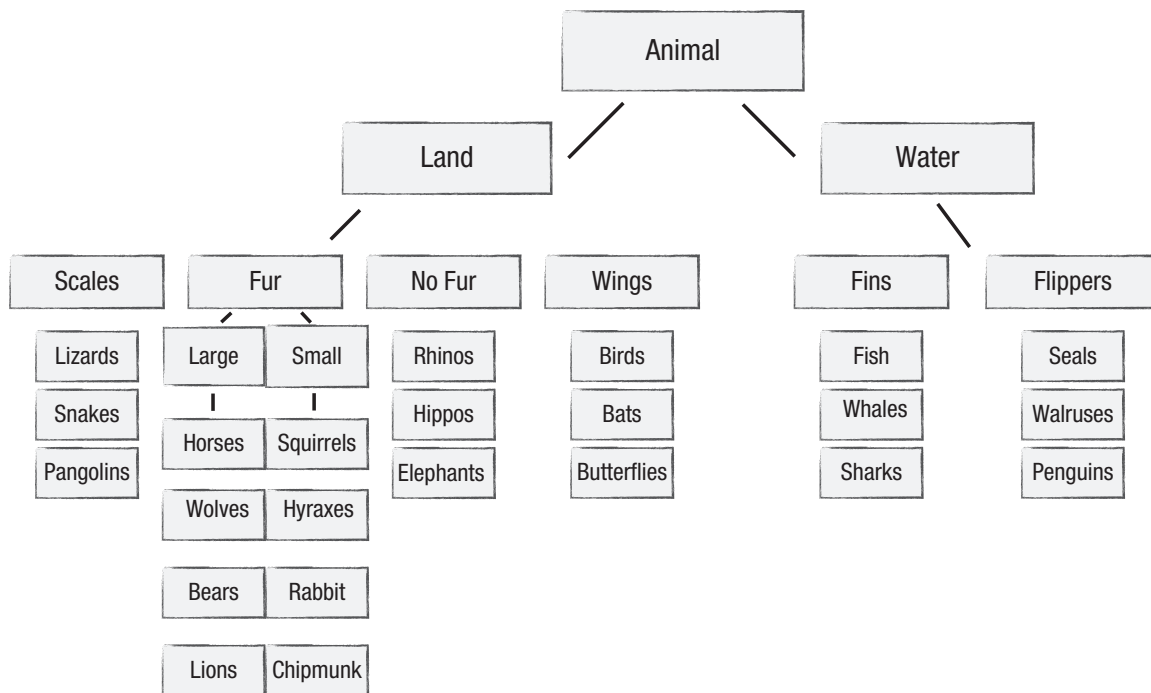
Fig. 1. What biological classification might look like if it were created using the same strategy as HiTOP.

of symptoms have the same form of psychopathology (which can be targeted by the same treatment because of shared etiology; Ruggero et al., 2019). As our animal classification example illustrates, HiTOP cannot account for equifinality (Cicchetti & Rogosch, 1996). In the case of equifinality, two individuals can reach the same phenotypic end state through different etiological processes (similarly to how birds and bats both developed wings). In HiTOP, these individuals would be considered "the same" despite the fact that they may have different disorders and need different treatments. There are numerous examples of equifinality in nature. For example, fatigue, body aches, pain, and headache are all symptoms common to influenza, rhinovirus, mononucleosis, and Lyme disease. Yet despite sharing the same phenotype, all of these medical problems have different etiologies (i.e., they are caused by distinct viruses) and are all treated differently. Likewise, chest pain and shortness of breath are common to acute coronary syndrome, pulmonary embolism, pneumonia, rib fracture, anxiety, and heart failure (McConaghy, 2020; Schwartzstein, 2020). Again, despite sharing the same symptom phenotype, these physical ailments are distinct and are also treated differently. Likewise, it is untenable to assume that people with depression and people with posttraumatic stress disorder (PTSD) should be grouped together (because of shared "distress" symptoms) without understanding their etiology. Meehl (1989b) noted that "a one-to-one correlation over individuals between

two things does not mean that the two things are actually identical . . . all animals with a heart have a kidney, but that does not show that the words heart and kidney designate the same concept!" (p. 938).

In summary, a taxonomy built on symptom covariation is unlikely to capture the complexity of nature. There is little evidence that HiTOP (a) is "modeled in nature" (Krueger et al., 2018, p. 286), (b) will "improve our ability to carve nature at its joints" (Conway et al., 2019, p. 429), or (c) can "explain the etiology of psychological problems" (Conway et al., 2019, p. 432).

Claim 2. HiTOP Will Solve the Problems of Comorbidity and Heterogeneity

> The hypotheses the statistician tests exist in a world of black and white, where the alternatives are clear, simple, and few in number, whereas the scientist works in a vast gray area in which the alternative hypotheses are often confusing, complex, and limited in number only by the scientist's ingenuity.
>
> —Bolles (1962, p. 639)

Comorbidity

The HiTOP approach "promises to resolve problems of comorbidity, heterogeneity, and arbitrary diagnostic thresholds" (Waszczuk et al., 2020, p. 12). In the case

of comorbidity, it is possible that HiTOP is waging a battle on a false front. Comorbidity is a problem when the co-occurring disorders represent the same condition and can be treated the same way (i.e., they are redundant). Without understanding the etiology of the disorders one diagnoses, it is impossible to know whether current comorbidity rates are artificially high.

Nature is complex, and etiologically distinct conditions can frequently co-occur. For example, 60% of Americans over the age of 65 years have two or more types of chronic medical conditions (43% have three or more; 24% have four or more; Centers for Disease Control and Prevention [CDC], 2019). Research shows that cardiovascular disease is highly comorbid with diabetes, chronic kidney disease, and depression (CDC, 2019). However, we suspect that most medical doctors and scientists would not dismiss the distinctiveness of these conditions and call for the eradication of this kind of comorbidity. In fact, level of comorbidity can be an important predictor of clinical outcomes such as adverse drug events, poor functioning, unnecessary hospitalizations, and even death (De Vries et al., 2019; Wolff et al., 2002). This kind of (valid) comorbidity is not inherently bad, nor does it invalidate a classification system.

That said, let us assume that comorbidity in the currently used diagnostic system (*Diagnostic and Statistical Manual of Mental Disorders* [*DSM*]) *does* reflect redundancies and inaccuracies. Does HiTOP solve the problem as promised by Conway and colleagues (2019)? The HiTOP solution is to lump diagnoses together and then give them a new label. This approach eliminates the need to provide more than one diagnosis for a cluster of symptoms, but this shell game does not create new knowledge or new theoretical explanations or identify new etiological pathways. Rather, it gives new labels to the same collection of symptoms. This creates larger, more heterogeneous groupings, which may not be clinically useful and can hinder our understanding of the etiology of mental illness. As noted by Smith and colleagues (2009),

> when it occurs that a previously recognized psychological construct is subdivided into more elemental components that have different etiologies, or different external correlates, or that require different interventions, it no longer makes sense to treat the original entity as a coherent, homogeneous construct. (p. 273)

Moreover, an implicit assumption of HiTOP is that people will fit neatly into one spectrum and a line of subfactors. However, research indicates that this is unlikely. Instead, people will "score high" on multiple subfactors and spectra (e.g., the co-occurrence of

internalizing and externalizing problems is substantial in both clinical and epidemiological studies; Pesenti-Gritti et al., 2008). Thus, people categorized using HiTOP are still going to carry an abundance of labels because a person might report internalizing, externalizing, substance use, distress, and antisocial behavior symptoms.

One might respond to this criticism by asking the following question: If HiTOP's hierarchical approach is not valid, then why do some treatments appear to cut across current diagnostic categories? This would seem to suggest that there are common etiologies cutting across the *DSM* categories that are being captured by HiTOP's "transdiagnostic" hierarchy. Unfortunately, the cause of a disorder does not always match up with the treatment of a disorder and vice versa (e.g., cigarette smoking is a causal risk factor for lung cancer, but stopping smoking is not an effective treatment for lung cancer). Exercise, good sleep, healthy diet, and cognitive expectations (placebo) are effective in mitigating and preventing nearly every human physical and mental ailment. The beneficial effects cut across hundreds of human problems (heart disease, depression, obesity, cancers, anxiety, etc.), but it does not mean that the problems they alleviate should be considered the same. Acetaminophen, naproxen sodium, and ibuprofen all are effective in treating headaches, pain, and fever associated with a variety of illnesses. Yet there is not a push in medicine to label these *transdiagnostic treatments*. Their efficacy also would not support the creation of a "headache" diagnostic category in a medical taxonomy. The point is that just because a treatment works for multiple problems, it does not mean those problems belong together in a taxonomy. Likewise, evidence of transdiagnostic treatments does not validate HiTOP or invalidate existing taxonomies.

Related to the idea of transdiagnostic treatments are transdiagnostic risk factors. Research shows that many risk factors are nonspecific. It is unclear what conclusions can be made about this kind of nonspecificity. It is not necessarily appropriate to conclude that the existence of common risk factors means that the disorders they influence should be considered the same. Again, research shows that smoking, poor nutrition, and low levels of exercise are the three most important predictors of common health problems in Americans, including heart disease and a variety of cancers (Khera et al, 2016). The lack of specificity for these risk factors does not invalidate the diagnoses that arise from them (or justify their lumping together). This is another example of the complexity of nature and a reminder that common contributors may ultimately lead to a variety of different outcomes. Trying to eliminate comorbidity because it is "messy" likely leads to an even more invalid and artificial taxonomy.

Heterogeneity

Another purpose of HiTOP is to resolve the problem of within-disorders heterogeneity (Kotov et al., 2018). The problem of heterogeneity is typically illustrated by showing that two people with the same *DSM* diagnosis may not share any of the same symptoms. For example, Conway and colleagues (2019) noted that there are 600,000 possible PTSD symptom combinations, which indicates that the *DSM* and its polythetic "menu" approach is not a valid taxonomy. First, it is important to recognize that just because it is mathematically possible to have a large number of symptom combinations, it does not mean that all those combinations are expressed in reality. For example, it may be possible to have a large number of genetic configurations (haplotypes), and yet all of those combinations are not expressed in nature. That said, even if all 600,000 combinations did exist in nature, it does not invalidate the diagnosis. It is possible for individuals with the same underlying problem to express completely different symptom profiles, as demonstrated by the principle of multifinality.

In the case of multifinality, the same causal agent (e.g., obesity) can lead to distinct outcomes or symptom profiles in people (e.g., diabetes or obstructive sleep apnea). Thus, it is possible for two people to express completely different symptom profiles yet share a common etiological pathway that can be targeted by the same treatment. There are numerous examples of this phenomenon in medicine. People with lupus often have completely different symptom presentations that include some combination of fatigue, fever, joint pain, rash, pericarditis, Raynaud phenomenon, vasculitis, blood clots, nephritis, shortness of breath, and anemia (Cojocaru et al., 2011; Wallace & Gladman, 2020). Systemic sclerosis is another disorder in which there may be no overlap in self-reported symptoms among people (symptoms can include things such as skin sclerosis, renal failure, interstitial lung disease, pulmonary hypertension, joint pain, pericardial effusion, erectile dysfunction, myopathy, and myocarditis; Adigun et al., 2002; Varga, 2020). These are just a few examples (others include COVID-19, hyperthyroidism, irritable bowel syndrome, etc.) that illustrate how people can express completely different symptom profiles without overlapping symptoms and yet suffer from the same underlying problem. HiTOP would miss these cases because the symptom profiles do not covary; it cannot deal with this kind of natural complexity (Kendler et al., 2011).

Symptom heterogeneity is a problem when the different symptoms do not share a common etiology. Strauss and Smith (2009) provided the following example to illustrate this point. According to these authors, neuroticism consists of six correlated but distinct constructs. Thus, it is possible for two people to have the exact same score on a general measure of neuroticism but for different reasons (e.g., one person may score high on hostility and low on self-consciousness, whereas another person may score low on hostility and high on self-consciousness). They argue that this kind of heterogeneity makes a total score on neuroticism imprecise, ambiguous, and an obstacle to theory testing. If we apply this example to HiTOP, we can see how its hierarchy may also hinder scientific progress. Depression appears to be a heterogeneous construct, likely reflecting multiple disorders with distinct etiologies (McGrath, 2005; Smith et al., 2009). Thus, an overall depression score is imprecise and may lead to uninterpretable findings. HiTOP compounds the problem by creating even larger groupings such as "distress," which includes not only depression but also syndromes like PTSD and generalized anxiety disorder. Distress is then combined with other heterogeneous groupings (e.g., fear, eating pathology, mania, sexual problems) under the umbrella of "internalizing." As one moves up the hierarchy, the scores become less and less useful. As noted by Littlefield and colleagues (2021), "currently, there is no clear consensus . . . regarding the utility of these common factors as a way to understand the potential structure of important constructs or to inform theoretical and clinical efforts" (p. 10).

In sum, it is premature to assume that a classification system is invalid because two people can have the same disorder without sharing the same symptoms (e.g., COVID-19 is a valid diagnosis despite highly heterogeneous symptom presentations). In fact, it may show that a classification system is scientifically progressive because it can account for multifinality. For example, after experiencing a life-threatening event, a small number of people will develop a clinically significant form of psychopathology (PTSD) that is expressed in a variety of ways. Despite the different symptom expressions, the *DSM* can identify these people as having the same problem, in part, by requiring the presence of a common contributory cause (life-threatening event).

Claim 3. HiTOP Is Empirical and Objective

> A statistical procedure is not an automatic, mechanical truth-generating machine for producing or verifying substantive causal theories. Of course we all know that, as an abstract proposition; but psychologists are tempted to forget it in practice. (I conjecture the temptation has become stronger due to modern computers, whereby an investigator may understand a statistical procedure only enough to instruct an R. A. or computer lab personnel to "factor analyze these data.")
>
> —Meehl (1992, p. 143)

The structural approach to classification is described as "quantitative," "empirical," "more accurate," and "derived strictly from data, free of political considerations" (Kotov et al., 2020, p. 165). Alternative approaches (e.g., *DSM*), in contrast, are described as the result of "authority and fiat" in which "experts gather under the auspices of official bodies and delineate classificatory rubrics through group discussions and associated political processes" (Krueger et al., 2018, p. 282). This characterization of HiTOP suggests that it is more objective and empirically valid than other classification systems; it is based on scientific facts, whereas taxonomies like the *DSM* are based on scientific opinions.

The insinuation that *DSM* committee members embrace politics over science is likely unjustified. As stated by Kendler (2018), "The procedures developed for change in DSM-5 by the American Psychiatric Association's Steering Committee are empirically rigorous and data driven" (p. 242). Likewise, the notion that HiTOP's 100-member consortium is immune to group dynamics is probably untrue. It is difficult to believe that decisions about HiTOP rely solely on the unthinking application of data.

Representation and structure

Politics aside, factor analysis does seem more objective than expert consensus. Data are entered into a statistical software package, analyses are specified, and a statistical solution appears without human interference. However, describing this approach as "empirical" and "data-driven" is somewhat misleading. Although HiTOP is derived from empirical data, its structure of symptom descriptors is not empirically supported. HiTOP uses a dimensional interpretation/simple structure procedure (Thurstone, 1947) in which stimuli are rotated to have high loadings on one dimension but low loadings on others in an effort to reduce cross-loadings and create unique factors; this is the same approach used by its predecessor, the five-factor model of personality. However, this mode of representation likely does not capture the complexity of the actual empirical structure of the data, which has yet to be actually tested (e.g., facet theory; Guttman, 1982). For example, the structure may be better represented by a radex, cylinder, circumplex, or simplex. As cautioned by Maraun (1997), "without a careful distinction being made between model, structure, representation, and mode of representation, and without the employment of appropriate methods for structural analysis, researchers are destined to confuse mere appearance with reality" (p. 646).

The dimensional interpretation/simple structure procedure leads to an infinite number of well-fitting models.[1] Choosing among these models is often based on

ease of interpretation and personal preference, not empirical veracity. And as statistical software packages have made it easier and easier to rotate solutions to simple structures, it has been "forgotten that the resulting dimensions were a post hoc MBA [meaningful but arbitrary] expediency, not a data-driven realization of a deeper scientific reality" (Turkheimer, 2017, p. 35). HiTOP is a mathematical solution constrained by an inadequate representation of the dimensional space of the symptoms of psychopathology. According to Maraun (1997), this ensures a "systematic misrepresentation of the structure" (p. 632). Supporting this claim, multiple studies show that the complexity of human personality descriptors may be better represented by a spherical three-dimensional model than the more widely endorsed five-factor model (e.g., Markey & Markey, 2006; Turkheimer et al., 2014).

In sum, the HiTOP model is not the result of some "truth-generating machine" (Meehl, 1992, p. 152). Rather, it is a human construction based on "meaningful but arbitrary" choices (Turkheimer et al., 2008, p. 1588). Fit indices are not an indicator of validity or even replicability (Littlefield et al., 2021; Watts et al., 2020). HiTOP may ultimately be a useful heuristic, but it is false to claim that it is an empirically validated or a data-driven realization of the structure of the symptoms of psychopathology. As noted by Turkheimer (2017),

> internalizing and externalizing are not substrates, with the implication of biological reality. They are dimensions, convenient statistical abstractions. We only think of rotated factors as being more natural than category boundaries because they emerge so effortlessly from the computer programs that rotate them into existence. (p. 41)

Data decisions

Another potential source of bias in factor analysis is the data; the validity of the model depends on the validity of the information used to create it. According to Barocas and Selbst (2016), "advocates of algorithmic techniques like data mining argue that these techniques eliminate human biases from the decision-making process. But an algorithm is only as good as the data it works with" (p. 671). Data decisions are easy when there is a well-defined and circumscribed body of data. For example, input decisions for the five-factor model of personality, from which HiTOP was derived, are based on the lexical hypothesis. According to the lexical hypothesis, the most frequently used descriptors in a given language represent socially important personality traits. The usage correlations among these words results in a factor structure of socially important traits

for a particular society. Here, the input decision is easy because it is possible to analyze an entire lexicon and compare between word types and languages.

Unfortunately, this type of breadth and inclusion is currently unavailable in the area of mental illness. This raises questions about the usefulness of the input used in HiTOP. Are the self-reported symptoms used to create the HiTOP factors all meaningful indicators of psychopathology (e.g., McGrane & Maul, 2020; Michell, 2000)? Furthermore, how many important indicators are missing from the model (Haroz et al., 2017; Huber et al., 2011; Keyes, 2007; van der Krieke et al., 2016)? And how many symptoms are included in the model that are superfluous or do not generalize across cultures, gender, and age (e.g., age-crime curve, Moffitt, 1993; Shulman et al., 2013)? For example, we already know that the data used by HiTOP are biased in terms of culture, race, age, and gender given that they come from studies using samples of Western, educated, industrial, rich, democratic (WEIRD) participants (Arnett, 2008; Henrich et al., 2010; Kaiser & Weaver, 2019; Kohrt et al., 2014; Kohrt & Mendenhall, 2016; Muroff et al., 2008; Neighbors et al., 1989; Weaver & Kaiser, 2015). There is at least one study to indicate that HiTOP will not be robust to changes in symptom input. For example, Wittchen and colleagues (2009) found that even the basic internalizing and externalizing structure was not robust when different ages and different diagnoses were considered. They concluded that "it seems unlikely that fairly simple and robust structural models will ever be derived, given the complexity of psychopathological features across the lifespan" (p. 201).

The lack of representation in psychological research is a problem for all taxonomies. However, it may be significantly more difficult for data-driven models like HiTOP to capture cultural nuance than it is for other approaches (in which it is possible to include cultural concepts of distress; Kaiser et al., 2015; Lewis-Fernandez & Kirmayer, 2019; Weaver & Kaiser, 2015). This is the case because cultural variability is effectively erased as it is dwarfed by the overwhelming amount of data arising from WEIRD samples (which Gone & Kirmayer [2010] called "conceptual imperialism"; see also Henrich et al., 2010). And when data fail to reflect heterogeneity of human experience (Fisher et al., 2018) in terms of race, gender, age, class, and culture, then systemic bias can arise (Cooper & Davids, 1986; Gelfand et al., 2002; Gone et al., 2010). For example, despite disparate symptoms and biological signatures of heart disease by gender (Chuang et al., 2012; Goldberg et al., 1998; Wenger, 1990), many clinical guidelines and practices (e.g., diet, physical activity, and aspirin) are derived from foundational research that was done on men (e.g., Caerphilly Heart Disease and Whitehall Studies of the 1970s and 1980s).

As the use of algorithms based on unrepresentative data has increased, so have the instances of systemic bias, including advertisements that are less likely to be presented to women, Black-sounding names being falsely linked to arrest records, face recognition algorithms failing to recognize the faces of Black people, photo software automatically lightening the skin tones of Black people, failure to identify poor people and Black people with complex health care needs, and predictive policing (Buolamwini & Gebru, 2018; Ferguson, 2019; Lee, 2013; Morse, 2017; Obermeyer et al., 2019). In fact, the first case of an incorrect facial recognition match leading to the arrest of an innocent man has been reported (Hill, 2020).

In summary, there is little evidence to support the claim that HiTOP is more "empirical," "accurate," or verisimilar than existing taxonomies. This is not necessarily a problem in and of itself. What is concerning is that the HiTOP consortium continues to promote its system as objective and empirically valid. As warned by Kleinberg and colleagues (2019), "it would be naïve—even dangerous—to conflate algorithmic with objective" (p. 9). Failing to acknowledge this fact (or worse, promoting the opposite) may lead to overconfidence in the validity of HiTOP and, in turn, promote a mindless application of the system, leading to systemic algorithmic bias for underrepresented groups.

Claim 4. HiTOP Will Lead to Genetic Discovery

> It will become apparent that seeking biology via factor analysis may be just tilting at a windmill.
>
> —Guttman (1992, p. 177)

According to Waszczuk and colleagues (2020), the lack of progress in identifying specific genetic variants that confer risk for psychopathology is due, in part, to poor *DSM* phenotypes. The authors claimed that HiTOP can "accelerate genetic discovery" (p. 8) and solve the problems "that impede progress in psychiatric genetics" (p. 12). In support of this claim, Waszczuk and colleagues reviewed a growing number of studies that have found high heritability estimates and genetic correlations with HiTOP dimensions.

There are at least two reasons HiTOP will not solve the problem of genetic discovery. First, HiTOP probably is not valid; it is a descriptive taxonomy based on symptom correlations. There is little reason to believe that these groupings reflect any natural kinds for which causal genetic variants can be discovered. Second, there is the "gloomy prospect" (Plomin & Daniels, 1987; Turkheimer & Waldron, 2000). Even if HiTOP somehow got everything right, it still would not lead to the

identification of any genetic mechanisms. That is because there are no specific genetic mechanisms to be found (i.e., no "mental illness genes"). Mental illness is too complex. Researchers are converging on the conclusion that complex behavioral phenotypes are likely the result of thousands of genes, each with a negligible effect (Turkheimer, 2016; Visscher et al., 2010). Furthermore, the myriad genes will likely combine and interact in ways that are different for each individual (e.g., intragenomic conflict; Kramer & Bressan, 2015). Genes do not directly cause psychopathology; rather, these genetic correlations are indicators of a general probabilistic influence—an uninterpretable confluence of genes and environment that influence behavior throughout the life span with a substantial random factor (e.g., Bierbach et al., 2017; Flint & Ideker, 2019; Turkheimer, 2016). In other words, even when genetic correlations are found, they may or may not reflect any direct etiological/causal influence on the phenotype.

If the slow progress in this area was caused by poor *DSM* phenotypes, as claimed by the HiTOP consortium, then we should see success in other areas of social science that have better theories and measurement tools. This is not the case; researchers have yet to discover the genetic mechanism for *any* complex human phenotype (intelligence, personality, etc.; Matthews & Turkheimer, 2021). Consider the example of human height. It is more heritable (.8–.9) than mental illness and can be precisely measured. Scientists (e.g., Boyle et al., 2017; Yengo et al., 2018) have identified more than 100,000 single nucleotide polymorphisms, accounting for less than 25% of variance in height (recent nonreplications suggest this percentage is inflated; Berg et al., 2018; Sohail et al., 2019). It remains unclear which, if any, of the identified genetic variants exert a causal/mechanistic influence on height (Boyle et al., 2017). As explained by Turkheimer (2012),

> the unspoken claim is that assiduous attention to statistical significance and population stratification will lead to discovery of an allele with an identifiable biological pathway extending through the many levels of analysis separating the allele from the complex phenomenon it is purported to explain. If I am correct that this is what the GWAS researchers intend, it is no wonder that they don't unpack the content of the claim, because on minimal examination it is so obviously false, false even for something not-really-so-complex as height, never mind delinquency. (p. 62)

Research on the five-factor model of personality has already shown us how genetic discovery will progress under HiTOP. Turkheimer et al. (2014) reviewed the literature on personality and heritability and concluded "that in the genetics of personality, a paradoxical outcome that has been looming for a long time has finally come to pass: personality is heritable, but it has no genetic mechanism" (p. 535). We suspect this conclusion also applies to psychopathology (as well as every other complex behavioral phenotype) regardless of how it is operationalized. Yes, psychopathology is "genetic," but there are no specific genetic mechanisms to discover.

It is also important to address the claim that heritability estimates and genetic correlations can be used to validate the HiTOP hierarchy (Waszczuk et al., 2020). Unfortunately, showing that HiTOP taxa are heritable is relatively meaningless. This is because everything is heritable—Turkheimer and Walkdron's (2000) first law of behavioral genetics. All measurable human differences have genetic correlations. Researchers have found that income, marital status, health insurance coverage, homophobia, military service, frequency of bread eating, and dog ownership are all heritable (Beaver et al., 2015; Fall et al., 2019; Hasselbalch et al., 2010; Hyytinen et al., 2019; Trumbetta et al., 2007; Wehby & Shane, 2019; Zapko-Willmes & Kandler, 2018). Obviously, human genes do not code for whether someone enrolls in health care coverage or joins the military. And yet the heritability estimates for phenotypes such as marital status and owning a dog are just as large as those found for mental illness (as operationalized by HiTOP facet or *DSM* diagnosis). Wicherts and Johnson (2009) showed that it is even possible to find genetic correlations using a random scale. They created a scale with random items from a multidimensional personality measure and then demonstrated that scores on it were heritable. If group differences on an artificial scale are heritable, then how noteworthy is it to show that HiTOP spectra are also heritable? It is not appropriate to use heritability estimates as a method for corroborating a taxonomy:

> Neither the magnitude nor new reports of the existence of heritability in previously unmeasured psychological or behavioural measures alone tells us much of anything. Most importantly, it is not useful as a criterion to judge the biological importance or even construct validity of a psychological measure. (Johnson et al., 2011, p. 263)

But what about genetic correlations? Conway and colleagues (2019) argued that it will be possible to identify specific genetic variants at different levels of HiTOP hierarchy; some variants influence nonspecific psychopathology risk and others confer risk for individual spectra, subfactors, or even symptoms. Waszczuk and

colleagues (2020) provided support for this statement by citing studies that have found an alignment between genetic architecture and the HiTOP structure. Conway et al. concluded that "although these specific genetic factors often are comparatively small, they provide etiological support for a hierarchy" (p. 425). It is a mistake to interpret this "alignment" as validation for HiTOP. Research shows that both genetic and environmental structures often align with the phenotypic structure (e.g., Loehlin & Martin, 2013). It is called the *puzzle of parallel structure* (McCrae et al., 2001; Turkheimer, 2016). One cannot conclude that it is the genetic structure that gives rise to (and validates) HiTOP's structure. In fact, it is likely the reverse, in that "phenotypic variation explains the genetic structure of behavior" (Turkheimer, 2014, p. 536).

In summary, it will be difficult for HiTOP to fulfill its promise to accelerate genetic discovery (Waszczuk et al., 2020). It is another descriptive taxonomy that lumps people according to similar symptom presentations. It proposes a unique hierarchy, but the symptom heterogeneity in the upper-level spectra will likely hinder genetic discovery (Smith et al., 2009). That leaves HiTOP's dimensional rating system as its primary route for facilitating genetic discovery (although the use of continuous measures is not exclusive to HiTOP). Dimensional ratings will make it easier to detect more significant genetic correlations because of increased statistical power (similarly to using larger samples). However, identifying a few hundred more statistically significant genetic correlations does not necessarily translate to a deeper understanding of the genetic causes of psychopathology.

Claim 5. HITOP Is Ready to Use Today

> Because the field of psychology has been reluctant to police itself, the consequences for mental health consumers and the profession at large have been problematic.
>
> —Lilienfeld (2007, p. 53)

According to Ruggero and colleagues (2019), HiTOP "is a viable alternative to classifying mental illness that can be integrated into practice today" (p. 1070). It is "poised to revolutionize the field's understanding of the structure of mental disorder and reshape how diagnostic assessments are performed and utilized" (Hopwood et al., 2019, p. 5). We were unable to find any published studies or empirical data to support these claims.

There is no evidence that practicing clinicians can reliably interpret a HiTOP profile. More than 50 years of research on the fallibility of human judgment (Garb, 2005; Grove et al., 2000; Meehl, 1954) indicates that it will be extremely difficult for clinicians to reliably and validly interpret a symptom report containing potentially dozens of subscale scores (Millon, 1991). Patients are going to score high on multiple spectra, subfactors, and disorders. How will a clinician interpret all of these scores? Currently, for example, there are no established norms or clinical cutoffs, no information for identifying primary or secondary problems, and no interpretation or treatment guidelines. To date, there is not even a standardized measure that can assess the entire HiTOP taxonomy, which means clinicians are on their own to piece together an assessment and then somehow interpret the patchwork of results.

Even if the HiTOP consortium eventually creates a standardized measure with interpretation guidelines, then practitioners will still need to predict which treatment will be most effective for which profile. To date, there are no studies to identify which specific HiTOP profiles respond to which empirically supported treatments.

Finally, there is no evidence that using HiTOP enhances diagnostic or treatment outcomes compared with using other taxonomies. There is not a single study in which clinicians were randomly assigned to use HiTOP or an alternative system to determine whether a particular classification system creates better treatment outcomes. There is at least one study that provides indirect evidence that using HiTOP may not enhance treatment outcomes. Using a manipulated assessment design, Lima and colleagues (2005) randomly assigned clinicians to either receive or not receive the Minnesota Multiphasic Personality Inventory symptom information for their patients. Results showed that the addition of symptom information did not improve treatment outcomes.

It is difficult to reconcile the HiTOP consortium's call for an "empirical" classification system with their recommendation for practitioners to start using a system for which there is no empirical data to support its usefulness. There is not a standardized measure of the entire HiTOP system, there are no empirically derived interpretation and treatment guidelines, and there is yet to be a single published study directly comparing the usefulness of HiTOP to other taxonomies. In fact, there is little, if any, research directly testing any aspect of HiTOP. As noted by Conway and colleagues (2019), "many of the analyses that we have reviewed were carried out using datasets that were not assembled with HiTOP in mind" (p. 428). In other words, support for HiTOP has not actually come from using HiTOP. The recommendation to use HiTOP for clinical purposes is premature at best and reckless at worst.

Table 1. Summary of Competing Approaches to the Classification of Psychopathology

Diagnostic system	Approach	Currently useful	Potential for progress	Strengths	Weaknesses
DSM	Descriptive	✓	✓	Information retrieval, prediction, nomenclature	Atheoretical
HiTOP	Descriptive	X	X	Dimensional ratings	Atheoretical, unfalsifiable
RDoC	Research framework	X	✓	Focused on etiology	Reductionistic
Taxometrics	Taxometrics	X	✓	Falsifiable, search for natural kinds	No evidence of latent taxa

Note: *DSM = Diagnostic and Statistical Manual of Mental Disorders*; HiTOP = Hierarchical Taxonomy of Psychopathology; RDoC = Research Domain Criteria.

A Comparison of Taxonomies

To be scientifically useful a concept must lend itself to the formulation of general laws or theoretical principles which reflect uniformities in the subject matter under study, and which thus provides a basis for explanation, prediction, and generally scientific understanding.

—Hempel (1965, p. 146)

In this section, we compare HiTOP with three alternative taxonomic approaches—the *DSM*,[2] the Research Domain Criteria (RDoc) initiative, and Meehlian taxometrics (see Table 1). We focus on the HiTOP and *DSM* comparison because these are the two taxonomies in direct competition. Both HiTOP and *DSM* are descriptive taxonomies, and HiTOP is promoted as a replacement for *DSM*.

DSM

HiTOP and *DSM* are more similar than different. They are descriptive taxonomies that share the same fundamental assumption: Symptom covariation is meaningful in nature (i.e., like goes with like). Both HiTOP and *DSM* are atheoretical and lump people together because they share the same self-reported symptoms. There is some empirical support for the factor structure illustrated by HiTOP (Conway et al., 2019), but there is also support for the distinctiveness of some *DSM* diagnoses (i.e., evidence against lumping; Gray et al., 2020; Jha et al., 2019; Korgaonkar, Fornito, et al., 2014; Korgaonkar, Williams, et al., 2014; Tung & Brown, 2020; Webb et al., 2019). That said, neither system is a long-term solution to the problem of classification in psychopathology given that both taxonomies are likely "wrong" (i.e., "splendid fictions"; Millon, 1991).

There are two primary differences between HiTOP and the *DSM*. The first difference is how the symptom groupings are created. HiTOP uses factor analysis, whereas the *DSM* uses expert consensus. Both approaches are fallible and rely on subjective decision-making. Expert consensus requires human decisions about how to interpret empirical findings and aggregate them into a coherent and usable taxonomy. Likewise, in factor analysis, there are decisions about mode of representation and how to deal with rotational indeterminacy, the consequence being that HiTOP is not any more "empirical" or "truthful" than the *DSM* approach. The choice between factor analysis and expert consensus is one of personal preference given that both strategies may ultimately lead to something that is clinically useful (e.g., communication, prognosis, treatment planning) even if not valid.

The second difference between HiTOP and *DSM* pertains to the rating system, which is dimensional in HiTOP and categorical in the *DSM*. It is important to underscore that the decision to parse the landscape of psychopathology into categories or facets is based more on expedience than empirical evidence (Turkheimer, 2017). HiTOP facets and *DSM* categories are both artificial delineations. That said, there is research showing that most forms of mental illness (self-reported symptoms) appear to differ in quantity rather than quality (Haslam et al., 2012; Markon et al., 2011; cf. Meehl, 1999). Furthermore, using dimensional ratings increases reliability and statistical power to detect correlations among symptoms and other constructs. Research shows that reliability estimates for specific HiTOP dimensions tend to be stronger than reliability estimates for *DSM* diagnoses. According to Waszczuk and colleagues (2020), 40% of *DSM* diagnoses did not meet acceptable levels of interrater reliability in the field trials of the fifth edition of the *DSM*, whereas reliability estimates for the same diagnoses were strong when rated dimensionally. This comparison is a bit misleading, however, because the field trials' estimates for the *DSM* used clinicians who received no training in the diagnostic categories and did not use structured interviews. Thus, it is not surprising that the reliability estimates would

be low. Consider the reliability estimates for diagnosing a broken bone if medical doctors were not allowed to use x-rays. Proper training and proper assessment tools (i.e., a structured interview) are needed to make reliable diagnoses. Reliability estimates for *DSM* diagnoses tend to be uniformly strong when structured interviews are used (e.g., Osório et al., 2019). That said, reliability estimates for HiTOP are probably going to be superior to diagnoses made using the *DSM* because of statistical necessity, not because it is more valid or scientific. As cautioned by Meehl (1992), "the intrinsic validity (empirical meaningfulness) of a diagnostic construct cannot be dismissed ipso facto on grounds of poor average clinician agreement" (p. 156).

Although symptom ratings tend to be more reliable when operationalized as dimensions rather than categories, note that their usefulness in clinical practice has yet to be validated. In the real world, dichotomous decisions often need to be made, such as to admit or not admit, to intervene or not intervene, or the picking of a diagnostic code for billing (Kendler, 2018). Moreover, there is at least some evidence that clinicians prefer categories to dimensions (Mullins-Sweatt & Widiger, 2009; Sprock, 2003). Furthermore, some have argued that mental illness can build over time until there is a tipping point (or a qualitative difference) in which impairment, symptom severity, or distress becomes too much to bear for an individual (e.g., Nelson et al., 2017). As noted by Kendler (2018), "while not all psychiatric disorders have such dramatic 'avalanche-like' transitions, they are fairly common in clinical psychiatry and challenge the authors' conclusions that there is little viable evidence that psychiatric disorders need to be understood from a categorical perspective" (p. 241).

It is important to evaluate the two taxonomies from a philosophy of science perspective. According to Hempel (1965), a scientifically progressive classification system is characterized by features such as operational definitions, open concepts, descriptions, explanations, predictions, and testable assumptions. It engenders assertions about origins and outcomes by weaving a nomological net of relationships between the taxa and their correlates (Meehl & Golden, 1982). A useful taxonomy should "tell us a lot about the patient—the course, the likely etiologic process, the best treatment, etc." (Kendler, 2018, p. 242), and it should have *generative power* and provide us with new attributes, relations, or taxa, that is, ones other than those used to construct it (Millon, 1991).

As imperfect as it is, the *DSM* exhibits many of the features found in a useful taxonomy: (a) It provides descriptive information and explanations about the disorders (e.g., discussion of course, severity, differential diagnosis, why specific disorders have been added or removed); (b) it distinguishes among symptoms, some of which are necessary to the syndrome (e.g., Criterion A) and some of which are supplementary to the syndrome; (c) it considers issues related to duration and persistence; (d) it integrates impairment ratings to reduce overpathologizing; (e) it specifies inclusion and exclusion criteria; (f) it allows for information retrieval (e.g., prevalence, comorbid conditions); (g) it allows for prediction (e.g., one can go to the literature to determine which treatment will work for which specific disorders); (h) it includes cultural considerations (cultural formulation and cultural concepts of distress); and (i) it contains at least some information related to risk and developmental factors (e.g., major stressor required for PTSD; identifies disorders developing in adulthood vs. childhood). In sum, the *DSM* provides hundreds of pages of information related to its categories.

HiTOP, on the other hand, exhibits few, if any, of the features found in a useful taxonomy. Its classification system is an interpretation of factor analytic results. It is a single picture. Absent one's knowledge and previous experience with *DSM* descriptions and disorders, HiTOP contains no additional information. It contains no explanations, no descriptive information (other than symptom labels and lists), no necessary symptoms, no inclusion or exclusion criteria, no information about how to integrate impairment severity, no information about prevalence, and no information on underlying developmental processes, and it ignores differences in culture, age, and/or gender. Furthermore, despite claims about eliminating comorbidity, it provides no information about how to interpret subscale comorbidity (i.e., when patients score high on multiple spectra, subfactors, and disorders).

It may be more accurate to think of HiTOP as a sorting algorithm (or multifaceted measurement tool) rather than a classification system. It does not feature information that lends itself to scientific discourse, disagreement, or progress. HiTOP is a statistical outcome from testing correlations among a large set of symptom items.

We acknowledge that HiTOP is much newer than *DSM*, and at some point, it may have a standardized measure with clinical cutoffs and interpretation guidelines and include descriptive information for the different symptom profiles (e.g., base rates, course of illness, etc.). If this happens, then the question is which of these two systems (HiTOP or *DSM*) is better positioned to evolve from a system based on observable characteristics to one based in theory (Hempel, 1965; Millon, 1991). We contend that the *DSM* has more potential for scientific progress than HiTOP. Ironically, the *DSM*'s most cited "weakness" may actually be its greatest strength with regard to potential for scientific change. The *DSM* is not bound by an analytic procedure but

rather is fueled by scientific debate (Zachar & Kendler, 2007). If scientific progress and self-correction come from disagreement (Lakatos, 1970; Meehl, 1978; Popper, 1959), then look no further than a group of human scientists. The *DSM* can be altered to incorporate more specific explanations and descriptions, additional open concepts, and even theory. There is a path for *DSM* in which "the various classes or categories distinguished now are no longer defined just in terms of symptoms, but rather in terms of the key concept of theories, which are intended to *explain* the observable behavior including the symptoms in question" (Hempel, 1965, p. 149). The *DSM* could be changed back to a theoretical system as quickly as it was changed from being one (the first and second editions of the *DSM* were theoretical; the third edition changed to a descriptive system).

HiTOP does not have a clear path for scientific and taxonomic progress. The main mode of change for HiTOP is to add or subtract symptom information in its analysis. This may lead to small changes in its structure or factor labels, but it will not lead to the type of scientific evolution that characterizes progressive taxonomies (description to theory). HiTOP was created using a statistic within a theoretical vacuum; there are few, if any, specific predictions and hypotheses that can be falsified, which would result in corrective change over time. Furthermore, HiTOP may even hinder progress because it may be creating larger, more heterogeneous factors that do not reflect meaningful etiological differences. This can obscure discovery and lead to more nonreplicable findings in the literature.

Research Domain Criteria initiative

Launched in 2009, RDoC is the National Institute of Mental Health's (NIMH) solution to the problems associated with descriptive taxonomies like *DSM* and HiTOP. Instead of focusing on symptom presentations, RDoC is concerned with etiology. Using an endophenotypic approach (Gottesman & Gould, 2003), RDoC specifies a set of intermediate constructs (negative and positive valence, cognitive systems, social process systems, and arousal systems) thought to form the link between mental illness and some biological or genetic process (Cuthbert & Insel, 2013).

RDoC is unusable in clinical settings. It cannot be used for diagnosis, case conceptualization, treatment choice, or billing options. However, this is to be expected because RDoC is not yet a taxonomy; it is a "framework for research on pathophysiology, especially for genomics and neuroscience" (Insel et al., p. 748).

Ostensibly, RDoC has more potential for scientific progress than HiTOP and *DSM*. Its goal is to characterize psychopathology in terms of etiology instead of description.

Furthermore, it is not tied to a particular clinical outcome or a statistical procedure. Thus, researchers are free to explore new syndromes. That said, RDoC does not explicitly promote theory building or the generation of falsifiable mechanistic explanations; instead, the focus is on identifying specific genes and/or markers of neurological dysfunction associated with its list of endophenotypes.

Unfortunately, the scientific potential of RDoC is limited by biological reductionism (e.g., Lilienfeld, 2014). In the RDoC framework, mental illness is a "brain disorder." The overriding purpose is to understand the biological and genetic basis of mental illness, not its psychological and environmental bases. This is a high-risk strategy because it is possible that low-level brain and genetic factors do not have a direct causal effect on higher level psychological phenotypes (Turkheimer, 2017). It also means that RDoC is wedded to neuroimaging tools such as MRI and functional MRI, which are "not currently suitable for brain biomarker discovery or for individual-differences research" (Elliott et al., 2020, p. 792; Weinberger & Radulescu, 2020). This has culminated in a research literature characterized by underpowered studies and nonreplicable findings (Button et al., 2013; Lilienfeld, 2014; Parnas, 2014; Szucs & Ioannidis, 2020). Even Thomas Insel, who launched RDoC, now questions its potential for success:

> I spent 13 years at NIMH really pushing on the neuroscience and genetics of mental disorders, and when I look back on that I realize that while I think I succeeded at getting lots of really cool papers published by cool scientists at fairly large costs—I think $20 billion—I don't think we moved the needle in reducing suicide, reducing hospitalizations, improving recovery for the tens of millions of people who have mental illness. I hold myself accountable for that. (Rogers, 2017, para 5)

Taxometrics

Bootstrap taxometrics (Meehl & Golden, 1982) was Meehl's response to the unfalsifiable and atheoretical nature of symptom-based statistical clustering (Meehl, 1978, 1989b). According to Meehl (1995),

> we admire Linnaeus, the creator of modern taxonomy, for discerning the remarkable truth—a "deep structure" fact, as Chomsky might say—that the bat doesn't sort with the chickadee and the whale doesn't sort with the pickerel, but both are properly sorted with the grizzly bear . . . I see classification as an enterprise that aims to carve nature at its joints (Plato). (p. 267)

To this end, he created a mathematical method for testing the existence of latent taxa or "natural kinds." It should come as no surprise that Meehl's critique of cluster analysis (and psychological science more generally) has motivated much of this critical evaluation. The HiTOP approach to classification is history repeating itself all over again.

Meehlian taxometrics is not usable in clinical settings, but it is more scientifically progressive than HiTOP and *DSM*. It provides a method to corroborate or refute theories of mental illness. From a Popperian perspective, taxometrics has been hugely successful; nearly every proposed taxon has been refuted (falsified). This does not completely shut the door on the existence of mental illness taxons, but it raises serious doubts.

Recommendation

Without theory-driven models to guide the interpretation of data, it is not likely that any empirical truth will emerge.

—Follette and Houts (1996, p. 1131)

Scientifically progressive taxonomies tend to evolve over time from description to theory. Descriptive taxonomies, like *DSM* and HiTOP, can be useful, but they should be considered a stopgap. It is time for clinical psychology to put its resources and efforts into developing a theoretically derived system that can *explain* mental illness. A theory-based classification system would not be tied to a specific level of analysis or current diagnostic syndromes or rely on finding an association between some genetic/biological measure and a clinical outcome. Rather, the focus would be on creating and testing mechanistic explanations of mental illness.

The research process used in clinical psychology is often atheoretical and backward. Science usually starts with a theory to explain a particular outcome; then, experiments are conducted to test the predictions derived from that theory. But in clinical psychology, researchers focus on the outcome (diagnosis or symptom dimension) rather than the explanation. There appears to be more interest in obtaining the "hard to get" clinical sample than there is in proposing theories (i.e., falsifiable mechanistic explanations) to explain the development of the clinical problems. The dominant research design in clinical psychology is to compare people with varying levels of psychopathology to determine whether they differ on some measure (e.g., amygdala activation). And when between-groups differences are inevitably found (Meehl, 1978), they are assumed

to reflect an etiological process. This kind of post hoc conjecturing is a problem because any difference found in the clinical group (relative to control) could be a concomitant or scar of experiencing psychopathology rather than a part of its etiology.

Pursuing a theory-based classification system may help to curb clinical psychology's obsession with testing samples rather than theories. Furthermore, it would push researchers to use more rigorous research methodologies such as behavioral high-risk designs and targeted prevention interventions (in which participants are selected on individual differences in a hypothesized risk factor rather than the clinical outcome). Examples of this kind of theory-based research include the hopelessness theory of depression (Abramson et al., 1989) and Newman's (1998) attention-based theory of psychopathy. The hopelessness theory specifies a falsifiable etiological sequence that explains a clinical outcome: It specifies distal, proximal, contributory, and sufficient causes as well as both mechanisms (e.g., hopelessness) and moderators (e.g., stress, cognitive vulnerability) of the outcome of interest. It also proposes a theory-based clinical outcome that is not tied to the current descriptive system (hopelessness subtype of depression). Along these same lines, Newman's (1998) attention theory of psychopathy is an exemplar of a progressive theory (Lakatos, 1970) that can both explain existing findings and generate novel predictions that cannot be explained by competing theories (e.g., the low fear hypothesis).

Obviously, a theory-based taxonomy remains a pipe dream. The field still needs to build stronger explanatory theories, rigorously test them (alone and in competition), and somehow integrate the findings into a taxonomy. The question is what to do in the meantime. We recommend a bifurcation strategy. Clinicians should continue to use the *DSM* while researchers focus on theory development and testing. We choose the *DSM* not because we believe it to be particularly valid but because it is currently the most useful taxonomy in clinical practice. As theory development progresses, the information can be integrated into *DSM* (similarly to how intervention research has influenced treatment guidelines), or it can be used to create an entirely new system. Research using RDoC and taxometrics can complement the theory-driven approach and be used in parallel. Although the RDoC is limited by biological reductionism, it can still serve as a basis for theory development. Likewise, taxometrics can be used to try to corroborate new theoretical subtypes. In contrast, there appears to be limited incremental value in pursuing HiTOP, which is another descriptive system. *DSM* already meets the need for a useful descriptive

taxonomy that can be used in clinical practice. It is possible that HiTOP could also meet this need at some point, but it is ultimately handcuffed by its inability to evolve over time.

Conclusion

> Is there a named cognitive bias describing the preference for a concrete quantitative answer to a complex question, even if it is invalid?
>
> —Turkheimer (2020)

Factor analysis provides a straightforward, intuitive, and parsimonious solution to the problem of classification. Researchers can impose a hierarchical structure on mental illness with the push of a button. According to Waszczuk and colleagues (2020), the result of this button push "promises to resolve problems of comorbidity, heterogeneity, and arbitrary diagnostic thresholds" (p. 12). It is a "paradigm shift" (Kotov et al., 2018) that will transform mental health research (Conway et al., 2019), improve clinical practice (Ruggero et al., 2019), and advance genetic discovery (Latzman et al., 2020; Waszczuk et al., 2020).

The purpose of this article was to critically evaluate the HiTOP approach and its purported advantages. We conclude that the extraordinary claims about HiTOP are not matched by extraordinary evidence (Gillispie et al., 1999; Sagan, 1979); it appears the HiTOP consortium is writing checks their taxonomy cannot cash. Unless psychopathology plays by a different set of rules than nearly every other realm of nature, the result of pushing the factor analysis button is an incorrect answer. For HiTOP to be valid, it would mean that (a) self-reported symptom expressions are meaningful indicators of development processes and the etiology of psychopathology; (b) all of the symptom indicators are equally important (deserve equal weighting) for classifying psychopathology; (c) equifinality and multifinality do not apply to psychopathology; (d) the expression and reporting of symptoms are not influenced by sex, culture, or age (and failing to account for them does not lead to algorithmic bias); and (e) a dimensional interpretation/simple structure approach represents the structure of psychopathology symptom data. To date, there is little evidence to support any of these statements. Moreover, HiTOP does not lend itself to theory building. It does not feature the characteristics of a falsifiable, scientifically progressive, and evolving taxonomy. It is bound to a statistical procedure in which change comes from adding or subtracting symptom information rather than through the falsification of specific hypotheses.

More than 40 years ago, Meehl (1978) argued that psychology was not progressing like the hard sciences because of shoddy theorizing and an overreliance on null hypothesis testing. The problems he noted are currently exemplified by the push for atheoretical, statistically driven structural taxonomies of psychopathology. He tried to remind us that creating specific and falsifiable theory (e.g., Popper, 1959) is necessary for scientific progress. Psychology's statistically driven approach to classification seems to fail this critical requirement because it is difficult to "be wrong" in the absence of any specific theoretical hypotheses while reporting the output of factor analyses. Because of this and the limitations discussed in this article, replacing the *DSM* with HiTOP has the potential to hinder progress on understanding the etiology of psychopathology. We recommend a bifurcation strategy in which the *DSM* continues to be used in clinical settings because of its usefulness while researchers focus on creating and testing falsifiable theories of mental illness that can eventually inform the *DSM* or lead to a new theory-based classification system.

ORCID iDs

Gerald J. Haeffel (iD) https://orcid.org/0000-0002-4029-1493
Lesley Jo Weaver (iD) https://orcid.org/0000-0001-8104-0511
Aaron J. Fisher (iD) https://orcid.org/0000-0001-9754-4618
Ivan Vargas (iD) https://orcid.org/0000-0002-0787-5630

Notes

1. There is an extensive literature questioning the logic and appropriateness of factor modeling for understanding complex phenotypes (e.g., interpretation of models, failure to test assumptions of quantitative structure, etc.). For additional discussion

of these issues, please see Aristodemou and Fried (2020), Bornovalova et al. (2020), Heene (2013), Rhemtulla et al. (2020), van Bork et al. (2017), and Wittchen and Beesdo-Baum (2018).

2. This comparison would also apply to the International Statistical Classification of Diseases and Related Health Problems (ICD); one minor difference is that ICD focuses more on public health and applicability to a diverse worldly population.

References

Abramson, L. Y., Metalsky, G. I., & Alloy, L. B. (1989). Hopelessness depression: A theory-based subtype of depression. *Psychological Review*, *96*, 358–372.

Adigun, R., Goyal, A., Bansal, P., & Hariz, A. (2020). *Systemic sclerosis (CREST syndrome)*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK430875/

Aristodemou, M. E., & Fried, E. I. (2020). Common factors and interpretation of the p factor of psychopathology. *Journal of the American Academy of Child and Adolescent Psychiatry*, *59*(4), 465–466. https://doi.org/10.1016/j.jaac.2019.07.953

Arnett, J. (2008). The weirdest people in the world. *American Psychologist*, *63*(7), 602–614.

Barocas, S., & Selbst, A. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732.

Beaver, K. M., Barnes, J. C., Schwartz, J. A., & Boutwell, B. B. (2015). Enlisting in the military: The influential role of genetic factors. *SAGE Open*, *5*(2). https://doi.org/10.1177/2158244015573352

Berg, J. J., Harpak, A., Sinnott-Armstrong, N., Jørgensen, A. M., Mostafavi, H., Field, Y., Boyle, E. A., Zhang, X., Racimo, F., Pritchard, J. K., & Coop, G. (2018). Reduced signal for polygenic adaptation of height in UK Biobank. *eLife*, *8*, Article e39725. https://doi.org/10.1101/354951

Bierbach, D., Laskowski, K. L., & Wolf, M. (2017). Behavioural individuality in clonal fish arises despite near-identical rearing conditions. *Nature Communications*, *8*, Article 15361. https://doi.org/10.1038/ncomms15361

Bolles, R. C. (1962). The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports*, *11*(3), 639–645. https://doi.org/10.2466/pr0.1962.11.3.639

Bornovalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis in psychopathology research: Appreciating benefits and limitations. *Biological Psychiatry*, *88*, 18–27.

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell*, *169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 77–91. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Button, K., Ioannidis, J., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376. https://doi.org/10.1038/nrn3475

Centers for Disease Control and Prevention. (2019). *Chronic diseases in America*. https://www.cdc.gov/chronicdisease/pdf/infographics/chronic-disease-H.pdf

Chuang, M. L., Massaro, J. M., Levitzky, Y. S., Fox, C. S., Manders, E. S., Hoffmann, U., & O'Donnell, C. J. (2012). Prevalence and distribution of abdominal aortic calcium by gender and age group in a community-based cohort (from the Framingham Heart Study). *The American Journal of Cardiology*, *110*, 891–896.

Cicchetti, D., & Rogosch, F. (1996). Equifinality and multifinality in developmental psychopathology. *Development and Psychopathology*, *8*(4), 597–600. https://doi.org/10.1017/S0954579400007318

Cojocaru, M., Cojocaru, I. M., Silosi, I., & Vrabie, C. D. (2011). Manifestations of systemic lupus erythematosus. *Maedica*, *6*(4), 330–336.

Conway, C. C., Forbes, M. K., Forbush, K. T., Fried, E. I., Hallquist, M. N., Kotov, R., Mullins-Sweatt, S. N., Shackman, A. J., Skodol, A. E., South, S. C., Sunderland, M., Waszczuk, M. A., Zald, D. H., Afzali, M. H., Bornovalova, M. A., Carragher, N., Docherty, A. R., Jonas, K. G., Krueger, R. F., . . . Eaton, N. R. (2019). A hierarchical taxonomy of psychopathology can transform mental health research. *Perspectives on Psychological Science*, *14*(3), 419–436. https://doi.org/10.1177/1745691618810696

Cooper, R., & David, R. (1986). The biological concept of race and its application to public health and epidemiology. *Journal of Health Politics, Policy and Law*, *11*(1), 97–116. https://doi.org/10.1215/03616878-11-1-97

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, *11*, Article 126. https://doi.org/10.1186/1741-7015-11-126

Dawkins, R., & Wong, Y. (2016). *The ancestor's tale: A pilgrimage to the dawn of evolution*. Mariner.

De Vries, Y. A., Al-Hamzawi, A., Alonso, J., Borges, G., Bruffaerts, R., Bunting, B., Caldas-de-Almeida, J. M., Cia, A. H., De Girolamo, G., Dinolova, R. V., Esan, O., Florescu, S., Gureje, O., Haro, J. M., Hu, C., Karam, E. G., Karam, A., Kawakami, N., & Kiejna, A., . . . WHO World Mental Health Survey Collaborators. (2019). Childhood generalized specific phobia as an early marker of internalizing psychopathology across the lifespan: Results from the World Mental Health Surveys. *BMC Medicine*, *17*(1), Article 101. https://doi.org/10.1186/s12916-019-1328-3

DeYoung, C. G., Chmielewski, M., Clark, L. A., Condon, D. M., Kotov, R., Krueger, R. F., Lynam, D. R., Markon, K. E., Miller, J. D., Mullins-Sweatt, S. N., Samuel, D. B., Sellbom, M., South, S. C., Thomas, K. M., Watson, D., Watts, A. L., Widiger, T. A., Wright, A. & HiTOP Normal Personality Workgroup. (2020). The distinction between symptoms and traits in the Hierarchical Taxonomy of Psychopathology (HiTOP). *Journal of Personality*. Advance online publication. https://doi.org/10.1111/jopy.12593

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What is the test-retest reliability of common task-functional MRI measures? New empirical

evidence and a meta-analysis. *Psychological Science*, *31*, 792–806. https://doi.org/10.1177/0956797620916786

Fall, T., Kuja-Halkola, R., Dobney, K., Westgarth, C., & Magnusson, P. K. E. (2019). Evidence of large genetic influences on dog ownership in the Swedish Twin Registry has implications for understanding domestication and health associations. *Scientific Reports*, *9*, Article 7554. https://doi.org/10.1038/s41598-019-44083-9

Ferguson, A. G. (2019). *The rise of big data policing: Surveillance, race, and the future of law enforcement*. New York University Press.

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, USA*, *115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Flint, J., & Ideker, T. (2019). The great hairball gambit. *PLOS Genetics*, *15*(11), Article e1008519. https://doi.org/10.1371/journal.pgen.1008519

Follette, W. C., & Houts, A. C. (1996). Models of scientific progress and the role of theory in taxonomy development: A case study of the DSM. *Journal of Consulting and Clinical Psychology*, *64*(6), 1120–1132. https://doi.org/10.1037/0022-006X.64.6.1120

Garb, H. N. (2005). Clinical judgment and decision making. *Annual Review of Clinical Psychology*, *1*(1), 67–89.

Gelfand, M. J., Raver, J. L., & Holcombe Ehrhart, K. (2002). Methodological issues in cross-cultural organizational research. In S. G. Rogelberg (Ed.), *Blackwell handbooks of research methods in psychology. Handbook of research methods in industrial and organizational psychology* (pp. 216–246). Blackwell Publishing.

Gillispie, C. C., Gratton-Guinness, I., & Fox, R. (1999). *Pierre Simon Laplace: A life in exact science*. Princeton University Press.

Goldberg, R. J., O'Donnell, C., Yarzebski, J., Bigelow, C., Savageau, J., & Gore, J. M. (1998). Sex differences in symptom presentation associated with acute myocardial infarction: A population-based perspective. *American Heart Journal*, *136*(2), 189–195.

Gone, J. P., & Kirmayer, L. J. (2010). On the wisdom of considering culture and context in psychopathology. In T. Millon, R. F. Krueger, & E. Simonsen (Eds.), *Contemporary directions in psychopathology: Scientific foundations of the DSM-V and ICD-11* (pp. 72–96). The Guilford Press.

Gottesman, I. I., & Gould, T. D. (2003). The endophenotype concept in psychiatry: Etymology and strategic intentions. *American Journal of Psychiatry*, *160*, 636–645.

Gray, J. P., Müller, V. I., Eickhoff, S. B., & Fox, P. T. (2020). Multimodal abnormalities of brain structure and function in major depressive disorder: A meta-analysis of neuroimaging studies. *The American Journal of Psychiatry*, *177*(5), 422–434. https://doi.org/10.1176/appi.ajp.2019.19050560

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Guttman, L. (1982). Facet theory, smallest space analysis, and factor analysis. *Perceptual and Motor Skills*, *54*, 487–493.

Guttman, L. (1992). The irrelevance of factor analysis for the study of group differences. *Multivariate Behavioral Research*, *27*(2), 175–204.

Haroz, E. E., Ritchey, M., Bass, J. K., Kohrt, B. A., Augustinavicius, J., Michalopoulos, L., Burkey, M. D., & Bolton, P. (2017). How is depression experienced around the world? A systematic review of qualitative literature. *Social Science & Medicine*, *183*, 151–162.

Haslam, N., Holland, E., & Kuppens, P. (2012). Categories versus dimensions in personality and psychopathology: A quantitative review of taxometric research. *Psychological Medicine*, *42*(5), 903–920. https://doi.org/10.1017/S0033291711001966

Hasselbalch, A. L., Silventoinen, K., Keskitalo, K., Pietiläinen, K., Rissanen, A., Heitmann, B., Kyvik, K. O., Sørensen, T. I. A., & Kaprio, J. (2010). Twin study of heritability of eating bread in Danish and Finnish men and women. *Twin Research and Human Genetics*, *13*(2), 163–167. https://doi.org/10.1375/twin.13.2.163

Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology. *Frontiers in Psychology*, *4*, Article 246. https://doi.org/10.3389/fpsyg.2013.00246

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. Free Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Hill, K. (2020, June 24). Wrongfully accused by an algorithm. *The New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Hopwood, C. J., Bagby, R. M., Gralnick, T., Ro, E., Ruggero, C., Mullins-Sweatt, S., Kotov, R., Bach, B., Cicero, D. C., Krueger, R. F., Patrick, C. J., Chmielewski, M., DeYoung, C. G., Docherty, A. R., Eaton, N. R., Forbush, K. T., Ivanova, M. Y., Latzman, R. D., Pincus, A. L., . . . Zimmermann, J. (2019). Integrating psychotherapy with the hierarchical taxonomy of psychopathology (HiTOP). *Journal of Psychotherapy Integration*, *30*(4), 477–494. https://doi.org/10.1037/int0000156

Huber, M., Knottnerus, J. A., Green, L., van der Horst, H., Jadad, A. R., Kromhout, D., Leonard, B., Lorig, K., Loureiro, M. I., van der Meer, J. W., Schnabel, P., Smith, R., van Weel, C., & Smid, H. (2011). How should we define health? *BMJ*, *343*, Article 4163. https://doi.org/10.1136/bmj.d4163

Hyytinen, A., Ilmakunnas, P., Johansson, E., & Toivanen, O. (2019). Heritability of lifetime earnings. *Journal of Economic Inequality*, *17*, 319–335. https://doi.org/10.1007/s10888-019-09413-x

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, *167*(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Jha, M. K., Minhajuddin, A., Chin-Fatt, C., Greer, T. L., Carmody, T. J., & Trivedi, M. H. (2019). Sex differences in the association of baseline c-reactive protein (CRP)

and acute-phase treatment outcomes in major depressive disorder: Findings from the EMBARC study. *Journal of Psychiatric Research*, *113*, 165–171.

Johnson, W., Penke, L., & Spinath, F. M. (2011). Heritability in the era of molecular genetics: Some thoughts for understanding genetic influences on behavioural traits. *European Personality Reviews*, *25*, 254–266.

Jolly, E., & Chang, L. J. (2019). The flatland fallacy: Moving beyond low–dimensional thinking. *Topics in Cognitive Science*, *11*(2), 433–454. https://doi.org/10.1111/tops .12404

Kaiser, B. N., Haroz, E. E., Kohrt, B. A., Bolton, P. A., Bass, J. K., & Hinton, D. E. (2015). "Thinking too much": A systematic review of a common idiom of distress. *Social Science & Medicine*, *147*, 170–183. https://doi.org/10.1016/j.soc scimed.2015.10.044

Kaiser, B. N., & Weaver, L. (2019). Culture-bound syndromes, idioms of distress, and cultural concepts of distress: New directions for an old concept in psychological anthropology. *Transcultural Psychiatry*, *56*(4), 589–598. https:// doi.org/10.1177/1363461519862708

Kendler, K. S. (2018). Classification of psychopathology: Conceptual and historical background. *World Psychiatry*, *17*(3), 241–242. https://doi.org/10.1002/wps.20549

Kendler, K. S., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders? *Psychological Medicine*, *41*(6), 1143–1150.

Keyes, C. L. M. (2007). Promoting and protecting mental health as flourishing: A complementary strategy for improving national mental health. *American Psychologist*, *62*(2), 95–108. https://doi.org/10.1037/0003-066X.62.2.95

Khera, A. V., Emdin, C. A., Drake, I., Natarajan, P., Bick, A. G., Cook, N. R., Chasman, D. I., Baber, U., Mehran, R., Rader, D. J., Fuster, V., Boerwinkle, E., Melander, O., Orho-Melander, M., Ridker, P. M., & Kathiresan, S. (2016). Genetic risk, adherence to a healthy lifestyle, and coronary disease. *New England Journal of Medicine*, *375*, 2349–2358.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination in the age of algorithms. *Journal of Legal Analysis*, *10*, 113–174. https://doi.org/10.1093/ jla/laz001

Kohrt, B. A., & Mendenhall, E. (2016). *Global mental health: Anthropological perspectives*. Routledge.

Kohrt, B. A., Rasmussen, A., Kaiser, B. N., Haroz, E. E., Maharjan, S. M., Mutamba, B. B., de Jong, J. T., & Hinton, D. E. (2014). Cultural concepts of distress and psychiatric disorders: Literature review and research recommendations for global mental health epidemiology. *International Journal of Epidemiology*, *432*(2), 365–406.

Korgaonkar, M. S., Fornito, A., Williams, L. M., & Grieve, S. M. (2014). Abnormal structural networks characterize major depressive disorder: A connectome analysis. *Biological Psychiatry*, *76*(7), 567–574. https://doi.org/10.1016/j .biopsych.2014.02.018

Korgaonkar, M. S., Williams, L., Song, Y., Usherwood, T., & Grieve, S. (2014). Diffusion tensor imaging predictors of treatment outcomes in major depressive disorder. *British Journal of Psychiatry*, *205*(4), 321–328. https://doi .org/10.1192/bjp.bp.113.140376

Kotov, R., Jonas, K. G., Carpenter, W. T., Dretsch, M. N., Eaton, N. R., Forbes, M. K., Forbush, M. A., Widiger, T. A., Wright, A., Zald, D. H., Krueger, R. F., & Watson, D. (2020). Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): I. Psychosis superspectrum. *World Psychiatry*, *19*, 151–172. https://doi.org/10.1002/ wps.20730

Kotov, R., Krueger, R. F., & Watson, D. (2018). A paradigm shift in psychiatric classification: The Hierarchical Taxonomy of Psychopathology (HiTOP). *World Psychiatry*, *17*(1), 24–25. https://doi.org/10.1002/wps.20478

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., Clark, L. A., Eaton, N. R., Forbes, M. K., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E., Ivanova, M. Y., Lynam, D. R., Markon, K., . . . Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, *126*(4), 454–477. https://doi.org/10.1037/abn0000258

Kramer, P., & Bressan, P. (2015). Humans as superorganisms: How microbes, viruses, imprinted genes, and other selfish entities shape our behavior. *Perspectives on Psychological Science*, *10*(4), 464–481. https://www.doi .org/10.1177/1745691615583131

Krueger, R. F., Kotov, R., Watson, D., Forbes, M. K., Eaton, N. R., Ruggero, C. J., Simms, L. J., Widiger, T. A., Achenbach, T. M., Bach, B., Bagby, R. M., Bornovalova, M. A., Carpenter, W. T., Chmielewski, M., Cicero, D. C., Clark, L. A., Conway, C., DeClercq, B., DeYoung, C. G., . . . Zimmermann, J. (2018). Progress in achieving quantitative classification of psychopathology. *World Psychiatry*, *17*, 282–293. https://doi.org/10.1002/wps.20566

Lakatos, L. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge University Press.

Latzman, R. D., DeYoung, C. G., & HiTOP Neurobiological Foundations Workgroup. (2020). Using empirically-derived dimensional phenotypes to accelerate clinical neuroscience: The Hierarchical Taxonomy of Psychopathology (HiTOP) framework. *Neuropsychopharmacology*, *45*, 1083–1085.

Lee, D. (2013, February 4). Google searches expose racial bias, says study of names. *BBC*. https://www.bbc.com/ news/technology-21322183

Lewis-Fernandez, R., & Kirmayer, L. J. (2019). Cultural concepts of distress and psychiatric disorders: Understanding symptom experience and expression in context. *Transcultural Psychiatry*, *56*, 786–803.

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science*, *2*(1), 53–70. https://doi.org/10.1111/j.1745-6916.2007.00029.x

Lilienfeld, S. O. (2014). The Research Domain Criteria (RDoC): An analysis of methodological and conceptual challenges. *Behaviour Research and Therapy*, *62*, 129–139. https:// doi.org/10.1016/j.brat.2014.07.019

Lima, E. N., Stanley, S., Kaboski, B., Reitzel, L. R., Richey, A., Castro, Y., Williams, F. M., Tannenbaum, K. R., Stellrecht, N. E., Jakobsons, L. J., Wingate, L. R., & Joiner, T. E., Jr.

(2005). The incremental validity of the MMPI-2: When does therapist access not enhance treatment outcome? *Psychological Assessment*, *17*(4), 462–468. https://doi.org/10.1037/1040-3590.17.4.462

Littlefield, A. K., Lane, S. P., Gette, J. A., Watts, A. L., & Sher, K. J. (2021). The "Big Everything": Integrating and investigating dimensional models of psychopathology, personality, personality pathology, and cognitive functioning. *Personality Disorders: Theory, Research, and Treatment*, *12*(2), 103–114. https://doi.org/10.1037/per0000457

Loehlin, J. C., & Martin, N. G. (2013). General and supplementary factors of personality in genetic and environmental correlation matrices. *Personality and Individual Differences*, *54*(6), 761–766. https://doi.org/10.1016/j.paid.2012.12.014

Maraun, M. D. (1997). Appearance and reality: Is the big five the structure of trait descriptors? *Personality and Individual Differences*, *22*, 629–647.

Markey, P. M., & Markey, C. N. (2006). A spherical conceptualization of personality traits. *European Journal of Personality*, *20*, 169–193. https://doi.org/10.1002%2Fper.582

Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological Bulletin*, *137*(5), 856–879. https://doi.org/10.1037/a0023678

Matthews, L. J., & Turkheimer, E. (2021). Across the great divide: Pluralism and the hunt for missing heritability. *Synthese*, *198*, 2297–2311. https://doi.org/10.1007/s11229-019-02205-w

McConaghy, J. R. (2020). Outpatient evaluation of the adult with chest pain. *UpToDate*. http://222.247.54.203:1057/contents/outpatient-evaluation-of-the-adult-with-chest-pain

McCrae, R. R., Jang, K. L., Livesley, W. J., Riemann, R., & Angleitner, A. (2001). Sources of structure: Genetic, environmental, and artifactual influences on the covariation of personality traits. *Journal of Personality*, *69*, 511–535. https://doi.org/10.1111/1467-6494.694154

McGrane, J., & Maul, A. (2020). The human sciences, models and metrological mythology. *Measurement*, *152*, Article 107346.

McGrath, R.E. (2005). Conceptual complexity and construct validity. *Journal of Personality Assessment*, *85*(2), 112–124. https://doi.org/10.1207/s15327752jpa8502_02

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press. https://doi.org/10.1037/11281-000

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Meehl, P. E. (1989a). Paul E. Meehl. In G. Lindzey (Ed.), *A history of psychology in autobiography* (Vol. 8, pp. 337–389). Stanford University Press.

Meehl, P. E. (1989b). Schizotaxia revisited. *Archives of General Psychiatry*, *46*, 935–944. https://doi.org/10.1001/archpsyc.1989.01810100077015

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality*, *60*(1), 117–174. https://doi.org/10.1111/j.14676494.1992.tb00269.x

Meehl, P. E. (1995). Bootstraps taxometrics. Solving the classification problem in psychopathology. *The American Psychologist*, *50*(4), 266–275. https://doi.org/10.1037//0003-066x.50.4.266

Meehl, P. E. (1999). Clarifications about taxometric method. *Applied & Preventive Psychology*, *8*, 165–174.

Meehl, P. E., & Golden, R. R. (1982). Taxometric methods. In P. Kendall & J. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 127–181). Wiley.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*, 639–667. https://doi.org/10.1177/0959354300105004

Millon, T. (1991). Classification in psychopathology: Rationale, alternatives, and standards. *Journal of Abnormal Psychology*, *100*(3), 245–261. https://doi.org/10.1037/0021-843X.100.3.245

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, *100*(4), 674–701. https://doi.org/10.1037/0033-295X.100.4.674

Morse, J. (2017). App creator apologizes for 'racist' filter that lightens skin tones. *Mashable*. https://mashable.com/2017/04/24/faceapp-racism-selfie/#zeUItoQB5iqI

Mullins-Sweatt, S. N., & Widiger, T. A. (2009). Clinical utility and DSM-V. *Psychological Assessment*, *21*, 302–312. https://doi.org/10.1037/a0016607

Muroff, J., Edelsohn, G. A., Joe, S., & Ford, B. C. (2008). The role of race in diagnostic and disposition decision making in a pediatric psychiatric emergency service. *General Hospital Psychiatry*, *30*(3), 269–276. https://doi.org/10.1016/j.genhosppsych.2008.01.003

Neighbors, H. W., Jackson, J. S., Campbell, L., & Williams, D. (1989). The influence of racial factors on psychiatric diagnosis: A review and suggestions for research. *Community Mental Health*, *25*(4), 301–311. https://doi.org/10.1007/BF00755677

Nelson, B., McGorry, P. D., Wichers, M., Wigman, J. T., & Hartmann, J. A. (2017). Moving from static to dynamic models of the onset of mental disorder: A review. *JAMA Psychiatry*, *74*(5), 528–534. https://doi.org/10.1001/jamapsychiatry.2017.0001

Newman, J. P. (1998). Psychopathic behavior: An information processing perspective. In D. J. Cooke, R. D. Hare, & A. Forth (Eds.), *Psychopathy: Theory, research and implications for society* (pp. 81–104). Kluwer Academic Publishers. http://doi.org/10.1007/978-94-011-3965-6_5

Nickels, M. K., & Nelson, C. E. (2005). Beware of nuts & bolts: Putting evolution into the teaching of biological classification. *The American Biology Teacher*, *67*(5), 283–289. https://doi.org/10.1662/0002-7685(2005)067[0283:BONBPE]2.0.CO;2

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*, 447–453.

Osório, F. L., Loureiro, S. R., Hallak, J. E. C., Machado-de-Sousa, J. P., Ushirohira, J. M., Baes, C. V. W., Apolinario, T. D., Donadon, M. F., Bolsoni, L. M., Guimarães, T., Fracon, V. S., Silva-Rodrigues, A. P. C., Pizeta, F. A., Souza, R. M., Sanches, R. F., dos Santos, R. G., Martin-Santos, R., & Crippa, J. A. S. (2019). Clinical validity and intrarater and test–retest reliability of the Structured Clinical Interview for DSM-5 – Clinician Version (SCID-5-CV). *Psychiatry and Clinical Neurosciences*, *73*, 754–760. https://doi.org/10.1111/pcn.12931

Parnas, J. (2014). The RDoC program: Psychiatry without psyche? *World Psychiatry*, *13*(1), 46–47. https://doi.org/10.1002/wps.20101

Pesenti-Gritti, P., Spatola, C. A. M., Fagnani, C., Ogliari, A., Patriarca, V., Stazi, M. A., & Battaglia, M. (2008). The co-occurrence between internalizing and externalizing behaviors. *European Child & Adolescent Psychiatry*, *17*(2), 82–92. https://doi.org/10.1007/s00787-007-0639-7

Petto, A. J., & Mead, L. S. (2009). Homology: Why we know a whale is not a fish. *Evolutionary Education Outreach*, *2*, 617–621.

Plomin, R., & Daniels, D. (1987). Why are children in the same family so different from each other? *Behavioral and Brain Sciences*, *10*(1), 1–16.

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson of London.

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Rogers, A. (2017, May 11). Star neuroscientist Tom Insel leaves the Google-spawned verily . . . For a startup? *Wired*. https://www.wired.com/2017/05/star-neuroscientist-tom-insel-leaves-google-spawned-verily-startup/

Ruggero, C. J., Kotov, R., Hopwood, C. J., First, M., Clark, L. A., Skodol, A. E., Mullins-Sweatt, S. N., Patrick, C. J., Bach, B., Cicero, D. C., Docherty, A., Simms, L. J., Bagby, R. M., Krueger, R. F., Callahan, J. L., Chmielewski, M., Conway, C. C., De Clercq, B., Dornbach Bender, A., . . . Zimmermann, J. (2019). Integrating the Hierarchical Taxonomy of Psychopathology (HiTOP) into clinical practice. *Journal of Consulting and Clinical Psychology*, *87*(12), 1069–1084. https://doi.org/10.1037/ccp0000452

Sagan, C. (1979). *Broca's brain, reflections on the romance of science*. Random House.

Schwartzstein, R. M. (2020). *Approach to the patient with dyspnea*. UpToDate.

Shulman, E. P., Steinberg, L. D., & Piquero, A. R. (2013). The age–crime curve in adolescence and early adulthood is not due to age differences in economic status. *Journal of Youth and Adolescence*, *42*(6), 848–860. https://doi.org/10.1007/s10964-013-9950-4

Smith, G. T., McCarthy, D. M., & Zapolski, T. C. (2009). On the value of homogeneous constructs for construct validation, theory testing, and the description of psychopathology. *Psychological Assessment*, *21*(3), 272–284. https://doi.org/10.1037/a0016699

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., Chiang, C. W. K., Hirschhorn, J. N., Daly, M. J., Patterson, N., Neale, B. M., Mathieson, I., Reich, D., & Sunyaev, S. R. (2019). Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. *eLife*, *8*, Article e39702. https://doi.org/10.7554/eLife.39702.001

Sprock, J. (2003). Dimensional versus categorical classification of proto-typic and nonprototypic cases of personality disorder. *Journal of Clinical Psychology*, *59*, 991–1014.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*, 1–25.

Szucs, D., & Ioannidis, P. A. (2020). Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *NeuroImage*, *221*, Article 117164. https://doi.org/10.1016/j.neuroimage.2020.117164

Thurstone, L. L. (1947). *Multiple factor analysis*. University of Chicago Press.

Trumbetta, S. L., Markowitz, E. M., & Gottesman, I. I. (2007). Marriage and genetic variation across the lifespan: Not a steady relationship. *Behavior Genetics*, *37*(2), 362–375.

Tung, E. S., & Brown, T. A. (2020). Distinct risk profiles in social anxiety disorder. *Clinical Psychological Science*, *8*(3), 477–490. https://doi.org/10.1177/2167702620901536

Turkheimer, E. (2012). Genome wide association studies of behavior are social science. In K. Plaisance & T. Reydon (Eds.), *Philosophy of behavioral biology. Boston studies in the philosophy of science* (Vol. 282, pp. 43–64). Springer. https://doi.org/10.1007/978-94-007-1951-4_3

Turkheimer, E. (2016). Weak genetic explanation 20 years later: Reply to Plomin et al. *Perspectives on Psychological Science*, *11*(1), 24–28.

Turkheimer, E. (2017). The hard question in psychiatric nosology. In K. S. Kendler & J. Parnas (Eds.), *Philosophical issues in psychiatry IV: Classification of psychiatric illness* (pp. 27–44). Oxford University Press. https://doi.org/10.1093/med/9780198796022.001.0001

Turkheimer, E. [@ent3c]. (2020, January 9). *Different topic: Is there a named cognitive bias describing the preference for a concrete quantitative answer to a complex question* [Tweet]. Twitter. https://twitter.com/ent3c/status/1215291870168977410

Turkheimer, E., Ford, D. C., & Oltmanns, T. F. (2008). Regional analysis of self-reported personality disorder criteria. *Journal of Personality*, *76*, 1587–1622.

Turkheimer, E., Pettersson, E., & Horn, E. E. (2014). A phenotypic null hypothesis for the genetics of personality. *Annual Review of Psychology*, *65*, 515–540.

Turkheimer, E., & Waldron, M. C. (2000). Nonshared environment: A theoretical, methodological, and quantitative review. *Psychological Bulletin*, *126*, 78–108.

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

van der Krieke, L., Jeronimus, B. F., Blaauw, F. J., Wanders, R. B., Emerencia, A. C., Schenk, H. M., Vos, S. D., Snippe, E., Wichers, M., Wigman, J. T., Bos, E. H., Wardenaar,

K. J., & Jonge, P. D. (2016). HowNutsAreTheDutch (HoeGekIsNL): A crowdsourcing study of mental symptoms and strengths. *International Journal of Methods in Psychiatric Research*, *25*(2), 123–144. https://doi.org/10.1002/mpr.1495

Varga, J. (2020). *Clinical manifestations and diagnosis of systemic sclerosis (scleroderma) in adults*. UpToDate.

Visscher, P. M., Yang, J. A., & Goddard, M. E. (2010). A commentary on 'Common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin Research and Human Genetics*, *13*, 5117–5524.

Wallace, D. J., & Gladman, D. D. (2020). *Clinical manifestations and diagnosis of systemic lupus erythematosus in adults*. UpToDate.

Waszczuk, M. A., Eaton, N. R., Krueger, R. F., Shackman, A. J., Waldman, I. D., Zald, D. H., Lahey, B. B., Patrick, C. J., Conway, C. C., Ormel, J., Hyman, S. E., Fried, E. I., Forbes, M. K., Docherty, A. R., Althoff, R. R., Bach, B., Chmielewski, M., DeYoung, C. G., Forbush, K. T., . . . Kotov, R. (2020). Redefining phenotypes to advance psychiatric genetics: Implications from hierarchical taxonomy of psychopathology. *Journal of Abnormal Psychology*, *129*(2), 143–161. https://doi.org/10.1037/abn0000486

Watts, A. L., Lane, S. P., Bonifay, W., Steinley, D., & Meyer, F. A. C. (2020). Building theories on top of, and not independent of, statistical models: The case of the p-factor. *Psychological Inquiry*, *31*(4), 310–320. https://doi.org/10.1080/1047840X.2020.1853476

Weaver, L. J., & Kaiser, B. N. (2015). Developing and testing locally derived mental health scales: Examples from North India and Haiti. *Field Methods*, *27*(2), 115–130.

Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., Fava, M., McGrath, P. J., Weissman, M., Parsey, R., Adams, P., Trombello, J. M., Cooper, C., Deldin, P., Oquendo, M. A., McInnis, M. G., Huys, Q., Bruder, G., Kurian, B. T., . . . Pizzagalli, D. A. (2019). Personalized prediction of antidepressant v. placebo response: Evidence from the EMBARC study. *Psychological Medicine*, *49*(7), 1118–1127. https://doi.org/10.1017/S0033291718001708

Wehby, G. L., & Shane, D. (2019). Genetic variation in health insurance coverage. *International Journal of Health Economics and Management*, *19*, 301–316. https://doi.org/10.1007/s10754-018-9255-y

Weinberger, D. R., & Radulescu, E. (2020). Structural magnetic resonance imaging all over again. *JAMA Psychiatry*, *78*(1), 11–12. https://doi.org/10.1001/jamapsychiatry.2020.1941

Wenger, N. K. (1990). Gender, coronary artery disease, and coronary bypass surgery. *Annals of Internal Medicine*, *112*(8), 557–558.

Wicherts, J. M., & Johnson, W. (2009). Group differences in the heritability of items and test scores. *Proceedings Biological Sciences*, *276*(1667), 2675–2683. https://doi.org/10.1098/rspb.2009.0238

Widiger, T. A., Bach, B., Chmielewski, M., Clark, L. A., DeYoung, C., Hopwood, C. J., Kotov, R., Krueger, R. F., Miller, J. D., Morey, L. C., Mullins-Sweatt, S. N., Patrick, C. J., Pincus, A. L., Samuel, D. B., Sellbom, M., South, S. C., Tackett, J. L., Watson, D., Waugh, M. H., . . . Thomas, K. M. (2019). Criterion A of the AMPD in HiTOP. *Journal of Personality Assessment*, *101*(4), 345–355. https://doi.org/10.1080/00223891.2018.1465431

Wittchen, H. U., & Beesdo-Baum, K. (2018). "Throwing out the baby with the bathwater?" Conceptual and methodological limitations of the HiTOP approach. *World Psychiatry*, *17*(3), 298–299. https://doi.org/10.1002/wps.20561

Wittchen, H. U., Beesdo-Baum, K., Gloster, A. T., Höfler, M., Klotsche, J., Lieb, R., Beauducel, A., Bühner, M., & Kessler, R. C. (2009). The structure of mental disorders re-examined: Is it developmentally stable and robust against additions? *International Journal of Methods in Psychiatric Research*, *18*, 189–203. https://doi.org/10.1002/mpr.298

Wolff, J. L., Starfield, B., & Anderson, G. (2002). Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. *Archives of Internal Medicine*, *162*(20), 2269–2276. https://doi.org/10.1001/archinte.162.20.2269

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., Frayling, T. M., Hirschhorn, J., Yang, J., & Visscher, P. M., & the GIANT Consortium. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Human Molecular Genetics*, *27*(20), 3641–3649. https://doi.org/10.1093/hmg/ddy271

Zachar, P., & Kendler, K. S. (2007). Psychiatric disorders: A conceptual analysis. *American Journal of Psychiatry*, *164*, 557–565.

Zapko-Willmes, A., & Kandler, C. (2018). Genetic variance in homophobia: Evidence from self- and peer reports. *Behavior Genetics*, *48*(1), 34–43. https://doi.org/10.1007/s10519-017-9884-9