PSYCHIATRY'S SECOND VALIDITY CRISIS:

THE PROBLEM OF DISPARATE VALIDATION

Nicholas Gaddis Zautra

Submitted to the faculty of the Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Cognitive Science

Indiana University

May 2024

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Doctoral Committee

_____

Kirk Ludwig, Ph. D. (Co-Chair)

_____

Peter Zachar, Ph. D. (Co-Chair)

_____

Dan Kennedy, Ph. D.

_____

Gary Ebbs, Ph. D.

_____

Jordi Cat, Ph. D.

April 22nd, 2024

ACKNOWLEDGMENTS

Professional Acknowledgments

I would first like to offer my utmost gratitude to my co-chairs, **Peter Zachar** and **Kirk Ludwig,** for their tremendous and continuous mentorship throughout the dissertation process. Peter has been an outstanding and resourceful guide in helping me through the "dissertation box canyon" from start to finish. His ability to intellectually inspire and encourage high level work has been so instrumental to the development of this manuscript. Kirk is someone I've known for several years to be one of the most insightful and supportive philosophers in the academy today. I am forever grateful to him in aiding a path forward and supporting my doctoral work.

The dissertation would not have been possible without the administrative and departmental labor and support of **Peter Todd**, **Fritz Breithaupt**, and **Susan Palmer** in the Cognitive Science Department. I am so appreciative for their willingness to help in any way they could to ensure my candidacy and degree requirements were met.

I would also like to thank **Dan Kennedy, Gary Ebbs**, and **Jordi Cat** for their dedicated service and support as invaluable committee members.

I would also like to thank several academic colleagues who took the time to offer counsel and guidance through helpful correspondence and conversation, including **Jason Robert, Colin Allen, Anthony Guest-Scott, Michael Weisberg**, **Archie Fields**, **Berly Brumble**, **Paul Kelly**, **Ali Mirza**, **Phillip Honenberger**, **J.J. LaTourelle**, and **Rebecca Jackson**,

Next, I would like to express my sincerest appreciation to **Evan Arnet**, **Dan Li**, and **Siyu Yao** for their insightful comments and feedback on prior drafts, as well as great friendship during our weekly works-in-progress Zoom meetings.

Lastly, I would like to thank those who sat down for informational interviews on the concepts of validity and validation and who helped to shape much of this work, including **Kathleen Slaney**, **Gregory Cizek**, **Awais Aftab**, **Colin DeYoung**, **Sarah Morris**, **Eiko Fried**, **Richard McNally**, **Erik Turkheimer**, **Matcheri Keshavan**, **Payton Jones**, and **Laura Bringmann**.

Personal Acknowledgments

I would like to thank my father, **Alex Zautra**, and my mother, **Ann Gaddis**, for encouraging and inspiring me in their own individual ways to think creatively and to cultivate an attitude toward lifelong learning.

Most importantly, I would like to thank my incredibly loving and supportive wife, **Nicole "Duckie" Zautra**, my wonderful little boy, **Maxwell "Max" Zautra**, and my two adorable corgis **Moose** and **Minnie**. I am so grateful for our little family.

Nicholas Zautra

PSYCHIATRY'S SECOND VALIDITY CRISIS: THE PROBLEM OF DISPARATE VALIDATION

In response to the crisis of confidence in the validity of the *DSM*'s diagnostic categories, psychiatry has seen a proliferation of alternative research frameworks for studying and classifying psychiatric disorders in new ways. The big three alternative approaches—the Hierarchical Taxonomy of Psychopathology (HiTOP), the Network Approach to Psychopathology, and the Research Domain Criteria (RDoC)—have been characterized as healthy responses to the *DSM*'s crisis of validity.

A yet unexplored aspect of psychiatry's validity crisis is related to disagreements regarding the standards of validity. Each of the approaches have multiple distinct senses of validity, which point to a thornier methodological problem for psychiatry that I term the problem of "disparate validation." This two-part problem can be summarized as follows: scientific psychiatry aims at achieving empirically informed classifications that demonstrate validity in the sense that they correspond to real attributes of psychopathology. To achieve this, alternative research frameworks are now approaching the conceptualization, testing, organizing, and validation of different features of psychopathology by their own standards in the hopes of one day informing more valid systems of psychiatric classification. The first problem is given a system of classification, by whose standard of validity should such a system be validated? Is there a single validation procedure by which validation should proceed, or some other combination thereof? Second, when we attempt to validate classifications informed by differing standards of validity, will any such validation be capable of assessing a unified fundamental

sense of validity that exists across the various frameworks, or will they only be valid in their own narrow sense?

In this dissertation, I offer an assessment of the problem of disparate validation through faithful reconstructions of the Holy Quadrinity of distinct senses of validity in psychiatry: starting with diagnostic validity (*DSM*) and proceeding with psychometric validity (HiTOP), network psychometric validity (the Network Approach), and etio-pathophysiological validity (RDoC). I introduce commonalities across frameworks that have not been previously addressed, including how each framework employs expert curation, being the selection and justification of certain elements into their model based on compromises, and how the goal of each framework eventually becomes a return to the original validators of Robins and Guze to evaluate prognosis, biomarkers, and etiology of psychiatric classifications.

By evaluating psychiatry's distinct senses of validity, I argue that despite an appearance of a shared goal of informing more valid classifications, the existence of multiple frameworks in which each employs their own standards of validity and validation is a detrimental methodology to achieve any kind of unified validation work. At its core, fundamental disagreements concerning 1) the underlying phenomenon that researchers are attempting to make inferences about; 2) the sources of validating evidence; and 3) the very nature of validity and validation, move each framework further and further toward a state of unrecognized plurality, in which these frameworks are really not at all talking about the same thing and are in fact engaged in different projects with different aims. I conclude with a positive program that suggests in what ways such different frameworks with distinct validation procedures can achieve validity in their own specific sense while also coming to inform one another through a kind of complementary pluralism.

Table of Contents

Chapter 1: Validity and Psychiatry, and the Problem of Disparate Validation

1.1 Psychiatry's Validity Crisis

When I was a psychology undergraduate, my professor for Psychological Research Methods introduced the 4th Edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV*) with a word of caution, saying that the *DSM* was "just for insurance purposes." I recall finding this characterization of psychiatry's flagship classification system surprising, but one that made practical sense. On the one hand, I assumed that a central aim of the *DSM* was to represent psychiatry's best scientific understanding of what it took to be real underlying mental illnesses. If a diagnostic category like Major Depressive Disorder had been selected for inclusion in the evidence-based versions of the *DSM* (the *DSM-III* and beyond), it was reasonable to suppose that not only was there sufficient theory and evidence to warrant its inclusion but that such evidence supported the notion that Major Depressive Disorder really was a thing in the world. Given what was perceived as growing support relating the *DSM*'s diagnostic categories to specific changes in brain chemical messengers called neurotransmitters like serotonin (Major Depression, Generalized Anxiety Disorder) and dopamine (Schizophrenia, Bipolar Disorder), it appeared that the *DSM* must be getting at some truth about underlying mental disorders. So, thinking of the *DSM* as merely an instrument for medical billing seemed perplexing.

On the other hand, I knew the *DSM* had many practical aims as well: to serve as a reliable and useful classification system for clinicians to effectively diagnose and treat patients in an applied setting; to substantiate claims for medical reimbursement; to address stigmatization concerns; and to provide interpretable and less-exploitable psychiatric classifications for use in

forensic and legal settings. Thus, the diagnostic criteria of *DSM* categories, with their broad symptom-based clinical descriptions, was not the scientific end-all-be-all for psychiatric disorders. Additionally, I intuited that constructing scientific classifications that accurately represent the essence and complexity of mental illnesses was incredibly challenging, so much so that our best scientifically informed estimates were likely to be wrong to some degree. So, the idea of the *DSM* as being "just for insurance purposes," which I more charitably interpreted to mean "not necessarily representing our best scientific understanding of mental disorders," made sense. Still, I assumed like many of the public that psychiatry was making progress toward accurately describing and explaining these disorders, and that progress would eventually come to work its way into the *DSM*.

But as I now near the completion of my graduate training, and after having the opportunity to study and work alongside various researchers in psychopathology, I've come to observe an even stronger pessimism regarding the *DSM*'s ability to progress toward an accurate scientific nosology. Whereas the *DSM* was once treated like a bible, today's researchers might not even own a copy.

Persistent doubts concerning the scientific status of the *DSM* and subsequent responses to adopt alternative research approaches are consequences of psychiatry's longstanding validity problem. Philosopher of science Miriam Solomon (2022) refers to this problem, which has persisted since the inception of the *DSM-III* in 1980 as psychiatry's crisis in validity. More specifically, psychiatrist James Philips (2013) characterized the problem as a "crisis of confidence in the validity of our psychiatric diagnoses" (p. 143). Validity in different contexts and in different senses may refer to a variety of different things. In reference to psychiatry's crisis in validity, validity in the most fundamental sense—a notion of "big-V" validity as defined

by Zachar (2012)— is thought of as the degree to which psychiatry's diagnostic categories accurately reflect real underlying clinical syndromes. In other words, the more validity our diagnostic categories have, the more confidence we have that they refer to some real, underlying condition.

Significant evidence suggests our diagnostic categories are lacking in validity in this fundamental sense. First, a lack of validity is attributed to poor performance on the *DSM*'s *validators*, defined as sources of evidence first instantiated in Robins and Guze's (1970) method as a "five-phase" approach for achieving diagnostic validity. While their method has been expanded over the years (Kendler, 1980; Kendler, 1990; Andreasen, 1995; Kendler, 2009); the core of the procedure utilized by the *DSM* still amounts to a systematic process for evaluating a proposed clinical syndrome against various sources of evidence that are considered to provide validation of that syndrome. Robins and Guze's expectation of achieving what they referred to as *diagnostic validity* is very much in service of what psychologist Richard Bentall (2003) characterized as psychiatrist Emil Kraeplin's (1856–1926) big idea: that psychiatric disorders are discrete disease entities that can be accurately identified through clinical observation of their signs and symptoms, direct observation of their pathological anatomy, or study of their etiology (e.g., whether the disorder runs in families). Kraeplin's approach to identifying psychiatric disease entities attempted to mirror that of medicine, in which he posited that "judging from our experience in internal medicine it is a fair assumption that similar disease processes will produce identical symptom pictures, identical pathological anatomy and an identical aetiology" (Kraepelin, quoted in Bentall, 2003, p. 12). Thus, the big idea of Kraepelin, which would be adopted by the neo-Kraepelinians Robins and Guze along with John Feighner and George Winokur of the Washington University group, was to assume that each psychiatric disorder

possessed a typical symptomology that directly corresponded with a particular etiology and underlying brain pathology (Figure 1, Appendix). Under this assumption, the first step toward identifying discrete disease entities and thus valid classifications of psychiatric disorders was in the accurate clinical description and grouping of symptoms. A classification based first on symptoms could then act as a kind of Rosetta Stone on which the biological underpinnings of those symptoms (etiology and pathophysiology) could be discovered and expected to converge (Bentall, 2003). This convergence across sources of evidence of a symptom-based classification would come to be interpreted as validation of the classification as standing for a real discrete psychiatric disorder.

Unfortunately for the *DSM*'s symptom-based diagnostic categories, Kraepelin's big idea of discoverable discrete psychiatric disease entities, presently defined by the *DSM* as clinical syndromes, has yet to be realized. For example, consider three of Robins and Guze's five original standards of validation: clinical description (assessment of symptoms), laboratory studies (search for biomarkers), and delimitation from other disorders. First, in the clinical description, diagnostic categories are supposed to reflect a collection of symptoms aggregated into discrete syndromes, yet research into the *DSM* points toward a vast heterogeneity. For example, depression shows up in *many* different ways, such that most patients don't even fit into a diagnostic category's specific criteria. Second, in laboratory studies, despite an emphasis toward coming to feature the biological basis of psychiatric disorders in the *DSM*, a single clear biomarker has been added to only *one* out of nearly 300 hundred *DSM* diagnostic categories—namely, an abnormally low concentration of orexin-A (hypocretin-1) in cerebrospinal fluid (CSAF), which is indicative of Narcolepsy type 1 as a part of the *DSM*'s Sleep-Wake Disorders grouping. Third, in delimitations from other disorders, rates of comorbidity are so high that

researchers are questioning whether these rates are the result of an invalid classification rather than a feature of psychopathology.

There are three primary explanations for why symptom-based diagnostic categories perform so poorly on the validators. The first centers on the idea that the underlying hypothesis regarding psychiatric disorders as discrete disease entities as envisioned by Kraepelin and maintained by the *DSM* is just flat wrong. Thus, creating symptom-based classifications in this image and expecting disparate sources of evidence to converge on a single account of a disorder as in other areas of medicine will ultimately not pan out. The second explanation views the *DSM* in its present and foreseeable state as attempting to serve too many purposes in its competing scientific, professional, and practical aims. Even if discrete disease entities *do* exist, the *DSM*'s more expert-based than evidence-based curation process stands in the way. This position is summarized succinctly by former Editor-in-Chief of the *New England Journal of Medicine* Maria Angell:

> Given its importance, you might think that the *DSM* represents the authoritative distillation of a large body of scientific evidence. It is instead the product of a complex of academic politics, personal ambition, ideology and, perhaps most important, the influence of the pharmaceutical industry. What the *DSM* lacks is evidence. (Maria Angell, quoted in Lynch, 2018, p. 6).

The third explanation, and the one that motivates the focus of this dissertation, is that our current guidelines for developing, evaluating, and revising psychiatric classifications need an overhaul. In response to the crisis of validity, psychiatry over the past twenty years has seen a proliferation of alternative research frameworks for studying and ultimately classifying psychiatric disorders in new ways. The big three alternative approaches—the Hierarchical Taxonomy of Psychopathology (HiTOP), the Network Approach to Psychopathology, and the Research Domain Criteria (RDoC)—have been characterized by Solomon (2022) as a "healthy

response" to the *DSM*'s crisis of validity. By approaching psychiatry's validity problem in ways that are unbound by the constraints of the *DSM*, the hope is this will bolster the validity of future psychiatric classifications, resolving psychiatry's longstanding validity problem. While the alternative frameworks represent very big changes in approach and orientation to psychiatric research and classification, the *DSM* itself is simultaneously going through its own iterative revision that may address the current concerns regarding validity of its diagnostic categories and contribute to a more evidence-based scientific nosology.

Despite how the crisis of validity is typically presented, there is reason to suspect that the problem of validity isn't solely based on the inability of *DSM*'s diagnostic categories to perform on the validators. A yet unexplored aspect of psychiatry's validity crisis is related to disagreements regarding the standards of validity one adopts to validate psychiatric disorders in the hopes our psychiatric classification may one day achieve validity. One clear example of this is a disagreement between frameworks in evaluating the degree to which the *DSM*'s diagnostic categories lack validity in this sense and why. *DSM* proponents will argue that while poor performance on what are described as concurrent validators (as evidenced by a lack of biological markers such as genes or neural substrates) for *DSM*-based diagnostic categories exists, performance on certain predictive validators (e.g., differential response to treatment or diagnostic stability) justifies that the categories do have some degree of validity. In turn, proponents of the HiTOP framework believe the *DSM*'s symptoms-first approach to be fundamentally misguided and lacking in structural validity. Since the categories themselves do not reflect pure, dimensional constructs derived from empirically based, quantitative methods, not only do they not hold validity, but they should not be considered reliable, which HiTOP understands as necessarily a part of validity. The Network Approach thinks of mental disorders as "systems, not

syndromes" in which mental disorders stem from complex, dynamic interactions between symptoms as opposed to arising from underlying "common causes." Under this approach, the *DSM*'s diagnostic categories do hold some reliability and clinical utility, but they cannot be tested in the way the Network Approach deems critical, and thus ultimately should not be interpreted as having validity. For RDoC, the *DSM*'s failure to locate biological underpinnings in its categories is the primary reason the National Institute of Mental Health Director Thomas Insel stated at a 2005 meeting of the American Psychiatric Association that the *DSM* has "0% validity" (Lynch, 2018, p. 5). Lack of validity in this approach is not an objective result of a failure of performance on the validators but is in part due to disagreements on the interpretation and meaning of the evidence.

The disagreements on standards that led to multiple distinct concepts of validity point to a more challenging methodological problem for psychiatry that I term "disparate validation." This two-part problem can be summarized as follows: scientific psychiatry aims at achieving empirically informed classifications that demonstrate validity in the sense that they correspond to real attributes or features of psychopathology. To achieve this, alternative research frameworks are now approaching the conceptualization, testing, organizing, and validation of different features of psychopathology by their own standards in the hopes of one day informing more valid systems of psychiatric classification. The first problem is, given a system of classification, by whose standard of validity should such a system be validated? Is there a single validation procedure, e.g., Robin's and Guze's method for achieving diagnostic validity by which validation should proceed, or some other combination thereof? Second, when we attempt to validate systems of classifications informed by different frameworks constructed from differing standards of validity, will any such validation process be capable of assessing a unified

fundamental sense of "validity" that exists across the various frameworks, or are the frameworks ultimately incompatible such that the classifications they produce will only be valid in their own narrow sense?

Before going further into the implications of disagreements over standards of validity in psychiatry, let me introduce the concept and purpose of validity more generally from a distinct but related psychological context in which it is traditionally discussed.

1.2 What Is Validity, and Why Should We Care?

Validity is often conceived as "the most fundamental consideration" in developing and evaluating measurement instruments referred to as tests. A test is defined by Cizek (2020) as "a sample of information about some intended characteristic of persons that is gathered under specified conditions" (p. 2). Tests can take many formats, including surveys, questionnaires, and scales—essentially any systematized collection of a sample (e.g., a single one-item question) may be thought of as a test. Tests are developed and administered by the "tester" to the "test taker" to obtain information regarding some underlying characteristic "that cannot be directly observed but which leave indications of their presence or magnitude in situations designed to elicit them" (Cizek, 2020, p. 10). Such underlying characteristics are typically referred to as constructs, understood as individual differences in the underlying psychological attributes of interest. Examples of constructs in educational and psychological measurement include math problem-solving ability, teamwork, introversion, clinical competence, and anxiety, among many others. The responses and scores on a test designed to detect variations in a construct are not thought to be conclusive, as any test only provides a limiting sampling of information. Instead, test scores are understood as supporting inferences, which are acts of interpreting the observed

test scores and forming a conclusion about where the test taker stands in relation to the underlying construct. The received definitions of validity and validation, as interpreted by Cizek (2020) are as follows:

> **Validity:** the degree to which scores on an appropriately administered test support inferences about variation in the construct that the instrument was developed to measure.
> **Validation:** the ongoing process of gathering, summarizing, and evaluating relevant evidence concerning the degree to which that evidence supports the intended meaning of scores yielded by an instrument and inferences about standing on the characteristic it was designed to measure. (Cizek, 2020, pp. 29-31)

There are two critically important reasons that validity, a technical measurement concept referring to the degree to which test scores of our measurement instruments mean what they are intended to mean, became the most fundamental consideration for psychological testing. The primary reason is consequential. Several decisions significant to both individual and organizational stakeholders are based on test scores, which, while not being the only source of consideration, are often considered to be of the most critical, and thus, should be of the highest quality possible (Cizek, 2020). For example, an individual may only gain admission to a university subject to sufficiently high standardized test scores, accreditation or licensure by professional programs based on an entrance exam, eligibility for medical treatment based on a diagnostic exam, access to job opportunities based industrial and organization personality and skills assessment, the ability to legally operate a motor vehicle based on a driver's license test, or the ability to immigrate to the United States based on the US Citizenship test. At an organizational level, a policymaker might decide to reform programs based on perceived effectiveness, or to supply or cut funding to a particular school system based on standardized test scores. The second and less emphasized reason is epistemic. While the purpose of validity as a

measurement concept is generally understood for making practical decisions and thus places an emphasis on consequences, validity continues to be applied and co-opted in scientific practice toward instruments utilized in scientific investigation by social and behavioral scientists for the sole purpose of basic empirical research. In these contexts, psychological instruments may not be validated for some practical use but instead validated to emphasize best research practices that contribute to the knowledge-building process about underlying constructs (e.g., an anxiety scale used by researchers for research on anxiety). Selection and justification for the use of one instrument over another in a research setting is often based predominantly on that instrument's perceived status of validity. Researchers may further maintain a higher degree of confidence in the evidence procured from instruments that they have appropriately deemed to demonstrate high degrees of validity.

Given the consequential nature of testing and a perceived need to ensure that validity is reliably and properly assessed, professional standards for validity and validation appear in an authoritative publication called the *Standards for Educational and Psychological Testing*. The purpose of the *Standards* is to serve as a guide of best practices for developing and evaluating educational and psychological measurement instruments. Since 1954, a new edition has been published every ten to twelve years by the American Psychological Association (APA), most recently in 2014. Since 1999, the *Standards* has been jointly sponsored by the American Educational Research Association (AERA) and National Council on Measurement and Education. Each edition is intended to reflect general areas of consensus within contemporary scholarship on validity. *Standards* is a widely accepted and drawn upon guide in various professional and research settings, including industrial and organizational psychology,

standardized professional examinations and licensures, and broadly within basic research within the social and behavioral sciences, including clinical psychology and psychiatry.

The current conception of validity as featured in the *Standards* represents a unified sense of validity, such that there aren't different types of validity but instead different sources of validating evidence that all contribute to a single coherent account of validity. The most recent edition of *Standards* (2014) represents the current thinking of contemporary validity theory, including general principles on which validity theorists tend to agree, along with professional standards for drawing upon and evaluating various sources of validating evidence from which testers and social and behavioral scientists may utilize. In terms of general principles, Cizek (2020) identifies what he refers to as "six foundational tenets of contemporary validity theory":

1. Validity pertains to test score inferences.
2. Validity is not a characteristic of an instrument.
3. Validity is a unitary concept.
4. Validity is a matter of degree.
5. Validation involves gathering and evaluating evidence bearing on intended test score inferences.
6. Validation is an ongoing endeavor. (p. 37)

All validity evidence is integrated into a single, unified sense of validity for psychological tests in educational and psychological measurement, referred to as construct validity. Construct validity falls within construct validity theory (CVT), defined by Slaney (2017) as the "general theoretical approach and set of methods for judging whether empirical inferences and decisions made based on quantitative data are licensed by the most current theory regarding the construct purportedly measured by the test or assessment tool in question" (p. 1).

1.3 Validity in Psychiatry Is Not Construct Validity (For the Most Part)

Returning to the concept of validity in psychiatry, we may ask: just how and in what way do validity and validation as they are featured in educational and psychological measurement (i.e., as construct validity) relate, if at all, to validity in psychiatry? A motivation for this inquiry is the fact that when scholars in psychiatry and the philosophy of psychiatry characterize psychiatry's crisis in validity, they may emphasize that the primary sense of validity that *DSM*'s diagnostic categories lack is that of "construct validity":

> Most *DSM* diagnostic categories do not have *construct validity*, that is, they do not "carve nature at the joints" by picking out just one kind of condition with a distinctive etiology. Rather, current categories are syndromes that encompass many different etiologies. In the long run, the goal of diagnostic research is construct validity because that yields the most insight and the most chance for developing novel and carefully targeted empirically supported treatments. (Wakefield, 2013, p. 826)

> There is scarce evidence that any *DSM* diagnostic categories—other than a small handful (viz., "schizophrenia," "bipolar disorder," "intellectual disability," "neurocognitive disorders")—possess construct validity. To have construct validity, a diagnostic category should accurately represent a construct as defined by theory. (Tsou, 2021, p. 69)

There are several instances of "construct validity," a property reserved for the evaluation of a measurement instrument, representing this fundamental sense of validity that our diagnostic categories are lacking. For one, there have been attempts to apply concepts from validity in psychological measurement to Robins and Guze's method for achieving diagnostic validity. In an article by Cloninger (1989), which includes a brief interview with Robins (who ultimately doesn't comment on the matter), Cloninger relates Robins and Guze's (1970) initial five phases of validating evidence to various "types" of validity in psychological measurement (Table 1, Appendix), arguing that Robins and Guze's method was "consistent with" validity and validation in psychological measurement.

Some researchers have gone a step further to interpret diagnostic validity not only as being consistent with psychometric validity but essentially being in service of establishing construct validity. For example, psychologist Catherina Hartman and colleagues (2001) claimed that "the hallmark of construct validity is external construct validity…through differential relations of current clinical concepts with aetiology, course, prognosis, or dysregulations in the neurobiological or cognitive system," suggesting the five phases contribute to a unified sense of construct validity. In the article "The Validity of Psychiatric Diagnoses Revisited," Aboraya et al. (2005) claim that "Robins and Guze actually were the first to articulate the elements of construct validity in psychiatry" (p. 50), and that "construct validity, consisting of validity criteria, is the core of psychiatry" (p. 55). Psychologists Mullins-Sweatt and Widiger (2009) claimed that "the authoritative statement on construct validity for psychiatric diagnoses was provided by Robins and Guze (1970), in which they compare "the conceptualization of a disorder…to the theoretical articulation of a construct (Cronbach & Meehl, 1955; Smith, 2005) and its delimitation from other disorders (E. Robins & Guze, 1970)" and how "predicting a future course concerns the validation of theoretically derived hypotheses concerning the construct (Cronbach & Meehl, 1955; Smith, 2005) through follow-up studies (E. Robins & Guze, 1970)" (p. 303). Philosopher of science Kenneth Shaffner, who has provided very thoughtful work in the philosophy of psychiatry and validity in psychiatry, has also drawn a close connection between diagnostic validity and construct validity:

> For our purposes, the notion of "diagnostic validity" is of special importance. This concept comes from Robins and Guze's classic and extraordinary influential 1970 article noted earlier. In a way, this article adapted the construct validity notion to psychiatric diagnosis by using the term "diagnostic validity" (Robins and Guze, 1970), though there is no reference to the term "construct validity" nor to Cronbach and Meehl's (1955) article in their 1970 paper. (Shaffner, 2012, p. 169)

While these interpretations of relating construct validity as either being consistent with or in some sense being the basis of validity in psychiatry may appear feasible, such interpretations are ultimately inaccurate. The *DSM*'s standards of validity, i.e., the process of establishing diagnostic validity based on an expansion of Robins and Guze's method, have essentially remained consistent in establishing a sense of validity that is distinct from the concepts of validity and validation in psychological measurement, i.e., construct validity. Two examples demonstrate this distinction. First, in an article by personality psychologists John Livesly and Douglas Jackson (1992) titled "Guidelines for Developing, Evaluating, and Revising the Classification of Personality Disorders," the authors proposed a novel validation system for the *DSM*'s personality disorders and, by extension, other forms of psychopathology for future editions of the *DSM* (*DSM-5*). Their proposal amounted to constructing and evaluating the *DSM*'s diagnostic categories in such a way that validating evidence from psychological measurement, such as content, criterion, predictive, convergent, and divergent sources of validity evidence would be the new required basis for establishing the validity of a diagnostic category. Their proposal, which was published just after the release of the *DSM-IV*, was motivated by the fact that up to and through the development of the *DSM-IV*, such a validation process had never been done before, and was intentionally presented in contrast to the standard that was Robins and Guze's method. Second, in the chapter "Five Criteria for an Improved Taxonomy of Mental Disorders" in *Defining Psychopathology in the 21st Century: DSM-V and Beyond*, Dr. Robert Kendell, one of the leading authorities on psychiatric classification in the *DSM*, clarifies the distinction between validity and validation of the *DSM*, as well as validity concepts typically associated with construct validity, which would ultimately not come to be adopted by the *DSM*:

> Psychologists are accustomed to distinguishing several different kinds of validity—construct, concurrent, content, predictive, and so on. Although these are useful distinctions in many settings, in the context of clinical medicine statements about diagnostic validity are essentially statements about predictive power and hence practical utility. The more information a diagnosis provides about outcome and response to treatment—and thus about which treatments are appropriate—the higher its validity and the greater its utility. (Kendell, 2002, p. 7)

These examples demonstrate that it is simply not the case that validity (or a lack thereof) in the *DSM* is, at its core, construct validity.

Others have suggested that the crisis in validity is an issue of the absence of construct validity. First, construct validity is just not what the validation of *DSM*'s diagnostic categories is based on. Second, construct validity in its current form follows the work of Samuel Messick (1984) and has little to do with the testing of hypotheses or the theoretical articulations of a construct. Interpretations comparing Robins and Guze's method with construct validity seem to focus on what amounts to selective interpretations of Cronbach and Meehl's (1955) "strong" version of construct validity which centered on theoretical constructs defined exclusively in terms of formal theories. Thus, the common notion that the *DSM*'s diagnostic categories are (unsuccessful) attempts to "carve nature at its joints" is one most appropriately reserved for diagnostic validity, in which validity in this fundamental sense is thought to depend on "drawing boundaries between adjacent syndromes, and between these syndromes and normality, where there are genuine discontinuities either in symptomology or in etiology" (Kendell, 2002, p. 7). The fact that there is little if any acknowledgment in psychology or the philosophy of science that construct validity is not the same sense of validity as diagnostic validity is an example of *unrecognized plurality* in scientific psychiatry. Given the complexities of validity and validation as well as a tendency to put forth underspecified or inconsistent conceptions of validity, we have

yet to realize the extent to which these disparate validation procedures support very distinct senses of validity. I will return to this notion in the concluding chapter.

While something like the diagnostic validity of the *DSM* upon closer examination may be ultimately distinguished from that of construct validity, there are aspects in which validity in psychiatry is reflective of certain aspects of CVT that warrant further consideration. First, there does appear to be some overlap in general principles of validity and validation in psychiatry with those in psychological measurement. Conceiving of validity as a matter of degree, as a unitary concept that is supported by various sources of validity evidence, and that validation is considered an ongoing endeavor are emphasized in both guides for evaluating classifications of the *DSM* and in newer versions of the *Standards* (2014), which suggests that certain surface level standards of validity associated with more modern validity theory are currently being adopted by those working to articulate validity and validation in psychiatry. Furthermore, recent attempts to make the *DSM*'s process of iterative revision more scientific, while still being based in diagnostic validity, do appear in some ways to model the more ambitious goals of the more theory-driven conceptions of CVT. Second, clinical instruments that are utilized in psychopathology research and/or clinical practice such as trait measures, inventories, and self-report scales are all subjected to the validity and validation criteria of psychological measurement to assess their construct validity. In 1998, Richter et al. reviewed the content, factorial, convergent, and discriminant validity of the Beck Depression Inventory, a depression self-report scale, and found that it had been employed in over 2,000 studies.

The more important aspect to consider in relation to CVT, however, is the proliferation of alternative frameworks in psychiatry outside of the *DSM* and diagnostic validity. Such frameworks adopt very different models of validity and validation for their initial research

process, especially those like the alternative frameworks of HiTOP and RDoC, which focus on the development of constructs. As such, construct validity is likely to feature to some degree in these new accounts. Understanding to what degree construct validity is reflected in each of the frameworks, and in what sense, will be useful in distinguishing between different frameworks and getting to the heart of the problem of disparate validation, to which we now return.

1.4 Overview of the Dissertation

The brief detours into the concepts of validity and validation in psychological measurement and misinterpretations of construct validity onto diagnostic validity provides an important normative lesson: the concept of validity is complicated, be it for the experts in modern validity theory who contribute to the *Standards*, or for the psychiatrists and psychologists taking concepts and applying them to psychiatry. It is very easy to engage in validation work without really having a sense of what validation amounts to and, as Slaney (2017) has observed in her study of validity practices in experimental psychology, it is even easier to apply validity concepts and procedures in inconsistent and illogical ways. As scientific psychiatry is now faced with not one but several senses of validity in psychiatric research in what I argue amounts to an even more daunting validity problem, coming to a more precise understanding regarding psychiatry's different methods, procedures, and standards of validity is paramount.

To address psychiatry's problem of disparate validation, this dissertation has two primary aims. The first aim will be to offer a diagnostic assessment of the problem through faithful reconstructions of what I call the "Holy Quadrinity" of distinct senses of validity in psychiatry: starting with diagnostic validity (*DSM*) and proceeding with

structure-first psychometric validity (HiTOP), network psychometric validity (the Network Approach), and etio-pathophysiological validity (RDoC). With each reconstruction, I provide a detailed account of the distinctive sense (or senses) of validity associated with each and explain why advocates consider their validity account to provide the best standards for informing valid classifications in psychiatry. I trace the historical development of each sense of validity and interpret how each may be reflective of different interpretations of CVT or other meanings of validity from educational and psychological measurement. I describe what each account takes to mean by the concepts of "validity," "validation," "and "construct," and how they view them in relation to validation concepts such as "reliability" and "utility." Lastly, I describe the underlying philosophical commitments of each validation process as they relate to various scientific realist positions and evaluate whether these positions are consistent with the overarching framework in which they feature.

In chapter 2, I begin by reconstructing the *DSM*'s newly updated empirically grounded plan for achieving validity in psychiatry based in diagnostic validity, the *DSM*'s particular sense of validity and validation. I provide a general overview of diagnostic validity and explicate its historical development from its initial development by the neo-Kraepelinians in 1970 and through the various iterations of the *DSM*. I then analyze its current relation to CVT and conclude with a discussion of its hybrid set of philosophical foundations intended to support the *DSM*'s two distinct aims of prediction and representation. In doing so, I aim to provide an anchor by which other senses of validity in scientific psychiatry described in chapters 3–5 may be distinguished.

In chapter 3, I introduce the Hierarchical Taxonomy of Psychopathology (HiTOP), a data-driven approach to psychopathology and its distinct sense of validity based in *psychometric*

validity, a type of validity associated with the validation of psychological instruments or tests designed to measure psychological constructs. I connect HiTOP's sense of psychometric validity to psychometrics research based in factor analysis in the early 20th century, the formation and development of construct validity theory (CVT), and quantitative approaches to personality research in psychology. I describe some refinements to HiTOP's revision process, distinguish this refined sense from specific senses of CVT within psychometric validity, and explain its scientific realist positions. In doing so, I show how HiTOP's validation process doesn't just offer a different path of achieving diagnostic validity, but instead offers a distinct conception as to what should count as validity in psychiatry altogether based in a hybrid set of its psychometric validation standards.

In chapter 4, I turn to the Network Approach to Psychopathology, a theory-driven approach that rejects the common cause models of the *DSM* and HiTOP and instead draws on a systems model that posits that mental disorders are like dynamic complex systems. I characterize the network approach's sense of validity, network psychometric validity, as being in a form of pre-validation, meaning a stage in which there is still ongoing discussion as to what it is that requires validation and how it should take place. I then turn toward motivations for the Network Approach based on the shortcomings of the use of data models, its development with the use of validity-adjacent concepts such as testability and falsifiability, and detail how in contrast to other approaches, seeks to achieve validity far more implicitly than other approaches by developing standards for theory development and theory construction.

In chapter 5, I focus on the Research Domain Criteria (RDoC), a biologically-driven research framework that is based in etio-pathophysiological validity. I trace the development of RDoC's etio-pathophysiological validity beginning from prior cognitive neuroscience-based

initiatives that informed RDoC's development through to its most recent changes with "RDoC 2.0," an unofficial term used internally among RDoC Unit members to denote RDoC's recent shift in approach. I argue that while RDoC has long hailed itself as an integrative approach in that it combines multiple scientific disciplines, what is truly integrative about RDoC is its attempt to rebrand itself as a research framework that is open to integrating all of the various senses of validity in psychiatry, including broader understandings of validity as a desirable quality or feature, and a willingness to allow researchers to employ their own standards of validity based on their specific aims and stages in their research.

In chapter 6, I turn toward the second goal of the dissertation, which will be to provide an overarching analysis of the Holy Quadrinity in terms of commonalities and differences across frameworks that have not been previously addressed to offer a positive program to the problem of disparate validation.

The first commonality I will introduce is that despite alternative approaches adopting new validation concepts, principles, and procedures for re-organizing or re-clustering psychopathology in their vision, the goal of each framework eventually becomes a return to the original validators of Robins and Guze to evaluate things like prognosis, biomarkers, and etiology of the newly formed classifications. A second commonality is that each framework employs expert curation—the selection and justification of certain elements into their model based on compromises, which contrasts with the notion that such features in the data-driven alternative frameworks are only selected based on evidence. A third commonality is that while validity is elaborately presented as empirically or scientifically based and is accompanied by long lists of validators or intricate sequencing of how evidence should be evaluated, validity for

each approach in the broadest of senses boils down to simply mean that which is considered to be good or desirable for that approach.

Next, I turn to analyze the differences that are, at their core, fundamental disagreements concerning 1) the underlying phenomenon that researchers are attempting to make inferences about; 2) the sources of validating evidence; and 3) the very nature of validity and validation. Such differences, I will argue, impede the capacity for progress in informing and developing valid psychiatry classifications by creating difficulties in evaluating between frameworks, coordinating and integrating between frameworks, and ultimately in developing any single unified sense of validity for psychiatry.

By thoroughly assessing and evaluating psychiatry's Holy Quadrinity of distinct senses of validity, I will argue that despite the appearance of a shared goal of informing more valid classifications, the existence of multiple frameworks in which each employs their own standards of validity and validation is ultimately a troubling situation methodologically for trying to do any kind of unified validation work. The opposing standards of validity amount to a state of non-complementary unrecognized plurality. Just as we had failed to recognize that diagnostic validity is not construct validity, we also have yet to fully realize to what extent these additional frameworks are really not at all talking about the same thing and are engaged in different projects with different aims.

I will conclude with a positive program that suggests in what ways such different frameworks with distinct validation procedures can achieve validity in their own specific sense while also offering suggestions as to how the approaches may eventually come to inform one another through a kind of complementary pluralism.

References for Chapter 1

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andreasen, N. C. (1995). The validation of psychiatric diagnosis: New models and approaches. *The American journal of psychiatry*, *152*(2), 161–162.

Aboraya, A., France, C., Young, J., Curci, K., & LePage, J. (2005). The validity of psychiatric diagnosis revisited: The clinician's guide to improve the validity of psychiatric diagnosis. *Psychiatry (Edgmont)*, *2*(9), 48.

Bentall, R. P. (2003). *Madness explained: Psychosis and human nature*. Penguin UK.

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.

Cloninger, C. R. (1989). Establishment of diagnostic validity in psychiatric illness: Robins and Guze's method revisited. *The validity of psychiatric diagnosis*, 9-18.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

First, M. B., Regier, D. A., & Kupfer, D. J. (2002). *A research agenda for DSM-V*. American Psychiatric Pub.

Hartman, C. A., Hox, J., Mellenbergh, G. J., Boyle, M. H., Offord, D. R., Racine, Y., ... & Sergeant, J. A. (2001). *DSM-IV* internal construct validity: When a taxonomy meets data. *The journal of child psychology and psychiatry and allied disciplines*, *42*(6), 817–836.

Kendell, R. E. (2002). Five criteria for an improved taxonomy of mental disorders. In J. E. Helzer & J. J. Hudziak (Eds.), *Defining psychopathology in the 21st century: DSM-V and beyond*, (pp. 3–17). American Psychiatric Publishing, Inc.

Kendler, K. S. (1980). The nosologic validity of paranoia (simple delusional disorder): A review. *Archives of general psychiatry*, *37*(6), 699–706.

Kendler, K. S. (1990). Toward a scientific psychiatric nosology: Strengths and limitations. *Archives of general psychiatry*, *47*(10), 969–973.

Kendler, K., Kupfer, D., Narrow, W., Phillips, K., & Fawcett, J. (2009). Guidelines for making changes to *DSM-V*. Unpublished manuscript.

Livesley, W. J., & Jackson, D. N. (1992). Guidelines for developing, evaluating, and revising the classification of personality disorders. *The journal of nervous and mental disease*, *180*(10), 609–618.

Lynch, T. (2018). The validity of the *DSM*: An overview. *The Irish journal of counselling and psychotherapy*, *18*(2), 5–10.

Mullins-Sweatt, S. N., & Widiger, T. A. (2009). Clinical utility and *DSM-V*. *Psychological assessment*, *21*(3), 302.

Phillips, J. (2013). The conceptual status of *DSM-5* diagnoses. In J. Paris & J. Phillips (Eds.), *Making the DSM-5: Concepts and controversies* (pp. 143–157). Springer.

Richter, P., Werner, J., Heerlein, A., Kraus, A., & Sauer, H. (1998). On the validity of the Beck Depression Inventory: A review. *Psychopathology*, *31*(3), 160–168.

Schaffner, K. F. (2012). A philosophical overview of the problems of validity for psychiatric disorders. In K. S. Kendler & J. Parnas (Eds.), *Philosophical issues in psychiatry II: Nosology* (pp. 169–89). Oxford Academic.

Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. Springer.

Solomon, M. (2022). On validators for psychiatric categories. *Philosophy of medicine*, *3*(1), 1–23.

Tsou, J. Y. (2021). *Philosophy of psychiatry*. Cambridge University Press.

Wakefield, J. C. (2013). The *DSM-5* debate over the bereavement exclusion: Psychiatric diagnosis and the future of empirically supported treatment. *Clinical psychology review*, *33*(7), 825–845.

Zachar, P. (2012). Progress and the calibration of scientific constructs: the role of comparative validity. In K. S. Kendler & J. Parnas (Eds.), *Philosophical issues in psychiatry II: Nosology* (pp. 21–40). Oxford Academic.

Chapter 2: Diagnostic Validity

2.1 Introduction: The *DSM* and the "Box Canyon Problem"

At the second international conference in Copenhagen in the Philosophical Issues in Psychiatry series, which took place just following the creation of the *DSM-5*'s Scientific Review Committee in 2010, Kenneth Kendler introduced the concept of an "iterative box canyon" to help illustrate an undesirable situation that psychiatry may be facing, in its quest for developing valid etiologically based psychiatric classifications. Consider the following analogy: imagine scientific progress in psychiatry as akin to being on a long and difficult hike toward the top of a mountain, which in this case is valid psychiatric disorders that stand for real underlying disease entities. While our hike may be slow going and can sometimes feel as if we are making very little progress, if stay on the trail we can be assured that the incremental steps we take are leading us forward. But what if we accidentally veer onto the wrong path into a box canyon—a narrow passage enclosed on all sides by steep vertical walls of which there is no easy way out? How do we know whether to press on, confident that our small, iterative steps will take us through and (hopefully) out of the canyon and toward our goal, or whether we're stuck, and that it's time to turn around and head out the way we came?

The so-called box canyon problem of psychiatric nosology, initially introduced as a simple thought experiment, has come to be viewed as an apt metaphor for the kind of real trouble we're in concerning the validity of our psychiatric disorders. For many, the *DSM*'s lack of validity is a sign that we are most certainly stuck in a box canyon. And if we are truly stuck, then remaining on the same path will be of no help in getting us out. It is the *DSM* system itself that serves many masters in its competing scientific, professional, and practical aims that may have

psychiatry boxed in. Thus, to simply continue might inevitably keep us hiking endlessly in our search for safe passage out of the canyon, never allowing psychiatry to achieve any real sense of scientific progress to which it once aspired.

Assuming that psychiatry is currently stuck in a box canyon of sorts, there are two primary solutions. The first solution, which will be addressed in chapters 3, 4, and 5, is to retrace our steps out of the canyon back to or near where we initially began, and ultimately choose a different path. This solution is characterized by Kendler (2014) as "thinking our way out of the box canyon" in that it amounts to a consensus-based selection and development of an alternative theoretical framework that we hope will lead us toward developing more valid psychiatric classifications. The risks of such an approach, however, are twofold. First, if our original path from which we departed was the right and only correct path, then we may journey even further away from our desired goal. Second, in leaving our current path, we may encounter practical consequences by adopting a new framework that is unable to be utilized by clinicians and other professionals who depend on our existing diagnostic framework.

The second solution, and the focus of this chapter, is to remain on the empirically rigorous path on which the current editions of the *DSM* are currently set. The *DSM*'s evolutionary approach to achieving validity by way of small, incremental improvements is described as a type of epistemic iteration on our psychiatric classifications based on their performance on specific validators (Kendler & Parnas, 2012). Proponents of this approach argue that despite it being more effortful and still involving compromises with the *DSM*'s clinical and practical aims, it will nevertheless allow us to successfully iterate our way out of the canyon and continue along our path toward valid psychiatric diagnoses.

The goal of this second chapter is to faithfully reconstruct the *DSM*'s empirically grounded plan for achieving validity in psychiatry. I will describe the *DSM*'s particular sense of validity and validation, referred to as diagnostic validity, first providing a general overview, then going briefly back over its historical development within the *DSM* and its key features. I then analyze its current relation to and implementation of construct validity theory (CVT) and conclude with a discussion of its philosophical foundations. In doing so, I aim to provide an anchor by which other senses of validity from the aforementioned Holy Quadrinity of validity in psychiatry may be distinguished.

2.2 Overview of Diagnostic Validity

Validity within psychiatric classification as featured in the *Diagnostic and Statistical Manual of Mental Disorders* (5th Edition, Text Revision) is referred to as *diagnostic validity*. Diagnostic validity can be understood as the extent that a *diagnostic category*, comprised of a set of operationalized diagnostic criteria intended to represent an underlying mental disorder, is supported by a specific set of *validators*. Validators are understood as acceptable sources of validating evidence for a diagnostic category. According to Kendler et al. (2009) "the 'bottom line' question to be evaluated is whether we have any confidence in the validity of this syndrome based on the set of validators" (p. 8). Diagnostic validity is based on an evaluation of "the overall strength of evidence across all validators," and an evaluation of the "strength of evidence for each of the validators" (p. 3).

Evaluation of the accumulation of evidence from the validators, i.e., the aggregation of the validators, contributes to a judgment regarding the overall degree of diagnostic validity of a diagnostic category. Evidence conceived as a validator may be drawn upon for uses beyond

diagnostic validity—for example, validators may demonstrate evidence for a diagnostic category's predictive and clinical utility, as well as be informative for carrying out additional studies (Solomon, 2022). Within a validity context, however, what makes a type of evidence validating evidence is its designated role in establishing diagnostic validity. Table 2, Appendix lists the current set of validators for the *DSM*, of which those denoted by an asterisk (*) are deemed high priority by the *DSM* and given the greatest emphasis, followed by a brief description of each validator class.

Validation within the *DSM* is an iterative process by which researchers collect and evaluate evidence to judge whether a diagnostic category sufficiently stands for some underlying disorder. The process for achieving diagnostic validity has both preserved and significantly expanded upon the original method first articulated by Eli Robins and Samuel Guze (1970) and Feighner et. Al. (1972). Its expansion typically proceeds in conjunction with periods of development preceding the publication of the latest edition of the Diagnostic and Statistical Manuel of Mental Disorders (*DSM*), the psychiatric research and classificatory framework most notably associated with and based in diagnostic validity since its third edition (*DSM-III*, 1980), which has since undergone several iterations (*DSM-III-R*, 1987, *DSM-IV*, 1994, *DSM-IV-TR*, 2000, *DSM-5*, 2013, *DSM-5-TR*, 2022). Official publications by the *DSM* of an agenda or guidelines for a subsequent *DSM* edition outline guiding principles and currently accepted validation standards that are intended to inform the *DSM*'s revision process of its diagnostic categories. Most recent guidelines include Kendler et al.'s (2009) "Guidelines for Making Changes to the *DSM*" and the American Psychiatric Association *DSM* Team's "Guide to Submitting Proposals for Changes to the *DSM-5*" (American Psychiatric Association [APA], 2021).

The *DSM*'s diagnostic categories are presently organized into one of twenty different disorder groupings. This section takes the example of "Major Depressive Disorder," which is a diagnostic category in the "Depressive Disorders" grouping. In clinical practice, the diagnosis of mental disorders relies on operationalized descriptions of the diagnostic categories referred to as diagnostic criteria. Diagnostic criteria "identify symptoms and signs comprising effects, behaviors, cognitive functions, and personality traits along with physical signs, symptom combinations (syndromes), and durations" (APA, 2013) (Table 3, Appendix). Diagnostic criteria "approved for routine clinical use" are accompanied by descriptive text that includes additional information such as "coding and recording procedures," "diagnostic features," "associated features," "prevalence," "development and course," "risk and prognostic factors," and "culture and gender-related diagnostic issues."

In the context of validity and validation, a diagnostic category's set of diagnostic criteria are infallible indicators of the underlying clinical syndrome. The symptoms only index the disease, they do not constitute the underlying syndrome. This distinction between the diagnostic criteria (i.e., the symptoms) and the underlying clinical syndrome is important, as it supports two overlapping yet distinct senses of diagnostic validity that may be utilized in practice. The first sense, which we may describe as *small-V diagnostic validity*, relates to judgments regarding whether a diagnostic category has validity in the sense that its diagnostic criteria can accurately identify, measure, or refer to some underlying concept. In this more instrumental or pragmatic sense of diagnostic validity, diagnostic criteria are treated as measurement instruments as such to determine the presence of a disorder. Specifically, small-V diagnostic validity amounts to judgments about how well a diagnostic criteria set is at identifying, differentiating, and predicting homogeneous diagnostic groupings. The goal in achieving small-V diagnostic validity

is to contribute to knowledge about the diagnostic category and thus the diagnostic criteria. The second sense of diagnostic validity, which we might describe as *big-V diagnostic validity*, relates to whether the diagnostic category stands for some real underlying clinical syndrome. In this more realist sense of diagnostic validity, the goal is to provide new information about the *focal phenomena*, the underlying clinical entity, that the diagnostic criteria *actually* represent, to establish the existence of the disorder. It is this sense of diagnostic validity with which the crisis of confidence in validity in psychiatry is primarily concerned (Phillips, 2013).

The overall strategy by which the *DSM* aims to achieve both small-V diagnostic validity and big-V diagnostic validity is characterized as an iterative model of nosological change. In this model, higher degrees of validity are achieved via small iterations of presently established diagnostic criteria. Any changes, no matter how small, must first be vetted through validating measures, which are evaluations of the proposed changes' performance on the validators. According to Kendler and First (2010), the iterative model operates under the assumption that by "using increasingly rigorous empirical methods, each subsequent revision of our diagnostic system will produce improvements over its predecessor" (p. 263). Such improvements, however, have undoubtedly been very slow-going. But the deliberately cautious nature of this approach, one that has resulted in what some might consider to be a painstakingly slow output of big-V diagnostic validity and a reason to shift gears, is defended by its proponents as a virtue that may minimize impediments to clinical practice and research.

Often mixed in with questions of diagnostic validity is the definition of a psychiatric disorder. Some interpretations of diagnostic validity have even conceived of validity as the degree to which a diagnostic category meets the criteria for a received definition of mental disorder, e.g., whether Major Depressive Disorder constitutes a harmful dysfunction (Willcut &

Carlson, 2005). For the *DSM* and the iterative model, however, whether a diagnostic category meets the criteria for a psychiatric disorder is more than just a question of its diagnostic validity and should be interpreted as a separate matter. Whether a diagnostic category has sufficient diagnostic validity is one of many considerations as to whether a diagnostic category may be judged to meet criteria for what the *DSM* refers to as a "Mental (Psychiatric) Diagnosis" (Table 4, Appendix). Proposed criteria for psychiatric disorders such as "reflects an underlying psychobiological disturbance," or represents "a behavioral or psychological syndrome or pattern that occurs in an individual," while potentially related to desired aspects of big-V diagnostic validity, are not to be interpreted as questions of validity, and are instead evaluated separately.

Furthermore, whether a diagnostic category meets criteria for a mental disorder is, in turn, one of several other considerations for the evaluation of a category's ultimate placement in the *DSM*. A host of other considerations by which a diagnostic category may be judged for inclusion are evidence for the category's *reliability*, defined by the *DSM* as "the degree to which two clinicians could independently arrive at the same diagnosis for a given patient" (APA, 2022, p. 8); *clinical utility*, defined as the ability to "help clinicians to determine prognosis, treatment plans, and potential treatment outcomes for their patients" (p. 14); as well as many practical considerations such as the need for the category, any potential harm caused by the category, and evidence that there would be available treatments if a category were to be accepted. Under the *DSM*, diagnostic validity is assessed independently of and weighed in conjunction with, many other more practical considerations for a diagnostic category's inclusion.

The specified relation of diagnostic validity as playing a supporting role in a diagnostic category's evaluation as a psychiatric disorder to support its potential inclusion in, modification, or deletion from the *DSM* is reflective of diagnostic validity having advanced toward a stage of

an application-specific validation process. By this, I mean the ongoing development of diagnostic validity and the sense in which something may be interpreted as achieving diagnostic validity is now tightly coupled with the practical aims, practices, and standards of the application for which it serves, the *DSM*'s categorical classification framework. While the original conception of diagnostic validity by Robins and Guze was developed independently of the *DSM* and with broader applications in mind (Robins & Barrett, 1989) so that a host of different classifications might be subjected to its method, the *DSM*'s appropriation and continuous updating of the diagnostic validation process in close step with the *DSM*'s official guidelines for revision has made the method inseparable from that of the application for which it is now utilized. The choice of specific validators and their level of priority by the *DSM* to other factors influencing classification decisions have now come to constitute diagnostic validity. While other validation procedures exist within alternative frameworks for establishing specific senses of validity, there is no validation process for achieving diagnostic validity outside of the *DSM*.

To better understand the assumptions and reasoning behind the development of diagnostic validity as an application-specific model for validity and validation of the *DSM*, and how its development has been shaped by the apparent crisis of confidence in the validity, so that an "iterative" model of revision to resolve the "box canyon problem" is warranted, I now briefly turn to its historical origins. I begin from its initial articulations by Robins and Guze (1970) and Feighner et al. (1972), followed by the various updates in method in all editions of the *DSM*.

2.3 Origins of Diagnostic Validity: Robins and Guze's Method

Diagnostic validity may be traced to the neo-Kraepelinians—a group of biologically oriented psychiatrists at Washington University in St. Louis during the 1960s and 1970s whose

leaders included Eli Robins (1921–1994) and Samuel Guze (1925–1996)—who came to adopt psychiatrist Emil Kraepelin's (1856–1926) psychiatric nosology. Most notably, the group came to embody what psychologist Richard Bentall (2003) referred to as Kraepelin's "big idea": that psychiatric disorders are discrete disease entities that can be accurately identified through clinical observation of their signs and symptoms, direct observation of their pathological anatomy, or study of their etiology. By assuming that each psychiatric disorder possessed a typical symptomology that directly corresponded with a particular etiology and underlying brain pathology, the development of classifications based first on symptoms—the most readily available and accessible data at the time—could serve as a viable starting point for the future identification of the presently more inaccessible and elusive features of psychiatric disorders (i.e., their etiology and pathophysiology). The overall strategy was simple: 1) identify discrete and homogenous diagnostic groups or clusters" i.e., syndromes, based on signs and symptoms, and 2) further study those homogenous clusters which may serve as stepping-stones to discovering their underlying and corresponding pathological processes.

In developing their strategy, Robins and Guze would come to adopt additional Kraepelinian empirical standards that would inform their own standards for reliability and validity. Like Kraepelin, they were unsympathetic to the psychodynamic and psychoanalytic theories prevalent within psychiatry, whose contributions they viewed as merely providing poor speculations about the etiology of which they believed was unknown for most psychiatric diseases (Decker, 2013). Robins and Guze had observed the incredibly poor inter-rater reliability of the *DSM-I* (APA, 1952) and *DSM-II* (APA, 1968), in other words, when multiple psychiatrists use the same criteria to come to a shared conclusion regarding the diagnosis of a patient (Weber, 2007). They believed the unscientific nature of psychiatry that had come to be dominated by

"opinion or tradition" rather than "data" was to blame (Feighner et al., 1972, p. 62). Neo-Kraepelinians viewed the diagnoses within psychiatry as effectively "ideas without data" (both mathematical and etiological), and resultingly, saw little if any validity in them. The terms validity or validation do not appear within *DSM-I*, and validity is mentioned only once in the *DSM-II* in brief about the importance of appropriate diagnostic terminology. While validity in the sense of the accuracy or truth of these diagnostic categories remained a point of discussion, the development of a formal empirical process of validation that would support the construction of diagnostic criteria was not readily present.

To make psychiatry more scientific, the Washington University group adopted the use of operational criteria to construct their diagnoses. The concept of operational criteria in psychiatry is attributed to the work of Mandel E. Cohen (1907–2000), a Harvard clinical psychiatrist who trained Robins while Robins was a medical resident, and who ultimately recommended him to take on his position in the psychiatry department at Washington University (Decker, 2013). Cohen's formation of the concept was likely to have been influenced by Harvard physicist and mathematician, Percy Williams Bridgman (1882–1961) who in 1927 introduced the concept of operational analysis, which posited that concepts should be stipulated in terms of the operations used to establish their existence (Bridgman, 1927, p. 57). Applying Bridgman's operationalism to psychiatric diagnoses meant limiting psychiatric diagnoses only to empirically observable phenomena, i.e., their operational criteria. Developing clearly defined descriptive criteria of psychiatric disorders by which to organize and base psychiatric research, it was believed, would bring about a level of precision and scientific rigor required to realize a biologically oriented psychiatry and facilitate the validation of psychiatric diagnoses:

> …by providing operational definitions of the disorders [that] will permit family, treatment and outcome studies, as well as systematic inquiry into etiology, the ultimate predictive validity of the *DSM-III* diagnoses can be determined with an accuracy heretofore impossible. (Spitzer et al., 1977, p. 15, 17)

Thus, to reorient psychiatry toward a scientific (and medical) discipline that would align Kraepelin's big idea with more rigorous empirical standards, Robins and Guze contended that outlining an explicit empirical method or process to support the construction of valid operationalized diagnostic criteria—one based in a "no-nonsense data-oriented approach" Decker, 2013, p. 56)—would be essential. In addition to several other shared positions with Kraepelin such as descriptive observations of clinical features and a focus on the longitudinal course of disorder (Robins & Barrett, 1989), Robins and Guze and the other neo-Kraepelinians shared two specific positions regarding reliability and validity that would inform their method as summarized by Kramer (1978). The first is that "a legitimate and valued area of research should be to validate such criteria by various techniques." The second is "in research efforts directed at improving the reliability and validity of diagnosis and classification, statistical techniques should be utilized" (Kramer, 1978, p. 105). Resultingly, diagnostic criteria would come to be supported by having achieved validity based on evaluations of "systematic studies" (Robins and Guze, 1970, p. 983).

Robins and Guze (1970) outlined their "method for achieving diagnostic validity in psychiatric illness" (p. 983) in their paper "Establishment of Diagnostic Validity in Psychiatric Illness: Its Application to Schizophrenia." Their method amounted to a five-phase approach for establishing valid classifications and, as a result, diagnostic validity. Each phase included specific validating evidence that was applied to the "homogenous diagnostic groupings" (p. 984)

of patients to establish the validity of diagnoses. The complete five phases are summarized in order in Table 5, Appendix.

Two years later, Robins and Guze would co-author a paper (1972) with lead author John Feighner and George Winokur (fellow neo-Kraepelinians of the Washington University group) that repeated their method while introducing the "most efficient currently available" diagnostic criteria for fourteen psychiatric illnesses (p. 57). The conditions, which became the Feighner Criteria, were developed via clinical expert consensus much like the diagnostic categories in the *DSM-II*. However, where the Feighner criteria differed significantly was in the way they were validated, namely by follow-up studies, family studies, and other systematic studies from Robins and Guze's five phases. The Feigner criteria were expanded upon in the publication of the *Research Diagnostic Criteria* by Spitzer, Endicott, and Robins (1978), which would become the road map for the development of 265 diagnostic categories in the *DSM-III* (1980), and inform the modification of the International Classification of Diseases, Ninth Revision (ICD-9) for use in the U.S.

The *DSM-III* steering committee or Task Force "relied, as much as possible, on research evidence relevant to various kinds of diagnostic validity" (APA, 1980, p. 3). Although an explicit method for achieving diagnostic validity was not detailed in the *DSM-III*, decisions on proposals for the inclusion/exclusion of diagnostic categories relied primarily on inter-rater reliability, and secondarily on "treatment planning, course, and familial pattern" (p. 3) and other "validity data" (p. 378) reminiscent of Robins and Guze's method when such validity information was available. An explicit goal of the *DSM-III* considered proposals that demonstrated "consistency with data from research studies bearing on the validity of diagnostic categories" (p. 2), and a "lack of

validity evidence" (p. 8) reportedly resulted in the removal of several newly proposed diagnostic categories (APA, 1980).

In addition to supporting valid diagnostic criteria, a central aim of the *DSM-III* was to support the creation of reliable and clinically useful diagnoses. Aside from supporting a greater sense of reliability and clinical utility which prior editions of the *DSM* had been lacking for use in clinical settings, having reliable diagnoses was viewed as a separate yet essential taxonomic step for researchers to further establish additional degrees of diagnostic validity: "…communication is meant to provide common ground for different research groups so that diagnostic definitions can be emended constructively…The use of formal diagnostic criteria by a number of groups should expedite psychiatric investigation. (Feighner et al., 1972, p. 57) At the same time, prioritization of reliability and clinical utility was by and large intended to meet the various clinical and practical concerns of a diagnostic classification system. The introduction of descriptive, operationalized diagnostic criteria via the Feighner and Research Diagnostic Criteria would allow clinicians and researchers everywhere to draw on the same diagnoses, thus increasing their reliability and usefulness. Factoring into the acceptance of a diagnostic category would additionally include evaluations of the perceived need for a diagnostic category, foreseeable consequences such as potential harm, and ability and access to treatment. Additionally, there was an ongoing clash in the development of the *DSM-III* between the needs of clinicians who, coming to be represented and fiercely defended by Spitzer himself, argued for the inclusion of certain diagnostic categories based on a specific interpretation of diagnostic validity as clinical utility (Spitzer, 2001, p. 354). This was at odds with those of the researchers who, represented by the Washington University group, only wanted to include diagnoses in the *DSM* if (and only if) they were judged to be valid under Robins and Guze's more etio-

pathophysiological method. Ultimately, acceptance of a diagnostic category in the *DSM-III* would come to be based on compromises between competing aims, with a primary goal still being the development of psychiatric classifications by which big-V diagnostic validity may eventually be achieved.

Despite the inclusion of certain diagnostic categories such as those in the personality disorder (PD) groupings that were judged by some to be more clinically useful than valid, the use of Robins and Guze's method within the *DSM* was still anticipated to re-align psychiatry with the medical model from which it had been distanced by psychoanalysis. The diagnostic validation of the diagnostic criteria, i.e., small-V diagnostic validity, would provide a steppingstone to a distinctive kind of *etio-pathogenic* validation of the underlying psychiatric phenomena that they assumed would bear out—one that would ultimately amount to big-V diagnostic validity:

> [T]hese five phases interact with one another so that new findings in any one of the phases may lead to modifications in one or more of the other phases. The entire process is therefore one of continuing self-rectification and increasing refinement leading to more homogeneous diagnostic grouping. Such homogeneous diagnostic grouping provides the soundest base for studies of etiology, pathogenesis, and treatment. (Robins & Guze, 1970, p. 984)

2.4 Diagnostic Validity II: The Concept of Validators and Matters Nonempirical

In the years following the publication of the *DSM-III* (1980) and a significant revision led by Spitzer, the *DSM-III-R* (American Psychiatric Association, 1987), concerns regarding a lack of validity of the *DSM*'s diagnostic criteria were already being raised. This "lack of validity" was generally referring to an absence of big-V diagnostic validity, or the absence of validity of the focal phenomena*,* conceptualized as "the disease entity":

> In the past 20 years, however, the disease entity assumption has been increasingly questioned as evidence has accumulated that prototypical mental disorders such as major depressive disorder, anxiety disorders, schizophrenia, and bipolar disorder seem to merge imperceptibly both into one another and into normality. (First et al., 2002, p. 12)

Two primary explanations for a lack of validity in this sense were offered. The first centered around the idea that the types of validating evidence being used to validate the diagnoses, i.e., the diagnostic criteria, were incomplete. As a result, diagnostic criteria based on incomplete validating evidence did not turn out to be an all that relevant basis for supporting the validation of the focal phenomena. One solution would be to maintain the overall approach of Robins and Guze but to expand their steps into a list of types of validating evidence. The inclusion of additional validating evidence that was considered more relevant for exploring and assessing the underlying disease construct would provide greater interaction between the validity of the diagnostic criteria and the validity of the construct still assumed to underly the criteria. A need for a reorganization and expansion of Robins and Guze's method was already underway at the time of the *DSM-III*'s publication, with Kendler (1980) introducing an "adaptation and enlargement" of the method as well as the concept of a validator, designating a particular source of validating evidence (p. 700).

A second explanation for the apparent lack of validity was how the validating evidence for diagnostic criteria were being evaluated and interpreted in practice. Specifically, the application of the method of achieving diagnostic validity was deemed unscientific. Disagreements about what is being validated among researchers, and which validators to prioritize when forming a judgment regarding their validity resulted in "*fundamentally* nonempirical" aspects of psychiatric nosology that, in the context of validation, needed systemization and standardization. (Kendler, 1990, p.972, emphasis in original). To move toward a greater scientific nosology—one that

would be more empirically informed but necessarily contain value judgments—Kendler (1990), based on his experience observing the many changes made in the *DSM-III-R*, felt it necessary to provide a "clear criterion for which to evaluate nosological proposals" (p. 970) in which diagnostic validity would be a part. Departing from an "advocacy model" (p. 972) that permitted selective interpretation of the available evidence to support one's proposals, Kendler advocated for what he termed an "advisory model" for nosological change:

> The most appropriate use of empirical data in our nosologic process, then, is the *advisory model*. In this approach, the "scientific" information relevant to a given nosologic issue is systematically gathered and objectively evaluated. This information is used to advise and inform the committee, which then makes decisions guided by this information, but also considering other, nonempirical, issues. (Kendler, 1990, p. 972, emphasis in original)

Thus, the next iteration of diagnostic validity would be marked by the expansion of validating evidence, designation of such evidence as a validator which would come to help with the organization and specificity of validating evidence, and standards for how to properly assess the validators. Such standards, while presented as separate from the validation process, still heavily involve an evaluative component, and thus, may be conceived as now a part of the expanded model of diagnostic validity. With the goal being the development of a more scientifically based psychiatry nosology, the central aim of whether a diagnostic category had diagnostic validity was more and more beginning to be interpreted as the degree to which the diagnostic criteria reflected the underlying disease entity, i.e., big-V diagnostic validity.

2.5 Diagnostic Validity 3: An Iterative Model for Nosological Change

The *DSM-IV* (American Psychiatric Association, 1994) would come to reflect a deepened emphasis and discussion on the use of empirical studies in validating diagnostic criteria. Allen

Frances would replace Spitzer as the chair of the *DSM* Task Force, and in doing so dictate that all changes to the *DSM* would be based on evidence and that no other major changes would be permitted. A methods and applications conference held in November 1988 took place to determine the specific evidence-based criteria for updating diagnostic categories in the *DSM-IV*. In addition to utility and reliability, the primary domains for making decisions included descriptive validity and external validators. (Kendler, 1988). Descriptive validity would include empirical data on features such as predictive power, co-occurrence rates, and longitudinal comorbidity data. Validators would include antecedent variables (e.g., demographic data, family history), concurrent variables (e.g., biological and psychological variables), and predictor variables (e.g., diagnostic consistency, treatment responses). Additional consideration went toward considering how to aggregate and analyze the data from different studies to come to an objective consideration of the empirical literature (Frances et al., 1989). Despite these advances in validation, criticisms of the *DSM-IV*'s continued use of descriptive-based classifications and subsequent lack of big-V validity would persist, leading for a call for a "paradigm shift" in psychiatric classification (Kendler & First, 2010, p. 263).

Between the publication of the *DSM-IV* and its text revision the *DSM-IV-TR* (American Psychiatric Association, 2000), Andreasen (1995) proposed a model of diagnostic validity that reflected "the need for a second structural program for validating psychiatric diagnosis" (p. 161). Meant to be complimentary to Robins and Guze's initial approach as well as Kendler's extension of their method, their model would add additional validators to be considered "that can be used to link symptoms and diagnoses to their neural substrates" and range "across multiple levels of conceptualization" (Andreasen, 1995, pp. 161-162). The program was meant to reflect the already ongoing research into the etiology and pathophysiology of psychiatric disorders. Such

validators could include findings from studies in molecular genetics and molecular biology, neurochemistry, neuroanatomy, neurophysiology, and cognitive neuroscience, to name a few. Under Andreasen's conception of diagnostic validity, the use of biologically based validators would go beyond the *DSM*'s labeling and demonstration of clinical stability of its diagnostic criteria, and in turn, provide powerful credibility to psychiatric diagnoses as real entities. Andreasen's proposal reflected the aspirations and hope of biologically oriented psychiatrists that future editions of the *DSM* (the *DSM-5*) would finally begin to incorporate various aspects of the underlying biology and develop etiologically based classifications.

The *DSM-5* Steering Committee officially adopted and expanded upon the sources of validating evidence with the inclusion of the validators concept. The standardization of the prioritization of validating evidence was adopted as a central tenet of the diagnostic validation process, so that "…once the critical validators are agreed on, only then can the process of formulating maximally valid criteria sets occur" (First, Regier, & Kupfer, 2002, p. 8). The original vision of the *DSM-5* Steering Committee as outlined in *A Research Agenda for DSM-V* (First, Regier, & Kupfer, 2002) was to promote bold and new proposals for the fifth edition of the *DSM* that would prioritize biologically based validators. To the disappointment of many, however, the *DSM-5* would stop well short of making any kind of paradigm shift that would replace the *DSM*'s descriptive classifications with those that were primarily etiologically based. Kendler and First (2010) explained the justification for the *DSM*'s conservative decision to maintain its prior approach, stating:

> Changing paradigms will place a considerable burden on the *DSM* user-community in terms of the costs of learning the new system, implementing new diagnostic and assessment procedures and creating a significant discontinuity in diagnostic data-sets. To sustain a scientific revolution, both the push of increasing dissatisfaction with the old paradigm and the pull of a new paradigm that can

> definitively address many of these concerns are needed. Although our field clearly has the 'push' for change, we do not yet have a strong enough 'pull' from a superior alternative paradigm that will successfully address our concerns. (Kendler & First, 2010, p. 265)

The *DSM-5* would instead adopt an iterative approach to revision, whereby incremental changes would be made to the previously existing paradigm of symptom-based classifications. One potential within-paradigm change, however, would come in what Regier et al. (2009) would describe as "the more prominent use of dimensional measures" (p. 649). A second would be in the prioritization of three specific validators: "Familial Aggregation," "Diagnostic Stability," and "Response to Treatment." Notably not prioritized as a part of the original vision for the *DSM-5* were Biological Markers, e.g., molecular genetics, neural substates" which the Steering Committee had soon realized would not be available for most of the diagnostic categories. The designation of critical validators to the method for achieving diagnostic validity came to reflect an important new aspect regarding validation process: that diagnostic validity was no longer intended to be only the steppingstone to knowledge of the focal phenomena but was to begin to directly contribute to the testing of theoretical knowledge of the focal phenomena underlying the diagnostic category. The hope was that the critical validators were prioritized in such a way as to draw greater insights that would ultimately lead to the development of theory surrounding psychiatric disorders. Further, the method of diagnostic validity was now being updated and re-written exclusively within the official guidelines of the *DSM*, for the purposes of the *DSM* categories, and thus had advanced to what I term an application-specific validation process. The process of establishment of diagnostic validity may no longer be viewed as a method that could be readily and easily applied outside the context of the application in which it is utilized. Additionally, in the spelling out of explicit guidelines for evaluating sources of

validator evidence, the diagnostic validity method would come to resemble Kane's argument-based validation framework, and his interpretation/use argument (IUA) (2013). For example, consider the following question posed in the Guide to Submitting Proposals for Change to the *DSM-5*: "For Type 1a (criteria set changes to improve validity), the question will typically be: is the validity of the proposed set of criteria for disorder X superior to the *DSM-5* criteria for disorder X?" (APA 2021, p. 6). To come to a determination regarding which disorder displays the more superior diagnostic validity, a researcher may be said to construct a *validity argument* based on the accumulation of validators. That is, validation is not simply objective hypothesis testing, but will necessarily involve some sense of organized evaluation and discernment of the evidence from the available validators, presented in the form of an argument.

2.6 Diagnostic Validity 3-TR: The Continuous Improvement Model

In the most recent version, diagnostic validity has maintained its validators with the addition of "Degree or nature of functional impairment," as well as an elevated prioritization granted to "Biological Markers." These revisions are reflective of the method of diagnostic validity further being utilized with the intent for it to contribute directly to the development of etiological and pathophysiological theory regarding the focal phenomena and thus big-V validity. The notion of "whether we have any confidence in the validity of this syndrome based on the set of validators" (Kendler, 2009, p. 8) is more and more interpreted as confidence of the diagnostic category as standing for some biologically based clinical syndrome.

Perhaps one of the biggest changes in the way diagnostic validity is incorporated in the *DSM* recently, however, is related to the *DSM*'s shift in how it is updated. The previous mechanism for changing the diagnostic criteria was to revise the entire *DSM* manual every ten or

so years, which had the advantage of facilitating agreed-upon standards of communication for clinicians and researchers. As of 2021, the *DSM* now operates under a continuous improvement model, whereby proposal changes for the *DSM* may be submitted at any time on its website and are reviewed on an ongoing basis. The goal of this model is to more readily incorporate "new scientific knowledge into the manual" as it emerges, coming to better approximate what Kendler had in mind with the iterative model in terms of it producing a "steady improvement in the validity of the diagnostic system" (Kendler & First, 2010, p. 263).

A Guide to Submitting Proposals for Changes to the *DSM-5* (APA, 2021) details the types of validating evidence that are expected to be submitted. Of note are more specific standards for prioritizing sources of validating evidence as they relate to the type of modification or refinement. For example, Type 1A is a "proposal for changes to an existing diagnostic criteria set that would markedly improve its validity" prioritizes evidence from and across different sources of the validators, whereas Type1C proposes "changes to existing diagnostic criteria set that would markedly improve clinical utility without an undue reduction in validity or reliability," which prioritizes evidence in favor of demonstrating clinical utility. All proposals are expected to include a 1) clear summary statement of the rationale for the proposed change, 2) historical context for the proposal, 3) discussion of possible negative consequences of the proposed change and consideration of arguments against the change, 4) magnitude of the proposed change, 5) evidence from the validators for the change, 6) evidence of reliability, 7) evidence of clinical utility, and 8) evidence of deleterious consequences.

2.7 Diagnostic Validity: Validator-Specific, or Anything Goes?

An additional point of change in the method of diagnostic validity is such that validating evidence not directly featured in the list of validators may, in certain instances, factor into the evaluation of the overall diagnostic validity of a particular diagnostic category. For example, in a study of the revisions of the substance-related disorders in the *DSM-5*, Zachar, First, and Kendler (2022) found that changes to specific aspects of diagnostic criteria were informed by psychometric evidence from an application of Item Response Theory, although such evidence is not considered part of the official method for achieving diagnostic validity. Slaney (2017) has observed discrepancies in the practice of validation research whereby the general aims of the model for validation do not always coincide with how the validation process is conducted in practice. And in some sense, according to modern validity theory, validity is not supposed to be a checklist, as the evaluation of available validity can change based on the purposes and intent of use (contrary to this, the *DSM* indeed publishes such validity checklists for its sources for validating evidence). It is unclear if this is evidence that strength and interpretation of the argument made based on the evidence is most important for diagnostic validity rather than the type of validating evidence itself. Further investigation as to why certain evidence not normally considered relevant for achieving diagnostic validity was accepted in this instance would be informative.

2.8 Understanding Diagnostic Validity Through the Lens of Construct Validity Theory (CVT)

Through an examination of the development of diagnostic validity, we have observed an active effort by those who have come to utilize its method to make it more scientific to address concerns regarding an apparent lack of validity of diagnostic categories. Another general

approach to validity that has dealt with its own scientific status is that of construct validity theory (CVT). While diagnostic validity may be conceived as distinct in sense and process from that of construct validity, there are aspects in which it is reflective of CVT. In fact, it would be difficult to find any application-specific model or account of validity in any other domain or field that has not at least to some degree come to draw on or relate with CVT. These surface-level similarities between diagnostic validity and CVT, however, are advantageous. Since most of the conceptual and theoretical work in relation to validity and validation has been done within general validity theory, we may utilize the concepts and language of CVT as a helpful lens through which to further analyze and more precisely characterize the aims of diagnostic validity that are associated with their particular senses, that is, small-V and big-V diagnostic validity.

In a study of construct validation practices, Slaney (2017) has observed how scientific aims in relation to construct validation generally fall into three categories: (1) validation of an instrument (e.g., a test or measure) that may involve an evaluation of the psychometric properties or the reliability and validity of the available data, which tells us something about the test or measure; (2) validation of a focal phenomena, with the aim to explore what the instrument is *actually* measuring, with the hope of providing new insights about the underlying focal construct; and (3) validation of a theoretical framework, whereby an evaluation of two or more models of a construct would contribute to knowledge of the construct of interest, and/or a mixture of all three aims to varying degrees.

The aims of the method for achieving diagnostic validity generally appear to be a mixture of Slaney's three categorizations. First, there is validation of the available evidence, often drawn from screenings and questionnaires but additional from other kinds of tests, that are utilized in developing the diagnostic criteria. That is, the validity of diagnostic criteria is based on the

validation of the tests or measures used to create them, as opposed to *DSM-II* being chiefly based on "the best clinical judgment and experience" of a committee (Feighner et al., 1972, p. 57). Second, their method aimed to provide a basis for exploration and validation of the constructs hypothesized to underly the diagnostic criteria, i.e., the real clinical syndromes. The relevant distinction to be made between "based on" and "a basis for" here is that validation of the focal construct, being the underlying clinical syndrome, is not completely "based on" valid diagnostic criteria, but that valid diagnostic criteria is thought to provide "a basis for," i.e., greater means to accumulate and test validating evidence, relevant for establishing the validity the underlying focal construct. Thus, at least initially, these two senses of validity—validity of the diagnostic criteria, and validity of the construct hypothesized to underly the criteria—remained separate. Third, their method was consistently framed as testing the theoretical framework regarding two or more models, e.g., in utilizing validating evidence to show that good prognosis "schizophrenia" is not mild schizophrenia, but a different illness as featured in the original Robins and Guze (1970) publication.

In terms of a particular account of construct validity, small-V diagnostic validity may be considered to most closely reflect what Cronbach (1988, 1989) would refer to as the "weak" program of CVT. Distinct from the "strong" program which required testable theories and which Cronbach (1989) viewed as being the "most appropriate to a scientific perspective that reaches centuries into the future" (p. 163), the "weak" program was amenable to the sciences with an absence of such theories (viz., the social and behavioral sciences), whereby "any evidence even remotely connected to the test scores is relevant to validity" (Kane, 2001, p. 326). The criticism of the "weak" program by Cronbach (1988) as constituting "sheer exploratory empiricism" (p. 12) related to the idea that any sort of external correlation could then later be interpreted as

contributing an overall sense of construct validation without a clear sense as to precisely how the evidence contributes to the validation of that construct. The utility of the "weak" program for Cronbach (1989), which aligns with the aims of Robins and Guze, was so that it may play a significant role in the beginning stages of the development and validation of constructs for both classificatory and research purposes which are notably absent of theory. For Cronbach (1989), the "weak" program "enables us to identify sensible alternatives in practical affairs and in the planning of research" (p. 163) much like that of small-V diagnostic validity. A consequence of basing one's process of validation on a "weak" program of CVT, however, is an absence of explicit recommendations in terms of how to determine what validity evidence is most relevant, as well as how much evidence is required. As a result, the evaluation of accumulated evidence can vary considerably depending on who is doing the evaluation, and for what specific purposes it is being used.

Within diagnostic validity, the lack of clear guidelines as to how to assess the totality of validating evidence was recognized shortly after its inception and throughout its development. Two changes in the method for achieving diagnostic validity (i.e., Diagnostic Validity-II) may be conceived as aligning with changes within CVT that partially address this issue. The first is Kendler's (1980) introduction of the concept of a "validator" as a type of validity evidence. While the second edition of *Standards* (1985) doesn't use the term "validator" in this capacity, what it does do is reflect the move by general validity theorists who called for the removal of distinctions between separate *types* of validity in favor of types of *evidence*. Whereas *Technical Recommendations* (1954) featured content *validity*, concurrent *validity,* predictive *validity*, and construct *validity*, *Standards* (1985) adopted content-, construct, and criterion-*related evidence*. The use of this terminology in *Standards* dissolved the notion of separate types of validity in

favor of different types of evidence, which supported a unified sense of validity. Similarly, the validator concept by Kendler (1980) organizes validity evidence into different types of validators, which denotes different types of evidence to support a unified sense of diagnostic validity (or *senses*, considering how small-V and big-V validity appear in practice). The result is the capacity to prioritize evidence across different validators without necessarily declaring that one "type" should necessarily be deemed more relevant in all cases.

The second change came in providing a call for specific guidelines by Kendler (1990) to create a more scientifically informed way of evaluating the validators. As it would turn out, some of the "fundamentally nonempirical" aspects that Kendler attributes to psychiatric psychology (e.g., determining which validators should be given priority for supporting a given diagnostic category) may be equally attributed to a feature of the model of psychometric validity, the "weak" program of CVT (e.g., a lack of explicit guidelines for determining which types of validity were most relevant for construct validation). Recommending that validating evidence be "systematically gathered" and "objectively evaluated" may be interpreted as attempts to make Robins and Guze's method for achieving diagnostic validity more empirically informed, and in a sense, attempting to shift diagnostic validity toward being based in a "strong*er*" sense of validation, viz., big-V diagnostic validity, and in a sense, like the "strong" program of CVT.

Evidence of diagnostic validity coming to reflect the "strong" program of CVT appears in the explicit framing of the evaluation of validity evidence in the form of *hypothesis generation and testing* by Kendler (1990)—a process which is *not* explicitly stated in Robins and Guze (1970) or Feighner et al. (1972)—but which is a key feature Cronbach's "strong" program of CVT:

> Thus, a scientific nosology would involve the generation of hypotheses about the reliability and validity of competing diagnostic schémas. These hypotheses would be tested by the examination of the research data that addresses the given hypotheses to determine whether the individual hypothesis (e.g., diagnostic criteria A are more valid than diagnostic criteria B) is or is not supported by the available evidence. (Kendler, 1990, p. 970)

A second aspect in how updated versions of diagnostic validity reflect "strong" CVT comes from the hierarchy and prioritization of the validators being set *in service of* theory building around the focal phenomena of the *DSM*'s diagnostic categories, notably around etiology and pathophysiology. Cooper (2018) notes that while *DSM-III* (1980) set out to be atheoretical and descriptive, in the *DSM-IV* (1994), the aim of supporting an atheoretical classification system was "quietly dropped" (p. 58). The *DSM-5* (2013) would subsequently aim toward a basis in theoretical knowledge concerning etiology. To this end, "Biomarkers*" are the only Concurrent Validator source being designated as a high-priority validator.

A third aspect relates to how diagnostic validity has developed in response to the issue of determining an overall evaluation of the validators. Referred to by Kendler and Solomon (2021) as *the problem of aggregating validators,* it is described as "how the validators should be aggregated to come to an overall conclusion about the strength of the evidence for a psychiatric category" (p. 9). The problem is framed not as an issue with the pragmatic choice of a diagnostic category, but as a problem of "underdetermination," whereby philosophers of science would describe the situation as occurring "when considering evidence for different and competing *theories*, as one of the underdetermination of *theory* by evidence" (p. 10, emphasis added). Under the "weak" program of CVT, theory building plays no immediate part, as this version of CVT was only intended as a steppingstone to future efforts of theory construction. Moreover, the lack of explicit guidance offered in *Technical Recommendation* as to how to evaluate validating

evidence in its totality is not interpreted as a problem of underdetermination at all, but an *intended feature* since, after all, the recommendations were only intended "as a basis for practical judgments rather than solely for research" (*Technical Recommendations,* 1954, p. 4). Therefore, if diagnostic validity were *not* moving toward being more reflective of a "strong" program of CVT, the aggregation of validators would *not* be viewed as a problem, or at least of problem of underdetermination in relation to theory.

2.9 Philosophical Underpinnings

Philosophical considerations regarding the *DSM*'s classificatory framework span a variety of issues and topics within the philosophy of nosology, including but not limited to the concept of mental illness (Wakefield, 1992), the boundary between the normal and abnormal (Lilienfeld and Marino, 1995), the relation between description and diagnoses (Zachar & Kendler, 2017), and the role values in diagnostic classification play (Sadler, 2005). Two philosophical questions that are particularly relevant for carving out diagnostic validity's philosophical backdrop are one, the nature of psychiatric disorders, and two, the nature of nosological progress in psychiatry.

The nature of psychiatric disorders, of which Kendler, Zachar, and Craver (2011) pose the question, "what kinds of things are psychiatric disorders?" (p. 1143) is ultimately a question of the nature of the underlying focal phenomena (i.e., the underlying clinical syndrome) and its relation to its psychiatric classification. Kendler, Zachar, and Craver (2011) consider four models or ways of conceptualizing psychiatric disorders. The first, an *essentialist kind*, views psychiatric disorders as essences, existing in the world independently of how they are classified. The second, a *socially constructed kind,* views the existence of psychiatric disorders as being entirely

dependent on how they are classified, so that they only exist by way of the culture and society that classify them. The third, a *practical kind*, views psychiatric disorders as helpful tools or instruments that can help us meet scientific and clinical aims. The fourth, a model that the authors argue for and that Zachar and Kendler (2017) recommend as an "aspirational goal for a future psychiatric nosology" (p. 57) is the *mechanistic property cluster kind* (MPC kind), which views psychiatric disorders as "sets of symptoms that are connected through a system of causal relations" (Kendler, Zachar, & Craver, 2011, p. 1144).

While *DSM*-based diagnostic categories are arguably much closer to practical kinds and are not readily amenable to fitting or corresponding neatly to MPC kinds, the advocacy for such a model is reflective of the realist aims of the *DSM* specifically, and psychiatry more generally. The *DSM*'s continued emphasis in developing a scientific nosology, as well as considerations of the best models of psychiatric disorders as a way of getting the focal phenomena "right" and thus facilitating the development of psychiatric knowledge, is indicative of the *DSM*'s realist theory-building aims. The *DSM*'s contribution to theory building is no longer conceived as primarily instrumentalist such that it would view the theoretical knowledge to which it contributes as mainly the conceptual or practical refinement of the diagnostic categories as supported by small-V diagnostic validity. Instead, the *DSM* aims for its classification system to constitute and contribute to the development of theoretical knowledge about the underlying real clinical syndromes as they truly exist, so that: "Over time, this process will slowly move our nosology from the rough constructs we now call 'disorders' towards a better and better approximation of the 'true' psychiatric diseases as they exist in nature." (Kendler & First, 2010, p. 263)

The notion of a scientific nosology or natural classification further illustrates a realist position toward the nature of nosological progress in psychiatry. Zachar and Kendler (2017)

emphasize how Kendler and Parnas (2012), adapting from Chang's (2004) concept of epistemic iteration have posited their own iterative model for psychiatry and the *DSM* whereby diagnoses may achieve stability over time so that various forms of evidence will converge on "stable, more scientifically grounded constructs" (Zachar & Kendler, 2017, p. 60). Distinct from Chang's coherentist program, this conception of nosological progress is considered by Chang (2017) to point to a pluralist realism that is more realistic for a scientific psychiatry. This model for nosological progress is further considered by Zachar and Kendler (2017) to share a similar "paradigm of naturalness" with Robins and Guze's method (p. 60).

Now turning directly to diagnostic validity—to what degree does Robins and Guze's method reflect the realist aims of the *DSM*? Right away, the answer is not so simple, for there are two overlapping yet distinct senses of diagnostic validity, each with relatively distinct aims and therefore, arguably different philosophical underpinnings. Thus, properly interpreting the philosophical backdrop of diagnostic validity will require a mix of two contrary philosophies.

Here, we may once again draw on prior work interpreting CVT to help us make sense of the underlying philosophy of diagnostic validity. Social psychologist Joseph E. McGrath (2005) argued that the conceptual complexity surrounding CVT has to do with the fact that psychological measurement and thus construct validation in practice is often centered around two distinctive aims, *prediction* and *representation*. Prediction, which concerns the ability of responses on some psychological measure to be predictive of some "external referent," is argued to occur "in the context of operationism" when the aim is "solely the maximization of an observable relationship, not the potential for intuitive understanding" of the construct (p. 120). As a result, prediction does not require any realist commitments. Representation, in turn, refers to the ability of some measure to accurately represent the construct. With representation, the goal

is not prediction, but the ability to assess responses on a measure as being causally attributed to the underlying construct that can be said to account for a certain performance or standing on the measure. According to McGrath, the aim of representation is "primary for the advancement of empirical knowledge" and is considered far more important to a measure's "scientific value" (p. 112–113). Thus, depending on the distinctive aims, McGrath would argue that construct validation may be viewed as a mix between both operationism and realism.

Drawing on McGrath's interpretation and what we know of the different senses of diagnostic validity, we may interpret the *DSM*'s validation process as maintaining both operationist and realist underpinnings. The aim of small-V diagnostic validity amounts to achievements relating to the predictive utility of the diagnostic categories, and thus its operationist leanings. The aim of big-V diagnostic validity amounts to successfully representing the focal phenomena, and requires a philosophy based in scientific realism.

Conclusion

In this chapter, I have argued that diagnostic validity, an application-specific validation process first articulated by Robins and Guze (1970) and subsequently developed within and alongside the various iterations of the *DSM* and in the context of clinical medicine and psychiatry, may be thought of as comprising two relatively distinct yet overlapping senses of validity—small-V diagnostic validity and big-V diagnostic validity. The former is intended to support the degree of confidence we have in a diagnostic category for its ability to serve as a measurement-like instrument that can offer useful predictions, while the latter is concerned with the ability of a category to accurately stand for or represent a real underlying disease entity. Now more than ever, diagnostic validity has been refined and expanded to support the development of

a scientific nosology, with the hope that sufficient validating evidence discovered via the *DSM*'s iterative and continuous improvement approaches may eventually help us out of the box canyon and back onto a path of nosological progress.

While the *DSM* remains relatively patient, others would prefer not to wait, instead asking, when, if ever, will we know enough to warrant a new diagnostic paradigm to get us out the canyon? In the next chapter, we turn toward the first alternative approach to psychiatric classification, the Hierarchical Taxonomy of Psychopathology, which, in departing from the *DSM*, has set its sights on researching, developing, and validating psychiatric classifications in a whole new way.

References for Chapter 2

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2, Pt.2), 1–38.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychiatric Association. (1952). *Diagnostic and statistical manual of mental disorders* (1st ed.).

American Psychiatric Association. (1952). *Diagnostic and statistical manual of mental disorders* (2nd ed.).

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.).

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., revised).

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.).

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.).

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.).

American Psychiatric Association. (2021). Guide to submitting proposals for changes to *DSM-5*. https://www.psychiatry.org/File%20Library/Psychiatrists/Practice/DSM/DSM5-Proposal-Submissions-General-Guidance.pdf.

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.).

American Psychiatric Association (2023). *Submit Proposals for Making Changes to DSM-5-TR*. https://www.psychiatry.org/psychiatrists/practice/dsm/submit-proposals

Andreasen, N. C. (1995). The validation of psychiatric diagnosis: New models and approaches. *The American journal of psychiatry*, *152*(2), 161–162.

Aragona, M. (2015. Rethinking received views on the history of psychiatric nosology: Minor shifts, major continuities. In P. Zachar, D. S. Stoyanov, M. Aragona, & A. Jablensky (Eds.), *Alternative perspectives on psychiatric validation: DSM, IRC, RDoC, and beyond* (pp. 27–46). Oxford University Press.

Bentall, R. P. (2003). *Madness explained: Psychosis and human nature*. Penguin UK.

Bridgman, P. W. (1927). *The logic of modern physics* (Vol. 3). Macmillan.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.

Chang, H. (2017). Epistemic iteration and natural kinds: Realism and pluralism in taxonomy. In K. S. Kendler & J. Parnas (Eds.), *Philosophical issues in psychiatry IV: Psychiatric nosology* (229–245). Oxford University Press.

Cronbach, L. J. (1980). Selection theory for a political world. *Public personnel management*, *9*(1), 37–50.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum Associates, Inc.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), Intelligence: Measurement, theory, and public policy (pp. 147–171). University of Illinois Press.

Cooper, R. (2018). Understanding the *DSM-5*: Stasis and change. *History of Psychiatry*, *29*(1), 49–65.

Decker, H. S. (2007). How Kraepelinian was Kraepelin? How Kraepelinian are the neo-Kraepelinians? —from Emil Kraepelin to *DSM-III*. *History of Psychiatry*, *18*(71 Pt 3), 337–360.

First, M. B., Regier, D. A., & Kupfer, D. J. (2002). *A research agenda for DSM-V*. American Psychiatric Pub.

Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of general psychiatry*, *26*(1), 57–63.

Hartman, C. A., Hox, J., Mellenbergh, G. J., Boyle, M. H., Offord, D. R., Racine, Y., ... & Sergeant, J. A. (2001). *DSM-IV* internal construct validity: When a taxonomy meets data. *The journal of child psychology and psychiatry and allied disciplines*, *42*(6), 817–836.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of educational Measurement*, *38*(4), 319–342.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kane, M. (2013). The argument-based approach to validation. *School psychology review*, *42*(4), 448–457.

Kendell, R. E. (2002). Five criteria for an improved taxonomy of mental disorders. In J. E. Helzer & J. J. Hudziak (Eds.), *Defining psychopathology in the 21st century: DSM-V and beyond*, (pp. 3–17). American Psychiatric Publishing, Inc.

Kendler, K. S. (1980). The nosologic validity of paranoia (simple delusional disorder): A review. *Archives of general psychiatry*, *37*(6), 699–706.

Kendler, K. S. (1988). Validators. Presented at the American Psychiatric Association *DSM-IV* Methods and Applications Conference; September 29, 1988; Washington, DC.

Kendler, K. S. (1990). Toward a scientific psychiatric nosology: Strengths and limitations. *Archives of general psychiatry*, *47*(10), 969–973.

Kendler, K.S., Kupfer, D., Narrow, W., Phillips, K., & Fawcett, J. (2009). Guidelines for making changes to *DSM-V*. Unpublished manuscript.

Kendler, K. S., & First, M. B. (2010). Alternative futures for the *DSM* revision process: Iteration v. paradigm shift. *The British journal of psychiatry*, *197*(4), 263–265.

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

Kendler, K. S. (2012). Epistemic iteration as a historical model for psychiatric nosology: Promises and limitations. In K. S. Kendler & J. Parnas (Eds.), *Philosophical issues in psychiatry II: Nosology* (pp. 305–322). Oxford Academic.

Kendler, K. S. (2014). *DSM* issues: incorporation of biological tests, avoidance of reification, and an approach to the "box canyon problem". *American Journal of Psychiatry*, *171*(12), 1248–1250.

Lilienfeld, S. O., & Marino, L. (1995). Mental disorder as a Roschian concept: A critique of Wakefield's "harmful dysfunction" analysis. *Journal of abnormal psychology*, *104*(3) 411–420.

McGrath, R. E. (2005). Conceptual complexity and construct validity. *Journal of personality assessment*, *85*(2), 112–124.

Phillips, J. (2013). The conceptual status of *DSM-5* diagnoses. In In J. Paris & J. Phillips (Eds.), *Making the DSM-5: Concepts and controversies* (pp. 143–157). Springer.

Regier, D. A., Narrow, W. E., Kuhl, E. A., & Kupfer, D. J. (2009). The conceptual development of *DSM-V*. *American Journal of Psychiatry*, *166*(6), 645–650.

Robins, E., & Guze, S. B. (1970). Establishment of diagnostic validity in psychiatric illness: Its application to schizophrenia. *American journal of psychiatry*, *126*(7), 983–987.

Robins, L. N., & Barrett, J. E. (1989). Establishment of diagnostic validity in psychiatric illness: Robins and Guze's method revisited. *The validity of psychiatric diagnosis*, *172*, 9.

Rodrigues, A. C. T., & Banzato, C. E. M. (2015). Reality and utility unbound: An argument for dual-track nosologic validation. In P. Zachar, D. S. Stoyanov, M. Aragona, & A. Jablensky (Eds.), *Alternative perspectives on psychiatric validation: DSM, IRC, RDoC, and beyond* (pp. 47–59). Oxford University Press.

Sadler, J. Z. (2005). *Values and psychiatric diagnosis*. Oxford University Press.

Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. Springer.

Solomon, M., & Kendler, K. S. (2021). The problem of aggregating validators for psychiatric disorders. *The journal of nervous and mental disease*, *209*(1), 9–12.

Solomon, M. (2022). On validators for psychiatric categories. *Philosophy of medicine*, *3*(1), 1–23.

Spitzer, R. L. (2001). Values and assumptions in the development of *DSM-III* and *DSM-III-R*: An insider's perspective and a belated response to Sadler, Hulgus, and Agich's "On values in recent American psychiatric classification". *The journal of nervous and mental disease*, *189*(6), 351–359.

Wakefield, J. C. (1992). The concept of mental disorder: On the boundary between biological facts and social values. *American psychologist*, *47*(3), 373.

Willcutt, E. G., & Carlson, C. L. (2005). The diagnostic validity of attention-deficit/hyperactivity disorder. *Clinical neuroscience research*, *5*(5–6), 219–232.

Zachar, P. (2012). Progress and the calibration of scientific constructs: the role of comparative validity. In K. S. Kendler & J. Parnas (Eds.), *Philosophical issues in psychiatry II: Nosology* (pp. 21–40). Oxford Academic.

Zachar, P., & Kendler, K. S. (2017). The philosophy of nosology. *Annual review of clinical psychology*, *13*, 49–71

Zachar, P., First, M. B., & Kendler, K. S. (2022). Revising substance-related disorders in the *DSM-5*: A history. *Journal of studies on alcohol and drugs*, *83*(1), 99–105.

Chapter 3: Psychometric Validity I

> *Is it really sensible to work with tests designed for the purpose of assigning individuals to psychiatric categories when these categories are largely arbitrary, have no scientific status, cannot be reliably assessed, and contradict in their very conception the strong evidence pointing in the direction of a dimensional rather than a categorical system of measurement (Eysenck 1970)? Should we not, as independent scientists, work out a system of classification based on empirical evidence, psychological theory, and experimental support, rather than accept more or less blindly a medical system whose only virtue (if that be the right term) seems to be that it is based on some form of consensus*.
>
> – Hans Eysench, Jerome Wakefield, and Alan Friedman (1983) in *Diagnosis and Clinical Assessment: The DSM-III*

3.1 Introduction

Referring to the *DSM*'s diagnostic categories as those with "no scientific status" (Eysench, Wakefield, & Friedman (1983), p. 168) was a radical claim when made in the early 1980s. Such a statement regarding psychiatry's most scientifically informed classification system to date reflected a minority viewpoint of psychologists and psychiatrists who were dissatisfied with the *DSM* being in their perspective more of a professional manual and "only secondarily, if at all, a scientific manual" (p. 184). their two main concerns were the *DSM*'s lack of validity due to its failure to include dimensional measures of psychopathology, along with keeping professional consensus as a criterion for inclusion of a diagnostic category in the *DSM*. Asserting that diagnostic categories were based "on foundations so insecure, so lacking in scientific support, and so contrary to well-established facts" (p. 189), Hans Eysench, Jerome Wakefield, and Alan Friedman (1983) argued that a significant overhaul to the revision process for the upcoming edition of the *DSM* would be required if there were to be any hope of bringing about a scientific nosology.

Somewhere between the publication of *DSM-IV-TR* (2000) and *DSM-5* (2013), when the realization hit that such an overhaul was never coming, what started as a minority, radical view regarding the *DSM*'s poor scientific standing became the norm. More and more, critics were rejecting the categorical model on which the *DSM* is based, judging it to be incompatible with the empirical observations of the continuous variations in psychopathology evident in their research. Given this incompatibility, many adopted the point of view that no amount of revision to the *DSM* could be of any help—that we are destined to remain forever in a validity box canyon. To move psychiatry toward a true scientific nosology with valid psychiatric classifications, a new dimensional paradigm to break up the existing diagnostic categories was needed (Krueger & Piasecki, 2002). In 2017, an application of such a paradigm was officially introduced that would pose the first real challenge to the *DSM* as a potential alternative classification system, referred to as the Hierarchical Taxonomy of Psychopathology, or HiTOP for short. (Kotov et al., 2017).

HiTOP, described as an empirical quantitative approach to psychopathology, is a paradigm shift model of nosological change (Kendler & First, 2010) in that it outright rejects the *DSM*'s iterative model and calls for a wholesale replacement of its current classificatory paradigm. Whereas the *DSM*'s development of its psychiatric classifications begins with the clinical description of signs and symptoms grouped into operationally defined, homogeneous diagnostic categories by which all other validating evidence is directed, HiTOP in turn clears the table and starts instead by quantitatively deriving theoretically pure dimensional constructs it considers to be the underlying structure of psychopathology. HiTOP identifies its key constructs by factor analysis, a group of statistical techniques that can be traced back to psychometric research of the early 20th century which seeks to reduce an initial set of variables (e.g., observed

symptoms) gathered via empirical instruments and tests such as surveys and questionnaires down to a smaller set of latent (unobserved) hypothetical variables, called factors, which point toward constructs that underly the observed variables. This method of quantitative classification, according to the HiTOP consortium, more accurately models the natural underlying structure of psychiatric disorders than the homogenous clinical syndromes of the *DSM*, which may provide a better starting point for more scientifically accurate and clinically useful research in and development of psychiatric classifications.

In this chapter, I describe and evaluate HiTOP's sense of validity based in *psychometric validity*, a type of validity associated with the validation of psychological instruments or tests designed to measure psychological constructs. I begin with an overview of HiTOP's application-specific sense of psychometric validity which I term *Structure-First Psychometric Validity* (SFPV). I then trace its historical foundations to early psychometrics research, the formation of construct validity theory (CVT), and scientific practice in psychology. Next, I analyze some of SFPV's refinements, in which I point toward HiTOP's recent attempts to advance SFPV beyond the validity of individual HiTOP constructs, and toward establishing the validity of the entire HiTOP hierarchy. I further distinguish this refined sense of SFPV from specific senses of CVT within psychometric validity and conclude with a reflection of some potential tensions within its philosophical foundations.

By offering a thorough reconstruction, I aim to show that SFPV doesn't just offer a different path out of the box canyon toward achieving diagnostic validity, but instead offers a specific conception as to what should count as validity in psychiatry based in a hybrid set of validation standards. I argue that such standards, in their present form, would benefit from

further specification regarding validation of a HiTOP construct vs. validation of the HiTOP hierarchy, and the concept of scientific accuracy.

3.2 Overview of Structure-First Psychometric Validity

HiTOP's specific account of validity resides within the general framework of psychometric validity*,* broadly understood as the degree to which a test, being a response to a standardized situation devised to measure a psychological construct has the desired psychometric features (e.g., various senses of validity, reliability, utility, etc.). SPFV is a comprehensive, systematized, and application-specific validity process centered on supporting the development of psychiatric classifications that carry both *scientific accuracy*, understood as the degree to which HiTOP's constructs represent true features of psychopathology, and *clinical utility*, being the degree to which classifications are considered practically and pragmatically useful in the clinic. Whereas the *DSM* understands "utility" as being one of several reasons for adding or subtracting specific criteria from a diagnostic category and thus views utility as being a part of a diagnostic category, HiTOP rejects this conceptualization. HiTOP instead maintains that utility should not be part of the definition of a HiTOP construct, and instead holds that utility follows from the correct definition, i.e., the scientific accuracy, of the construct. As a result, scientific accuracy is not only heavily prioritized by HiTOP but seemingly equated with the very concept of validity, so that "establishing the scientific accuracy of psychiatric classification systems is essentially the task of establishing their validity" (Forbes et al., 2023, p. 12). Thus, for SVPF, validity can be thought of as the degree to which validating evidence supports the empirical quantitative measurement of a psychiatric classification to establish scientific accuracy.

Unlike the *DSM*, SFPV is based in construct validity—the unified sense of validity that encompasses all contemporary psychometric validity—as it is HiTOP's dimensional constructs, quantitatively derived from psychological testing data, that serve as the basis of its psychiatric classifications and that are measured and subsequently validated. For HiTOP, a construct most closely follows Colman's (2006) definition as a "conjectured entity, process, or event that is not observed directly but is assumed to explain an observable phenomenon. It is not merely a summary of the relationship between observable variables but contains surplus meaning over and above such relationships" (p. 359). HiTOP constructs are understood within a reflective or common factor model, in which such constructs represent coherent and distinct underlying dimensions that hold a shared causal influence on a set of indicator (observed) variables, being the symptoms and/or symptom groupings of psychopathology. HiTOP constructs are latent (unobserved) variables that are assumed to explain the shared variance (i.e., correlation) between the indicator variables of interest. They adhere to a common cause model within psychopathology, which assumes that the underlying factor (the latent variable) causes the psychopathological symptoms (indicator variables). The symptoms, representing fallible indicators of the latent variable, are considered reflective when they are shown to vary with changes within the latent variable, the construct of interest.

HiTOP constructs are organized hierarchically and at various levels. SFPV, however, operates exclusively at the level of the HiTOP *spectra*, the hallmark organizing and psychometrically validated components of the HiTOP model. Spectra represent the main common factors and thus the core structure underlying all major forms of psychopathology. To date, there are six spectra empirically derived via first-order confirmatory factor analysis, including internalizing (or negative affectivity), thought disorder (or psychoticism), disinhibited

externalizing, antagonistic externalizing, detachment, and somatoform. The remainder of the HiTOP model includes narrow constructs populated below the spectra, as well as higher-order constructs populated above the spectra. Constructs lower in the hierarchy are derived from empirical measures of symptom data "not designed for structural research" (Kotov et al., 2017, p 15), whereas constructs one level below or one (to two) levels above the spectra are derived from 2nd (or 3rd) order factor analysis.

At the lowest level of the model are *Symptom*s, i.e., individual symptoms, signs, and maladaptive behaviors associated with psychiatric disorders. Symptom-level data inform the next level, *Homogeneous Symptom Components and Maladaptive Traits*, which are closely related groupings of individual observable behavioral traits into Symptom Components (e.g., disassociation, numbing, pure obsessions) and closely related groupings of Maladaptive Traits (specific pathological personality characteristics, e.g., withdrawal, (low) attention seeking, disaffiliativeness). Next, components and traits are grouped under *Syndromes*, defined as "composites of related components/traits" (e.g., social anxiety, substance-related disorders) that indicate a dimension as opposed to a category (Kotev et al., 2017, p. 4). At present, *DSM*-based disorders are not included in the Empirical Syndromes level (e.g., GAD, Anorexia nervosa), but symptoms and signs that constitute them are included. Closely related syndromes are grouped at the next level, *Subfactors,* e.g., "distress" includes depression, generalized anxiety, dysthymia, post-traumatic stress, and borderline personality disorder. Constellations of *Subfactors* make up the *Spectra*. At the highest-order levels of the HiTOP model, the six spectra are combined into broader *Super-Spectra*, representing higher-order dimensions of psychopathology that cut across psychiatric disorders (e.g., "Emotional dysfunction" as an aggregated super-spectra for both

somatoform and internalizing spectra), with the "p-factor" representing a general factor that is evident across all of psychopathology (Figure 2, Appendix).

Whether the general factor of psychopathology (p-factor) truly represents an underlying latent variable consistent with a common cause explanation for all of psychopathology has been contended (e.g., see Watts et al., 2023 for an extensive critical evaluation of the p-factor literature). For example, disagreements within the modeling of the general factor exist concerning technical recommendations for model comparison, as well as findings that the general factors do not exhibit uniform influence on the factors below it in the way one would expect in a hierarchical model (Kreuger et al., 2018). There is also currently a lack of specificity as to just what is required to establish causal influence, uncertainty as to the philosophical backdrop of causation assumed of the models, and debate as to whether such models truly necessitate a common cause interpretation. Nevertheless, HiTOP maintains a general assumption that potential sources for the observed relationship between indicator variables, as represented by its lowest level constructs up to the general factor, point to latent variables that causally influence the development of psychopathology.

Validation of HiTOP's spectra proceeds in three stages, as detailed in HiTOP's three Workgroup and Sub-Workgroup expert reviews (Kotov et al., 2020; Krueger et al., 2021; Watson et al., 2022). The first and by far most important stage, which HiTOP "has expressly prioritized….as a foundation" (Forbes et al., 2023, p. 9) focuses on establishing *structural validity*. Structural validity, characterized by Messick (1989) as the "Structural Component of Construct Validity" reflects an evaluation of how well a particular construct accounts for "the observed pattern of individual differences in item performance" based on "correlational structures of item responses and their interpretation in construct theory" (Messick, 1989, pp.

68).The current *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) has since adopted and interpreted this source of validity evidence as Evidence Based on Internal Structure. Along these same lines, HiTOP views structural validity as the degree to which a particular construct accounts for the empirically observed covariance (i.e., the direct relationship) between different signs and symptoms of psychopathology. For HiTOP, the central research question is, which symptoms tend to co-occur within the same people? The basic premise is if two things tend to co-occur, then this increases the likelihood that they share some underlying cause. So, information on the covariance structure provides information on the likely causal structure. Structural validity for HiTOP is the foundation for supporting claims about "how constructs are formed, how they relate to each other, and how they are distributed" (Forbes et al., 2023, p. 11). Hence, validation for HiTOP is structure-first.

Quantitatively derived, structural validity is interpreted as the degree to which the accumulated *structural evidence* supports patterns and associations of the features of psychopathology, i.e., the structure of psychopathology, which is required for understanding the causal origins of those features. Structural evidence is understood as the total assortment of empirical studies of co-occurrence drawn on to assess and establish the structural validity of a particular construct. Supporting evidence from these studies shows empirically observed covariance structures (i.e., the structure) that conform to predictions of the construct of interest. This evidence is based primarily in exploratory and confirmatory factor-analytic research with a preference for continuous latent variable models. Such models produce factor loadings, i.e., the correlation between the original variables and an underlying factor (the construct), which are assessed based on the goodness of fit of the model to the data. These quantitatively derived factors are drawn upon to support the existence and contribution of a construct's *coherence.* A

construct's coherence is understood as the degree to which a construct may be viewed as a *pure* construct, meaning the construct is more or less the same thing (e.g., "Detachment" spectra is *just* Detachment, not "Thought Disorder," "Internalizing," or any other factor). Coherence is understood as a highly desired property or value of structural validity. For HiTOP, only a construct with a high degree of structural validity may be regarded as coherent.

Once a construct has been shown to demonstrate a sufficient degree of structural validity and thus internal coherence, SFPV proceeds to its second stage, which is an evaluation of the *external validity* of the construct. External validity refers to the degree to which evidence for a certain construct correlates with other (relevant) indicators of that construct. For HiTOP, the assumption is that structural evidence should correlate with evidence from measures of other external criteria, thereby establishing validity "through the patterns of association with constructs external to the descriptive system" (Forbes et al., 2023, p. 11). Constructs are considered only to be hypothesized when they have yet to be subjected to external validating evidence. Strong external validity evidence is interpreted to support the coherence of constructs and a hierarchical conceptualization initially supported via structural evidence. Weak external validity evidence may reveal shortcomings in the original structure and motivate a reevaluation of the structural research. While various sources of external validating evidence have been differentiated in psychopathology research, HiTOP considers five sources of validating evidence assumed to be part of external validity to be essential (Table 6, Appendix).

Following an evaluation of sufficient external validating evidence that is deemed to be congruent with the structural evidence of the construct (Watson et al., 2022), an evaluation of evidence collected based on a construct's *reliability* and *utility* is assessed. Reliability, generally understood as a statistical measure used to assess the consistency and reproducibility of results

obtained from measuring a construct repeatedly, is typically distinguished as being a separate and unrelated property or process from validity. For example, to evaluate a diagnostic category, the *DSM* considers the evidence for *inter-reliability* (i.e., the degree of agreement between multiple raters) as distinct from its validity (Kendell & Jablensky, 2003). HiTOP, in contrast, recognizes that some measures of reliability have implications for validity. For example, *internal consistency reliability* is a reliability measure based on the extent to which all items on a test measure the same construct. Insofar as HiTOP is interested in coherent constructs that require that all items on a test measure the same construct, then reliability may be considered a relevant validity concept for SFPV.

Utility for HiTOP is divided into two parts—*clinical utility* and *predictive utility*. Clinical utility, more specifically beyond that which is clinically useful, refers to the degree to which a construct is "helpful in organizing clinical assessment, selecting treatment options, and communicating the nature of the problem(s) to patients and others involved in their health care" (Forbes et al., 2023, p. 7). Predictive utility, a concept notably utilized among personality psychologists and considered distinct from clinical utility, reflects the degree to which a construct is helpful in differentially predicting outcomes of interest. For example, showing that a construct within the HiTOP model (e.g., the general factor of psychology or p-factor super-spectra) predicts future suicide and self-harm (outcomes of interest) would be an indication of predictive utility. Along with reliability and clinical utility, the predictive utility of a particular construct is also thought to be in part determined by the validity of the construct. For HiTOP, a construct that lacks the necessary structural and external validity evidence within the HiTOP model would necessarily lead to suboptimal levels of reliability and utility.

In sum, SFPV is an ongoing three-stage validation process based in psychometric validity, with a strong foundation in structural validity, with each stage supporting the next (e.g., a construct with strong structural evidence should predict and support the existence of strong external evidence; a construct with both strong structural and external evidence should predict and support strong reliability and utility evidence).

To get a better sense of how SFPV is based in psychometric validity and thus distinct from diagnostic validity, I now turn to the origins of SFPV, beginning with how psychometric validity led to the development of construct validity theory (CVT). I then introduce SFPV in relation to CVT as well as additional standards of validity SFPV draws from other areas and extend SFPV's account with recent refinements from the HiTOP consortium.

3.3 Origins of Psychometric Validity and the Development of Construct Validity

Psychometric validity and related concepts such as reliability as featured within SFPV developed separately from that of diagnostic validity out of psychometrics, broadly understood as a scientific research tradition "that concerns itself with the study of measurement and human behavior" (Wijsen, 2021). The initial inception of the psychometrics discipline is typically traced to the introduction of the common factor model of general intelligence by Charles Spearman (1904), in which the relationship between observed variables, e.g., performance on certain tests, may be predicted by a common latent (underlying) variable, general intelligence, or simply "g." Shortly thereafter, the concepts of reliability, validity, and validation were frequently discussed in the writings of psychometricians during the 1910s and 1920s, as this period saw an influx of mental testers developing tests for assessing psychological attributes (Newton & Shaw, 2014). In 1927, psychometrician Truman Kelley developed a formal account of validity, now commonly

referred to as test validity, in which validity within a psychometric tradition is conceived as "whether a test measures what it purports to measure" (p.14).

In this regard, psychometric validity in its infancy was very much considered a property of the test. The measurement of a psychological attribute via a test (or battery of tests) would be the test taker's (aggregate) responses or scores on the test(s). To assess statistical tendencies among the test takers, the tester (e.g., the experimental psychologist) would necessarily have to choose specific attributes by which all test takers could be quantitatively compared while subsequently leaving certain factors out, meaning that "psychometrics, in other words, cannot deal with differences between individuals, but only with differences between performances in tests." (Vernon, 1933, p. 163). Nevertheless, psychometricians may be justified in their descriptive conclusions drawn from analysis of their testing data so long as the tests could be validated "by the ubiquitous correlational methods" during the final phases of a test's construction (p. 163). Given the prevalence of tests of intelligence and other psychological attributes at the time, both reliability and validity under the psychometric tradition were conceived as crucial steps in supporting any statistical conclusions made from the analysis of testing data (Wijsen, 2021). Moreover, tensions regarding the reliability and validity of such tests were reflective of early discussions regarding standards for psychological testing.

The concept of validity within psychometrics over the next thirty years centered on how accurate a test was in estimating or predicting the criterion value of an existing variable of interest. Referred to as the criterion (concurrent and predictive) model of test validity, this model evaluated validity based on how well scores on a test compared to the accepted criterion scores, of which the test was designed to estimate or predict. An often-cited example of criterion validity used in contemporary educational measurement is how accurate a college entrance exam (the

test) is at estimating and predicting academic success, with students' college GPA as an accepted criterion measure. With the general notion of validity being the degree to which a test measures what it purports to measure (Buckingham, 1921) and "how well a test does the job it is employed to do" (Cureton, 1951, the criterion model of validity was considered preferred model of validation at that time for tests that may be readily compared to a "uniquely pertinent" i.e. real criterion (Cronbach, 1971, p. 462).

But what about test cases for which there is no adequate, well-defined criterion measure that exists to estimate or predict? And what if the tester does not only want to know how valid their test is but wishes to gain a deeper understanding of the internal processes they think may be causally relevant in explaining the scores on the test itself (i.e., the correctness of the underlying constructs)? Given a consensus that validity evidence provided by the criterion model was insufficient in meeting the validation needs of psychological testing, in 1954 the American Psychological Association Committee on Psychological Tests decided to expand its conception of validity to include additional validity evidence. In the resulting *Technical Recommendations for Psychological Test and Diagnostic Techniques*, the Committee outlined recommendations that were "necessarily of a psychometric nature" (p. 204) and came to a working definition and set of standards for construct validity, a sense of validity whose initial conception and underlying philosophy of science is attributed to logical empiricist Herbert Feigl (1943; 1950). Construct validity was further elaborated by APA committee member Paul Meehl (1920–2003), a former student, colleague, collaborator, and friend of Fiegl's at the University of Minnesota. In *Technical Recommendations*, construct validity was to be evaluated "by investigating what psychological qualities a test measures, i.e., by demonstrating that certain explanatory constructs account to some degree for performance on the test" (APA, AERA, & NCME, 1954, p. 214).

From its formal introduction in *Technical Recommendations*, the authors were aware that the various types of validity (content, predictive, concurrent, and construct) all ultimately interact and contribute to the establishment of construct validity, so that "to analyze construct validity, our total background of knowledge regarding validity would be brought to bear" (APA, AERA, & NCME, 1954, p. 216). Construct validation was conceived as a two-step process whereby a test user would first make predictions as to the variation of scores on a test that may be (indirectly) attributed to an underlying theory (or construct), and second, gather empirical evidence to confirm those predictions. Construct validation necessitated "integrating evidence from many different sources" (APA, AERA, & NCME, 1954, p. 214), including setting up experiments that would utilize the test, correlations with other types of (indirect) measures, and psychometric techniques such as factor analysis. Moreover, construct validation was conceived as a process by which both the test and the underlying hypothesis regarding the construct of interest (e.g., a certain personality organization) are validated simultaneously; as the test is validated, so is the underlying construct of the test validated. This theory-building conception of construct validity was explicated by Meehl and then APA chairman Lee Cronbach (1916–2001), and would later be referred to by Meehl as the "strong" version of construct validity referenced in chapter 2 (Cronbach, 1988).

Cronbach and Meehl's (1955) "strong" version of construct validity centered on theoretical constructs defined exclusively in terms of formal theories. A construct, as later defined in a subsequent edition of *Standards,* was conceived as "an idea developed or 'constructed' as a work of informed, scientific imagination; that is, a theoretical idea developed to explain and organize some aspect of existing knowledge" (APA, AERA, & NCME, 1974, p. 29). Since psychological constructs like "anxiety," "aptitude," "intelligence," etc. do not have a

readily available criterion or sampling domain from which to test, their method instead proposed to validate the construct against some posited scientific theory that implicitly defined the construct in specific relation to other constructs in a network (the nomological network). Construct validation under the strong version of construct validity amounted to stating the theoretical assumptions and conclusions regarding one's constructs, and then subjecting them to empirical tests. If empirical evidence supported the relations of the construct with other constructs as stipulated by some scientific theory, that construct would then be considered well-validated.

The notion of theory adopted for construct validity was that of the Hypothetico-Deductive (HD) model of theories consistent with a logical positivist framework. Under the HD model, theories were axiomatic systems of connected theoretical constructs that may relate to observable variables via correspondence rules. Theories under a positive framework were used to test predictions about observable relationships, i.e., to test empirical laws among the theoretical constructs within a specific nomological network (the explicit theory) without needing to refer to anything outside of the theory itself. The entire nomological network consisted of the axiomatic system as well as the empirical laws derived from and explained by it (Hempel, 1965). Thus, in this particularly insightful formulation of construct validation, Cronbach and Meehl proposed that validation of psychological constructs could be conducted entirely in terms of how well the empirical evidence satisfies a given theory. So long as there are explicit theories (nomological networks) in psychology to test and empirical laws to be explained by such theories, observations consistent with the theory may be interpreted as providing validating evidence for the theory and the measurement procedures used to estimate and predict psychological constructs.

The strong version of construct validity, while considered the ideal form, was ultimately not sustainable following the realization that there were very few (if any) highly developed formal theories in psychology and the social sciences more broadly from which to test, nor were there empirical laws to be derived, and neither were likely to emerge in the coming years. Thus, this strict sense of construct validation that required the existence of nomological networks had very limited utility. To allow for more applicability, some general guidelines were maintained, but the strict reliance on theory was relaxed, leading to criticism from Cronbach as seeing a potential for construct validity to devolve toward "sheer exploratory empiricism" (Cronbach, 1988, p. 12).

Since its initial conception, construct validity, termed construct validity theory (CVT) by Slaney (2017) in a review of its conceptual history, underwent several developments, iterations, and competing interpretations between 1955 and 1989. It was first conceptualized as an additional tool in the validity tool kit alongside criterion, concurrent, and predictive validities, then ultimately followed the work of Loevinger (1957) and Messick (1989). At this point, it became the basis for "a general approach to validity that includes all evidence for validity, including content and criterion evidence, reliability, and the wide range of methods associated with theory testing" (Kane, 2001, p. 324). This approach shifted validity from the validation of the instrument and underlying construct(s) to the validation of a specified and purposeful interpretation of the instrument and underlying construct(s). The current conception of construct validity, updated by Samuel Messick (1987, 1989), rejects the focus on formal theory testing by instead emphasizing interpretive inference, in which "what is to be validated is not the test as such, but the inferences derived from test scores" (Messick, 1987, p. 1) and is defined as "the

degree to which multiple lines of evidence are consonant with the inference" (Messick, 1989, p. 13).

Kane (2001) identified four specific features in Messick's conception of construct validity. First, construct validity centers on an evaluation of the overall plausibility of a proposed interpretation of a test. Second, interpretations include analysis of inferences and assumptions with considerations of the reasoning behind the interpretation as well as alternative interpretations. Third, validation may include considerations of unintended or negative consequences of the use of some proposed interpretation of a test. Fourth and most importantly, construct validity has developed as a unified concept, encompassing all other validity forms, so that the general approach and criteria for judging validation efforts, that is "an extended analysis of evidence, based on an explicit statement of the proposed interpretation, and involving consideration of competing interpretations" (p. 339) applies to all forms and validation methods.

In identifying these features, Kane (1992; 1994; 2001) proposed the most recent refinement of construct validity from interpretive inferences to what he terms an *interpretive argument* that was adopted by the *Standards* (2014). The interpretive argument for Kane bridges some of the general guidelines for validity originally stipulated by Cronbach with the focus on the interpretation of the inferences by Messick. Furthermore, it plainly spells out all possible inferences from the assessment to the proposed interpretation of the assessment, as well as any actions or use of the assessment based on the proposed interpretation. The supposed benefit is in guiding the researcher in carrying out their validation efforts in a systematized fashion and may provide a method of evaluating the progress made by validation efforts.

3.4 Origins of Structure-First Psychometric Validity

Whereas diagnostic validity developed within the context of clinical medicine, SFPV developed separately out of CVT within psychometrics and the standards for psychological testing, as well as validity as featured in scientific practice. To the former, many facets of SFPV appear to reflect current conceptions of construct validity as introduced by Messick (1989) and further articulated by the most recent edition of *Standards* (2014). First, both understand validity as a unitary phenomenon, so that "the varieties of evidence are not alternatives but rather supplements to one another (Messick, 1989, p. 26). SFPV similarly does not distinguish separate kinds of validity, and instead conceptualizes validity as different forms of evidence (e.g., structural evidence, external validity evidence, evidence for reliability, evidence for utility), although HiTOP appears to prefer certain forms of evidence over others. Second, both see validity as a matter of degree rather than an all-or-nothing feature of a construct. A construct is not simply valid or invalid, but instead, the validity of a construct is supported to varying degrees based on the evidence. Accordingly, the process of validation for SFPV as reflected by the current standards for CVT is ongoing and "involves accumulating relevant evidence to provide a sound scientific basis" (AERA, APA 2014, p. 11) for determining the degree to which a construct may be valid. In this regard, the general conceptions of validity and validation have developed in part out of current conceptions of CVT.

Where SFPV puts a twist on contemporary CVT is in its placement of structural validity evidence as the foundation for assessing and establishing validity. There are two potential sources for this development. The first traces back to Blashfield and Draguns' (1976) theory of psychiatric classification, which publications from the HiTOP consortium cite as informing of SFPV (Forbes et al, 2023). HiTOP interprets Blashfield and Draguns' position that for a theory of psychiatric classification to be descriptively effective, it requires structural validity, which

Blashfield and Draguns originally called *descriptive validity*. To achieve descriptive validity, a classification system or taxonomy should utilize "new statistical methods for generating classifications" (p. 374) to provide a form of structural evidence for the structure of psychopathology. Subsequently, for a system of classification to be predictively effective, HiTOP requires "validity through the patterns of associations with constructs external to the descriptive system" (Forbes et al., 2023, p. 11), i.e., external validity, which HiTOP denotes as "Validity Evidence" in review papers following a presentation of the "Structural Evidence" from which quantitative methods have been employed (e.g., Kotov et al., 2020).

In addition to informing views on standards for structural and external evidence, Blashfield and Draguns' theory relates to SFPV in two important ways. First, they viewed systems of psychiatric classification as not just diagnostic tools for social communication or information retrieval (both criticisms of current classifications of the time) but that they support the scientific conceptualization of such classifications in their ability to describe and predict. To this end, Blashfield and Draguns saw the ability to increase the utility of a classification system was directly related to the testability of its constructs, implying the potential for utility as a type of validity concept. Second, the emphasis on an empirical approach to classification in their theory, and a criticism of the political power and social influence, i.e., non-empirical aspects, seen in the current systems of classification, is reflected in SFPV's noted lack of consideration of unintended or negative consequences for the use of a particular construct as being a conceivable part of validity. In other words, whether a particular construct would be used appropriately in a classification system is not considered a form of validity evidence in the sense that it would not contribute toward the descriptive and predictive aims of the classification system.

Returning to the importance of structural validity as the foundation of SFPV, the second source for HiTOP's emphasis on structural validity bears out of the methodological approaches found in personality and intelligence research. Examples include Raymond Cattell's (1943) Sixteen Personality Factor Questionnaire (16PF), Donald Fiske's (1949) analysis of Cattell's variables into what would become the Big Five personality factors, and Hans Eysenck's (1963) three-factor model of personality, all of which were developed based on factor analytic-research used to derive the basic structural elements of personality. Such research reflects rigorous attempts to understand a domain of features of people (personality traits) by thoroughly sampling and measuring that domain, and then, through quantitative, statistical methods, attempting to derive the structure of that domain via the covariance structure (John & Shrivasta, 1999). Additionally, HiTOP recognizes evidence suggesting dimensions of covariation among psychopathology symptoms may be largely the same as dimensions of normal personality variation, further encouraging HiTOP's aims and methodological techniques to follow this lead and align its validity standards with that of personality research.

Similarly, intelligence research, which produced the general factor of intelligence (called the "g" factor) that is thought to account for the positive correlation among high scores on intelligence tests (Caspi et al, 2014), has arguably served as a methodological model for thinking about and incorporating the general factor in psychopathology that is thought to account for positive correlations among various features of psychopathology. Thus, conceptions of structural validity and standards for empirical and quantitative structural evidence that developed from the bottom-up via the scientific practice of personality and intelligence research have found their influence in SFPV.

Lastly, SFPV can be thought in a broad sense to have developed out of and in response to diagnostic validity. SFPV adopts validity concepts such as the concept of a validator, a concept of validity evidence exclusively developed within diagnostic validity. Most significantly, the revision process for the *DSM-5*, which features detailed discussions on the appropriate weighting of different forms of validity evidence, was influential in the design of the HiTOP's recently updated revisions protocol procedure (Forbes et al., 2023). Thus, while SFPV and diagnostic validity ultimately amount to two very distinct senses of validity, one cannot envision a world where SFPV developed the way it has without the influence of diagnostic validity.

3.5 Refinements of Structure-First Psychometric Validity from Revising HiTOP

The HiTOP consortium has made recent concerted efforts to update SFPV in the form of official publications (Kotov et al., 2021; Kotov et al., 2022; Forbes et. al., 2023) that expand upon and explicitly detail HiTOP's principles and procedures for revising its model. The first such refinement of SFPV relates to an emphasis on the way HiTOP constructs, and by extension, the entire HiTOP hierarchy may be considered valid. When first launched in 2017 (Kotov et al., 2017), HiTOP emphasized the structural validity of the spectra over and above any other source of validity in identifying and selecting individual HiTOP constructs. While HiTOP still strongly preferences structural validity of its spectra (and hence, is structure-first), the HiTOP consortium has since officially clarified that "evaluation of validity beyond structural validity is an optional but strongly encouraged step" (Forbes et al., 2023, p. 25). As a result, HiTOP constructs above and below the spectra may be valid in two related senses. Constructs may first be valid in a *narrow* sense, which means that the degree of validity is based on a single source of validating evidence, e.g., they may be structurally valid, but not demonstrate a high degree of external

validity, and vice versa. Second, constructs may be valid in a *complete* sense in which all forms of validity evidence from its three-stage validation process (structural, external, and reliability and utility) are sufficiently demonstrated.

Most significantly, there is a *global* sense of validity HiTOP wishes to achieve, which is the validity of the broader system of multiple, hierarchically organized constructs, i.e., the entire HiTOP hierarchical model itself. Validation in the global sense is evidenced by HiTOP's emphasis on the structural validation of its higher-order constructs (spectra and above) in its expert reviews, and more recent large-scale meta-analyses (e.g., Ringwald, Forbes, & Wright, 2023). The validation process for this global sense of validity is assumed to follow the same procedures of SFPV as in the complete sense, although there are noticeably additional principles and procedures now at play considering SFPV is being applied beyond the level of HiTOP spectra to the entire hierarchy.

A second refinement relating to HiTOP's updated revision process has to do with the narrowing in on specific types of external validity to be evaluated beyond structural validity, and the standards by which those sources of evidence are validated. When SFPV first appeared, there was a sense that external validating evidence would feature in its validation procedure, as psychometric validity in general considers external validating sources beyond those related to the internal structure—but it was unclear specifically which sources of evidence would be considered and to what degree. With recent refinements, external validity beyond structural validity for HiTOP, it turns out, is now a return to some of the original validators of Robins and Guze (1970). For HiTOP, these include the magnitude of genetic and environmental influences from behavior genetic studies, molecular-genetic risks, specific environmental risks, cognitive-and-emotional-processing abnormalities, neural substrates, biomarkers, childhood-temperament

antecedents, trajectory/illness course, and treatment response. While portions of such evidence were previously mentioned in HiTOP's expert reviews on the validity and utility of HiTOP, HiTOP has since clarified in its most recent publications that the evaluation of such evidence "necessarily involves value judgments by the reviewers and committee members" based on additional organizing principles such as "the need to balance stability with flexibility" and "promoting inclusion rather than gatekeeping" (Forbes et al., 2023, p. 10). Thus, like the *DSM*, HiTOP does in part base its decisions on which constructs to feature in the model not solely based on quantitative output, but on compromises with other more practical considerations.

HiTOP attempts to set itself apart from diagnostic validity in its assessment of such validators by adopting a formal revision process it sees as distinct from the *DSM*, referred to as *evidence-based consensus*. Contrary to the *DSM*, which HiTOP interprets as relying too heavily on expert committees that are susceptible to individual committee members' preferences and special interests, HiTOP aims to systematize the validation process in such a way that it promotes a consensus that is 1) more data-driven than previous frameworks, and 2) begins with the right kinds of evidence. HiTOP claims to accomplish this in two ways. First, by placing structural validity as the foundation by which constructs in its model are subjected for further validation, it considers the data via the structural research, not the clinical experts, to be the deciding factor on which the components of its model are derived. Constructs that HiTOP seeks to further validate in more complete and/or global senses are thus not initially decided upon by committees who review the evidence and then make decisions based on an assigned level of priority to different classes of validators; rather, HiTOP constructs of interest emerge from the evidence gathered by the HiTOP reviewers.

Second, HiTOP adopts specific evidence-based criteria to evaluate revisions to HiTOP constructs that provide the strength of recommendations developed from an adapted version of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system (Balsham et. Al., 2011; Guyatt et. Al., 2011) to score the strength of evidence from studies relevant to a proposed revision of a construct. The GRADE system is an example of evidence grading, a systematic method associated with Evidence-Based Medicine (EBM), an approach toward the use of evidence in medicine that historically has ranked clinical expertise as the lowest quality of evidence. GRADE is used to assess and rate the quality of evidence from research studies, clinical guidelines, meta-analyses, and expert opinion. With HiTOP, the adaptive criteria that may contribute to a higher quality rating of the evidence (i.e., structural evidence) include "study design, risk of method bias, appropriateness of construct measurement, relevance of each study to the proposal (how directly each study targets the proposed change), discriminant validity (sufficient markers of diverse content included, and reasonable competing models tested), and effect size" (Kotov et al., 2021). The adoption of a formalized approach like the GRADE system is indicative of HiTOP's intent to reduce the role of experts in the validation process even as it adds on additional validators that it acknowledges are subject to a kind of consensus. Use of GRADE also serves to address criticisms that certain prior revisions in HiTOP did not undergo as much scrutiny, such as when the original HiTOP model's designation of Higher-Order Dimensions was updated in 2019 to "general psychopathology" and later to the "general factor of psychopathology (p-factor)" in a way that suggests the p-factor was grandfathered in to the HiTOP model rather than revised through a more rigorous process of evidence-based consensus (Watts et al., 2023).

3.6 Structure-First Psychometric Validity, CVT, and Scientific Accuracy

Thus far, we have seen how HiTOP's SFPV model can be thought of as a hybrid validity account—one that picks and chooses certain concepts and standards at the level of general validity theory from construct validity theory (CVT), draws on particular theories of psychiatric classification, brings in concepts and standards from traditions in psychometrics and affiliates research programs in the practice of research in personality psychology, and, lastly, reflects some procedural and evidentiary aspects of diagnostic validity. SFPV thus shares several general commitments of CVT, including the general nature of validity and validation. But as a hybrid account, SFPV differs in significant ways, so much so that the account is not simply a variation of contemporary CVT in practice, but arguably its own account. While application-specific accounts of CVT are common in practice, their significant differences do bring up relevant questions on how they relate to other concepts for which a particular sense of validity is said to relate to and/or support such as HiTOP's notion of "scientific accuracy."

The first significant difference lies in the formulation of SFPV as a mishmash of two distinct versions of CVT. The first CVT account, the initial "strong" version of construct validity first posited by Cronbach and Meehl (1955), proposed to validate constructs against some proposed scientific theory which implicitly defined the construct in specific relation to other constructs in a network (the nomological network). HiTOP takes this a step further, positing that "classification must go beyond an approach that establishes separate nomological networks for putatively independent constructs" (Forbes et al., 2024, p. 8) and instead works to validate the "broader system." The broader system can be understood as explicit theory regarding "how narrow constructs (e.g., individual features) associate to form higher order constructs, how they associate to form broader ones (i.e., a hierarchy), and how all are distributed among individuals

(e.g., dimensionally vs. categorically) and relate to other relevant and differential constructs" (Forbes et al., 2023, p, 8). Thus, what HiTOP assumes SFPV to be doing is essentially an iteration of Cronbach and Meehl's "strong" version of CVT: stating the theoretical assumptions and conclusions regarding one's constructs within a broader system, and then subjecting them to empirical tests as a method for establishing their validity.

On the other hand, SFPV also seems to adopt the more contemporary understanding of CVT, articulated by Messick (1989) and developed and maintained in *Standards* (APA, AERA, & NCME, 2014) that views validity as an integrated evaluative judgment of the empirical validity evidence. HiTOP's process of evidence-based consensus for revising the HiTOP model essentially designates validity as a property of the interpretation of various forms of validity evidence from psychological measurement (e.g., structural, external) rather than an invariant property of the construct itself. HiTOP thus makes no initial determination of whether a construct is valid only to accept it as a fixed property, instead putting forth principles and procedures for revising the validity of constructs based on updated evidence and interpretations of that evidence. As previously discussed, this version of CVT developed out of the apparent failures of the "strong" version of CVT. Without well-developed theories in psychology, construct validation in Cronbach and Meehl's "strong" sense which required the existence of nomological networks had very limited utility. Validity thus developed into a proposed judgment of the overall validity evidence, based on a never-ending process of validation. Considering these two versions of CVT rest on different philosophical foundations, it may be the case the philosophical presuppositions and underlying logic of SFPV are inconsistent and in need of further clarification and interpretation.

A second difference is regarding forms of validity evidence. CVT centers on five forms of evidence: Evidence Based on Test Content (e.g., content validity), Evidence Based on Response Processes (e.g., task deconstruction), Evidence Based on Internal Structure (e.g., structural validity), Evidence Based on Relations to Other Variables (e.g., external validity), and Evidence for Consequences of Testing (e.g., appropriate use of the construct). SFPV, however, draws on just two: Evidence Based on Internal Structure, reflecting HiTOP's descriptive aims, and Evidence Based on Relations to Other Variables, reflecting HiTOP's predictive aims. With "the specific aim of curtailing decision-making based on special interests, tradition, or politics" (Forbes et al, 2023; Krueger et al., 2018), the appropriate use of a construct (Evidence for Consequences of Testing) does not count as a form of validity evidence for SFPV. If anything, appropriate use for SFPV is a product of whether the construct is scientifically accurate, as indicated by structural and external validity evidence.

In terms of integrating the validity evidence, CVT relies on all available evidence without specifying any weighting of the evidence or relationship between different forms of evidence. SFPV, on the other hand, places Evidence Based on Internal Structure as the essential form of validity evidence. Under SFPV, if a construct is not interpreted to have structural validity, all the external validity evidence in the world will not lead to a positive judgment regarding the overall validity of that construct. Furthermore, in SFPV, external validity is thought to act in service of supporting judgments regarding structural validity. That is, there is an additional specified relationship between the external validity of the construct and its structural validity that is not found in CVT. In SFPV, external validity provides a sort of feedback loop for supporting interpretations of structural validity.

A third difference can be seen in how HiTOP interprets the appropriate use of a construct. Whereas recent accounts of CVT view validation of a construct as a two-step process in terms of establishing 1) accuracy and 2) appropriateness of use, SFPV seems to initially not distinguish between the two (or at least they *think* there is no distinction). For SFPV, only a construct that achieves a sufficient amount of scientific accuracy is an appropriate construct to be featured in the HiTOP model, and thus, all that is required for validating a construct is evaluating whether it is well-supported by the quantitative empirical validity evidence. What is ultimately included in the HiTOP model, however, may now depend on other guiding principles and value judgments of those who interpret and review the validating evidence based on its most recently published revisions protocol. Thus, HiTOP may be open to distinguishing between the concepts of the accuracy and appropriateness of the use of a construct, as well as the accuracy and appropriateness of the entire hierarchy, although such an acknowledgment has yet to be made.

Considering SFPV's mishmash of seemingly disparate senses of CVT, its additional application-specific differences, and recent refinements to its revisions protocol, HiTOP's recent claim that "establishing the scientific accuracy of psychiatric classification systems is essentially the task of establishing their validity" (Forbes et al., 2024, p. 8) appears underspecified and requiring additional clarification. At what point may a HiTOP construct versus the entire HiTOP hierarchy be thought to carry a sufficient degree of scientific accuracy? Is the minimum for scientific accuracy for either the base foundational sense of structural validity, which has been expressly prioritized, or are external sources of validity required? If the former, does structural validity depend exclusively on model fit (the degree to which a specific latent factor model fits the data), or does the testing of certain theories play a role, and if so, how? What amount of structural evidence warrants an evidence-based consensus that a HiTOP construct, or the entire

hierarchy even has sufficient structural validity? If additional validity sources are required to establish sufficient scientific accuracy (e.g., reliability/utility evidence), which sources specifically, how are they aggregated, and to what degree? Seeing how HITOP still includes certain provisional constructs in its model and thus indicates their lack of sufficient scientific accuracy, what types of evidence would include or exclude such a construct upon a subsequent revision? As HiTOP's revision process is said to be based on evidence-based consensus which, at the end of the day, is still a form of consensus, aren't experts ultimately still involved in making judgments concerning a construct's scientific accuracy? Should non-epistemic values of constructs such as flexibility, stability, and inclusion by which these consensus groups also weigh in their decisions be perceived as distinct from or contributing toward a construct's scientific accuracy? Depending on the answers to these myriad questions, it's possible, if not likely, that establishing the scientific accuracy of psychiatric classifications is not quite the equivalent task of establishing their validity as HiTOP understands it. As a result, the notion of scientific accuracy for HiTOP and SFPV would benefit from further specification.

3.7 Philosophical Underpinnings of HiTOP, SFPV and Scientific Accuracy

As described by Slaney (2017), competing historical interpretations of the philosophical foundations of CVT have tended to fall into three general frameworks: positivist, realist, and positive-realist hybrid accounts, which inform positions of ontology, epistemology, and causality, as they relate to the nature of constructs and the capacity to measure them. A consequence of the existing ambiguities of interpretations of the philosophical backdrop of CVT is that those inconsistencies may be replicated in the practice of those researchers who model their validity accounts in part based on CVT. That is, the researchers themselves may remain

faithful to a particular interpretation of CVT, which they incorporate into their account, that is ultimately inconsistent with the philosophical underpinnings of their respective system of classification or research framework. The degree to which the incompatibility of positions poses a problem for HiTOP is debatable, but at a minimum it may be said that it leads to difficulty in identifying the philosophical positions of hybrid validity accounts like SFPV, in which ambiguities may remain, while also having additional consequences for the concept of scientific accuracy.

Scientific Realism and HiTOP

The HiTOP consortium adopts a scientific realist stance, generally defined as a "positive epistemic attitude toward theories, including parts putatively concerning the unobservable" (Chakravartty, 2017) in terms of its quantitative empirical approach to psychopathology. HiTOP's scientific realist position is shaped across two dimensions, ontological and epistemological, as evidenced in recent HiTOP publications.

The ontology of HiTOP constructs is presented so that there exists mind-independent causes, entities, processes, and structures underlying HiTOP constructs, and consequently for every construct there exists at least one counterpart in nature. Consider the following claims: "The quantitative approach seeks consensus of studies on the natural organization of mental health" (Kotov et al., 2021, p. 86), and "HiTOP follows the quantitative approach to nosology that seeks to identify natural constellations of signs and symptoms" (Kotov et al. 2022, p. 1666). Here, the natural organization or natural constellation of signs and symptoms signifies the actual patterns of associations between psychopathology phenomena of interest, and not merely approximations or heuristic concepts or classifications. HiTOP constructs, which are the

dimensions and/or factors of psychopathology quantitatively derived from empirical studies, are considered to directly refer to the real psychopathological phenomena. HiTOP assumes the natural structure of psychopathology to also be dimensional, which prior classification systems have failed to accurately portray: “Importantly, imposition of a categorical nomenclature on naturally dimensional phenomena leads to a substantial loss of information and to diagnostic instability.” (Kotov et al., 2017, p. 5). While the structure of psychopathology for HiTOP in the most technical sense is interpreted to mean the empirically observed covariance structure of symptoms of psychopathology as determined by quantitative studies, ultimately the structure is the [actual] structure, and should be interpreted as such. Claims regarding constructs and their hierarchical organization are not intended in some instrumentalist sense to be used only for the prediction of manifest (observable) variables. Instead, a claim about constructs should be interpreted literally, e.g., that the overarching general factor psychopathology dimensions (the p-factor) “reflects shared liability for all forms of psychopathology” (DeYoung et al., 2023, p. 6).

Additionally, HiTOP maintains an epistemologically scientific realist position. In these cases, claims interpreted literally about their constructs establish knowledge of psychopathology as it exists in nature, when claims are produced with the appropriate epistemic criteria, viz., quantitative empirical methods and an appropriate validation process. For HiTOP, knowledge of psychopathology emerges from HiTOP’s process of nosologic discovery: “The natural organization of psychopathology can be discerned in the co-occurrence of its features.” (Kotov et al., 2021, p. 86). “A fundamental HiTOP aim is to move beyond traditional diagnostic categories to establish the empirical structure that emerges from quantitative analyses of comprehensive symptom-level and trait-level measurement of psychopathology.” (Forbes et al., 2023, p. 30). HiTOP asserts that the quantitative empirical approach is epistemologically robust

enough to produce such knowledge. Psychopathological phenomena and knowledge regarding those phenomena are discovered based on assumptions that 1) the structure of psychology exists, and 2) the quantitative tools of psychometrics are epistemologically capable of measuring it in a way that tells us what it actually is. The capacity of the quantitative empirical approach to be capable of such a feat—that is, discovering the actual structure that is not constructed but that emerges from the measurement—may lie in additional scientific realist commitments more generally identified by Maraun (2003) in psychological researchers who utilize factor analysis. In "Central Account," Maraun describes how researchers in quantitative psychology typically consider latent variables to be unobservable features that are the causal sources for the observed measurement outcome, implying latent variable modeling can readily detect these unobserved features.

Scientific Realism and SFPV

Considering HiTOP is based in realism, it is a reasonable assumption that SFPV is as well. The reality, however, is not so clear. For one, it appears that SFPV may rely in part on what others have called the positivist characterization of CVT, particularly regarding appeals to nomological networks, discussed by researchers and citing Cronbach and Meehl's original (1955) formulation of CVT: "The strategy for revising the HiTOP framework closely follows the standard analytic sequence for evaluation of nomological networks…" (Forbes et. al., 2023, p. 16). "The validity of hypothesized constructs ultimately depends upon locating them in nomological networks, that is, theoretically coherent pattern of linkages among the constructs, other constructs, and observable variables that account for their interrelationships." (Brislin et. al., 2022, p. 3052). Despite the subsequent critiques of the original formulation of CVT offered

by Cronbach (1975, 1989) that recognized how unfeasible an account of validity based on fully specified nomological networks in psychology is, it would appear that HiTOP researchers disregard later critiques of the positivist framing of CVT, haven't been exposed them, have different ideas in mind when it comes to nomological networks, or are just confused in thinking SFPV reflects specific positivist philosophical foundations that do not bear out when validation occurs in practice.

Even if HiTOP were to continue to argue for this original formulation, there exist some interpretations such that the formulation may be interpreted as a form of scientific realism and thus (more) compatible with HiTOP's scientific realist stance. For example, Slaney (2017) notes certain scholars such as Rozeboom (1984) have interpreted the original formulation of CVT to Feigel's empirical realism, under which the semantics of our scientific constructs are "about whatever features of the world have the observationally describable character that their defining theory says they have" (Rozeboom, 1984, p. 212). Additionally, Rozeboom claims that CVT has been previously mischaracterized as being based in positivism, stating "[i]n explicit opposition to positivist doctrine, the view that psychology's theoretical constructs designate real underlying causes through their conceptual roles in the 'nomological network' was forcefully articulated by Cronbach and Meehl (1955)" (p. 214). Similarly, SFPV could be interpreted to proceed along these same lines and that, as described by Slaney (2017), the appeal to nomological networks in CVT could also be understood for SFPV as "merely a methodological move that enable[s] the empirical testing of causal hypotheses" (p. 179).

SFPV, however, may also be interpreted as a mixed positivist-realist position. If SFPV also draws in part on Messick's contemporary articulation of CVT, it may further share what Messick referred to as a constructivist-realist account of CVT, which saw compatibility between positivist

and realist philosophies. Overall, the position essentially asserts that not all constructs are thought to refer to real counterparts in nature. Messick (1981) noted that "many useful constructs, especially higher order constructs…are employed within this framework as heuristic devices for organizing observed relationships with no necessary presumption of real entities underlying them" (p. 583). This interpretation would be more compatible if HiTOP were to take a realist stance toward narrow constructs (e.g., symptom components and maladaptive traits), and a constructivist stance toward higher-order constructs (e.g., General Factor of Psychopathology, the p-factor), which does not appear to be the case at this time.

Considering the potential disconnect between scientific realist positions with HiTOP and possible positivist, realist, and positivist-realist positions within SFPV of which some but not all may overlap, what does this mean for the concept of scientific accuracy? For HiTOP, what the consortium is truly aiming for in terms of scientifically accurate constructs are those constructs which, after having achieved what it understands to be validity, may be considered to stand for some real underlying feature of psychopathology. In this regard, we might think of these constructs as achieving a kind of "big-S" scientific accuracy insofar as both HiTOP and SFPV are scientific realist-based. At the same time, since aspects of SFPV may reflect non-scientific realist underpinnings based in a specific understanding of construct validity theory (CVT), then the scientific accuracy of certain constructs, and the entire hierarchy itself, is not simply something self-evident from the data/evidence, but requires inferences based on judgments made by informed experts of the accumulated evidence. While perhaps not as desirable for HiTOP, we might interpret this as a more conservative "small-S" conception of scientific accuracy that is more consistent with CVT as understood today.

Conclusion

In this chapter, I have argued that Structure-First Psychometric Validity (SFPV), HiTOP's validation process that is distinct from the *DSM*'s diagnostic validity insofar as it is based in psychometric validity and specific senses of CVT, may be interpreted as a three-stage validity process with the primary aim of establishing the scientific accuracy of its quantitively derived dimensional constructs, which stand for latent "common cause" features of psychopathology. As SFPV draws on various instances of CVT, it is not simply a variation of contemporary CVT in practice, but arguably its own hybrid validity account. Recent refinements of SFPV move validity in HiTOP beyond the structural validation of individual spectra and toward validation of the entire HiTOP hierarchy via evidence-based consensus, raising questions concerning how the concept of scientific accuracy relates to these two validation processes. Finally, despite HiTOP's numerous scientific realist assertions, there may be a disconnect between the underlying philosophy of the overall approach and the validation process, which warrants further evaluation.

We saw in chapter 2 how the *DSM*, in attempting to develop a more empirically-motivated validation process, updated its approach to achieving diagnostic validity to be more like the "strong" version of CVT in principle. In turn, HiTOP has undergone its own refinements to SFPV that, being actually based in CVT, appear to draw on certain aspects of the "strong" version with the appeal to nomological networks as well as other standards attempting to make its account more "scientific." Despite some surface similarities between the two, this is frankly where their commonalities end. The starting point and structure for validation are both based on a very different set of standards, and neither would accept the other's conception as scientific in their respective visions.

While HiTOP represents one application of psychometric validity based in a common cause understanding of mental illness, a second alternative framework based in psychometric validity but aligning with the theory of mutualism offers yet another approach and advocates for its own standards of a scientific validation process by testing formalized theories. In the following chapter, I introduce the network approach to psychopathology, whose understanding of validation challenges many of the underlying assumptions of both diagnostic validity and HiTOP's structure-first psychometric validity. I instead advocate for yet another very distinct sense of validity that I term *network psychometric validity*.

References for Chapter 3

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2, Pt.2), 1–38.

American Psychological Association. (1974). *Standards for educational & psychological tests*.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Balshem, H., Helfand, M., Schünemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., Vist, G. E., Falck-Ytter, Y., Meerpohl, J., Norris, S., & Guyatt, G. H. (2011). GRADE guidelines: 3. Rating the quality of evidence. *Journal of clinical epidemiology*, *64*(4), 401–406.

Blashfield, R. K., & Draguns, J. G. (1976). Toward a taxonomy of psychopathology: The purpose of psychiatric classification. *The British journal of psychiatry*, *129*(6), 574–583.

Brislin, S. J., Martz, M. E., Joshi, S., Duval, E. R., Gard, A., Clark, D. A., Hyde, L. W., Hicks, B. M., Taxali, A., Angstadt, M., Rutherford, S., Heitzeg, M. M., & Sripada, C. (2022). Differentiated nomological networks of internalizing, externalizing, and the general factor of psychopathology ('*p* factor') in emerging adolescence in the ABCD study. *Psychological medicine*, *52*(14), 3051–3061.

Buckingham, B. R. (1921). Intelligence and its measurement: A symposium— XIV. Journal of Educational Psychology, 12, 271–275.

Casey, B. J., Craddock, N., Cuthbert, B. N., Hyman, S. E., Lee, F. S., & Ressler, K. J. (2013). *DSM-5* and RDoC: progress in psychiatry research?. *Nature Reviews Neuroscience*, *14*(11), 810–814.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders?. *Clinical psychological science*, *2*(2), 119–137.

Chakravartty, A. (2017, June 12), Scientific Realism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/scientific-realism.

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.

Colliver, J. A., Conlee, M. J., & Verhulst, S. J. (2012). From test validity to construct validity… and back?. *Medical education*, *46*(4), 366–371.

Cronbach, L. J. (1971). Test validation. In R. Thorndike (Ed.), *Educational measurement* (2nd ed., p. 443). Washington DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum Associates, Inc.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Cureton, E. E. (1951).Validity. In: E. F. Lindquist (Ed.), Educational measurement, 621–694.Washington, DC: American Council on Education.

DeYoung, C. G., Blain, S. D., Latzman, R. D., Grazioplene, R., Haltigan, J. D., Kotov, R., ... & Tobin, K. E. (2023, May 5). The hierarchical taxonomy of psychopathology (HiTOP) and the search for neurobiological substrates of mental illness: A systematic review and roadmap for future research. https://doi.org/10.31234/osf.io/yatw7.

Eysenck, H. J., Wakefield Jr., J. A., & Friedman, A. F. (1983). Diagnosis and clinical assessment: The *DSM-III*. *Annual review of psychology*, *34*(1), 167–193.

Feigl, H. (1943). Logical empiricism. In D. D. Runes (Ed.), *Twentieth century philosophy: Living schools of thought* (pp. 373–416). Greenwood.

Feigl, H. (1950). Existential hypotheses: Realistic versus phenomenalistic interpretations. *Philosophy of science*, *17*(1), 35–62.

Forbes, M. K., Sunderland, M., Rapee, R. M., Batterham, P. J., Calear, A. L., Carragher, N., Ruggero, C., Zimmerman, M., Baillie, A. J., Lynch, S. J., Mewton, L., Slade, T., & Krueger, R. F (2021). A detailed hierarchical model of psychopathology: From individual symptoms up to the general factor of psychopathology. *Clinical psychological science*, *9*(2), 139–168.

Forbes, M. K., Ringwald, W. R., Allen, T., Cicero, D. C., Clark, L. A., DeYoung, C. G., Eaton, N., Kotov, R., Krueger, R. F., Latzman, R. D., Martin, E. A., Naragon-Gainey, K., Ruggero, C. J., Waldman, I. D., Brandes, C., Fried, E. I., Goghari, V. M., Hankin, B., Sperry, S., . . . Wright, A. G. C. (2024). Principles and procedures for revising the hierarchical taxonomy of psychopathology. *Journal of Psychopathology and Clinical Science, 133*(1), 4–19.

Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Glasziou, P., deBeer, H., Jaeschke, R., Rind, D., Meerpohl, J., Dahm, P., & Schünemann, H. J. (2011). GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *Journal of clinical epidemiology*, *64*(4), 383–394.

Hempel, C. G. (1965). *Aspects of scientific explanation* (Vol. 1). Free Press.

Hubbard, K., & Hegarty, P. (2016). Blots and all: A history of the Rorschach ink blot test in Britain. *Journal of the history of the behavioral sciences*, *52*(2), 146–166.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). Guilford Press.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin*, *112*(3), 527.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of educational measurement*, *38*(4), 319–342.

Kendell, R., & Jablensky, A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. *American journal of psychiatry*, *160*(1), 4–12.

Kendler, K. S., & First, M. B. (2010). Alternative futures for the *DSM* revision process: Iteration v. paradigm shift. *The British journal of psychiatry*, *197*(4), 263–265.

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., Miller, J. D., … Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of abnormal psychology*, *126*(4), 454–477.

Kotov, R., Jonas, K. G., Carpenter, W. T., Dretsch, M. N., Eaton, N. R., Forbes, M. K., ... & HiTOP Utility Workgroup. (2020). Validity and utility of hierarchical taxonomy of psychopathology (HiTOP): I. Psychosis superspectrum. *World Psychiatry*, *19*(2), 151–172.

Kotov, R., Krueger, R. F., Watson, D., Cicero, D. C., Conway, C. C., DeYoung, C. G., ... & Wright, A. G. (2021). The hierarchical taxonomy of psychopathology (HiTOP): A

quantitative nosology based on consensus of evidence. *Annual review of clinical psychology*, *17*, 83–108.

Kotov, R., Cicero, D. C., Conway, C. C., DeYoung, C. G., Dombrovski, A., Eaton, N. R., ... & Wright, A. G. (2022). The hierarchical taxonomy of psychopathology (HiTOP) in psychiatric practice and research. *Psychological medicine*, *52*(9), 1666–1678.

Krueger, R. F., & Piasecki, T. M. (2002). Toward a dimensional and psychometrically-informed approach to conceptualizing psychopathology. *Behaviour research and therapy*, *40*(5), 485–499.

Krueger, R. F., Kotov, R., Watson, D., Forbes, M. K., Eaton, N. R., Ruggero, C. J., ... & Zimmermann, J. (2018). Progress in achieving quantitative classification of psychopathology. *World psychiatry*, *17*(3), 282–293.

Krueger, R. F., Hobbs, K. A., Conway, C. C., Dick, D. M., Dretsch, M. N., Eaton, N. R., ... & HiTOP Utility Workgroup. (2021). Validity and utility of hierarchical taxonomy of psychopathology (HiTOP): II. Externalizing superspectrum. *World Psychiatry*, *20*(2), 171–193.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological reports*, *3*(3), 635–694.

Maraun, M. D. (2003). *Myths and confusions: Psychometrics and the latent variable model*. Unpublished Manuscript.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological bulletin*, *89*(3), 575.

Messick, S. (1987). Validity. *ETS research report series*, *1987*(2), i-208.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education; Macmillan.

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). *DSM-5* field trials in the United States and Canada, Part II: Test-retest reliability of selected categorical diagnoses. *American journal of psychiatry*, *170*(1), 59–70.

Ringwald, W. R., Forbes, M. K., & Wright, A. G. (2023). Meta-analysis of structural evidence for the hierarchical taxonomy of psychopathology (HiTOP) model. *Psychological medicine*, *53*(2), 533–546.

Rozeboom, W. W. (1984). Dispositions do explain: Picking up the pieces after hurricane Walter. In *Annals of theoretical psychology: Volume 1* (pp. 205–223). Springer.

Shaw, S. D., & Newton, P. E. (2014). *Validity in educational and psychological assessment*. SAGE Publications.

Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. Springer.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101

Vernon, P. E. (1933). The American v. the German methods of approach to the study of temperament and personality. *British journal of psychology*, *24*(2), 156.

Watson, D., Levin-Aspenson, H. F., Waszczuk, M. A., Conway, C. C., Dalgleish, T., Dretsch, M. N., Eaton, N. R., Forbes, M. K., Forbush, K. T., Hobbs, K. A., Michelini, G., Nelson, B. D., Sellbom, M., Slade, T., South, S. C., Sunderland, M., Waldman, I., Witthöft, M., Wright, A. G. C., Kotov, R., … HiTOP Utility Workgroup (2022). Validity and utility of

hierarchical taxonomy of psychopathology (HiTOP): III. Emotional dysfunction superspectrum. *World Psychiatry*, *21*(1), 26–54.

Watts, Ashley L., Ashley L. Greene, Wes Bonifay, and Eiko I. Fried. A critical evaluation of the p-factor literature. *Nature Reviews Psychology* (2023): 1–15.

Wijsen, L. (2021). *Characterizations of psychometrics* [Unpublished doctoral dissertation]. University of Amsterdam.

Zbozinek, T. D., Rose, R. D., Wolitzky-Taylor, K. B., Sherbourne, C., Sullivan, G., Stein, M. B., Roy-Byrne, P. P., & Craske, M. G. (2012). Diagnostic overlap of generalized anxiety disorder and major depressive disorder in a primary care sample. *Depression and anxiety*, *29*(12), 1065–1071.

Chapter 4: Psychometric Validity II—Network Psychometric Validity

4.1 Introduction

We saw in chapters 2 and 3 how despite their differences in approach and senses of validity, both the *DSM* and HiTOP share a similar commitment to explaining the covariation among symptoms of psychopathology (i.e., the structure of psychopathology) under a common cause model, whereby sets of symptoms tend to co-occur in syndromes because they share the same underlying cause (Fried & Cramer, 2017). For example, in the *DSM*, the fact that the symptoms of depressed mood and anhedonia, the latter defined as a "markedly diminished interest or pleasure in all, or almost all, activities of the day," (Heininga et al., 2019, p. 1), tend to show up together is because they are caused by Major Depressive Disorder, a *DSM*-based diagnostic category representing a distinct underlying disease entity. With HiTOP, the fact that depressed mood and anhedonia go together is because they are caused by the Internalizing spectra, a HiTOP dimensional construct representing an empirically derived, coherent statistical factor underlying psychopathology.

An alternative theory to explaining why symptoms co-occur is the theory of *dynamic mutualism* or *mutualism* of the network approach to psychopathology (Aristodemou et. al., 2023). Under mutualism, symptoms of psychopathology co-occur in syndromes due to the causal and mutual interactions between them. For example, a depressive episode is hypothesized to arise from the causal interaction between symptoms such as depressed mood, anhedonia, and others (e.g., insomnia, fatigue). Symptoms are not conceived as indicators of some underlying cause. Instead, it is the mutual interactions between symptoms, like agents in a complex system, that constitute "depression" itself. While the idea of symptoms mutually interacting has been

around since the *DSM-III* (Beck et al., 1979), mutualism has only recently found direct application among researchers within the network approach who have begun to construct network psychometric models of psychopathology from empirical data (e.g., Bringmann et al., 2013; Haslbeck & Waldorp, 2015; Epskamp, Borsboom, & Fried, 2017).

Given the network approach draws on psychometric models with a very different governing conception (Campbell, 2017) of psychopathology than common cause models, what does this mean for establishing validity? What is being validated, if not a diagnostic category or a dimensional construct meant to stand for some underlying cause? Are validity and validation conceptualized and utilized in similar psychometric concepts and terms? Can the same or separate sense of psychometric validity account for validation practices in the network approach, or is something else required?

In this chapter, I address the above questions by undertaking a faithful reconstruction of what I term *Network Psychometric Validity (NPV)* within the network approach to psychopathology. I first provide a general overview of NPV, along with an introduction to the network perspective framework. I then turn toward motivations for NPV based on the shortcomings of alternative methodologies, its development with the use of validity-adjacent concepts during its more exploratory phase, along with its more explicit conception of *node validity* in its confirmatory phase. I further describe how construct validity theory (CVT) connects with the concept of node validity, while also attempting to reconcile a more general shift away from contemporary CVT toward standards for theory development and construction. Lastly, I address some philosophical tensions within the network approach and NPV.

By reconstructing how the network approach primarily conceptualizes validity as being realized implicitly through the development and testing of network theories, I aim to highlight

how current and future research in psychopathology may draw on a notion of validity far broader than that of the *DSM*'s diagnostic validity or of HiTOP's Structure-First Psychometric Validity. When interpreted in this broad and implicit sense, validity may be understood as a concept that stands for some desirable characteristics or qualities a research framework wants for its models, constructs, theories, and ultimately, psychiatric classifications.

4.2 Overview of Network Psychometric Validity (NPV)

Validity within the *network approach to psychopathology*—defined as a preference to utilize a particular type of statistical model, a *network model*, to model and study the existence, development, and maintenance of psychiatric phenomena as "complex, biopsychosocial systems" (Fried, 2022, p. 501)—is also interpreted as residing within psychometric validity. Fortified with its preferences and understanding of quantitative methods, and specific standards and procedures for understanding the structure of psychopathology, the network approach draws on psychometric methods in a way that stands in stark contrast to the way HiTOP utilizes reflective models, being models where latent (unobserved) variables explain the shared variance (i.e., correlation) between indicator (observed) variables of interest. Assumptions regarding what is being measured, characteristics of the observed variables and relationship between those variables, and conceptions of a construct all differ quite considerably with network models, ultimately shaping what is thought to be required for assessing and establishing validity in the network approach.

Unlike the *DSM* or HiTOP, the network approach is not affiliated with any single governing body or research organization, nor does it offer a system of psychiatric classification or organizing research framework. The approach can be better categorized as an expanding

research tradition whose scientific aims may best be identified through close examination of the works and writings of network approach researchers. Resultingly, there has been very little explicitly written in terms of what the concepts and process for establishing validity in the network approach are, as ideas regarding validity are still in a formative stage and their development is ongoing. Thus, the validity account here is not reconstructed from the official writings of a consortium or standards committee, but instead from the activities of network researchers in scientific practice, along with some methodological papers from some of its main proponents.

I call the type of validation process adopted by researchers of the network approach the Network Psychometric Validity (NPV), which is characterized as an adaptable and conservative approach to validation, combining specific interpretations of psychometric validity, construct validity, and standards for the development and testing of network theories. Researchers working to identify and validate features and characteristics of psychometric networks, which includes the testing and validation of network theories, do so with a critical attitude toward their models and their results, as well as the kinds of claims of validity they make. In fact, upon conducting a literature search, one can find very little mentioned about a valid network model or valid network theory. The absence of typical validity concepts in network approach papers is not to be interpreted as an indication that assessments of validity and a kind of validation procedure are not taking place. Rather, most research within this approach may be interpreted as being in a stage of *pre-validation*, in which there is ongoing discussion as to what it is that requires validation. Whereas the *DSM* and HiTOP have a refined sense of the operationalized diagnostic categories or dimensional constructs they wish to validate and a process on how to validate them, the network approach has only very recently suggested that certain variables in a network and the

relations between them can be usefully interpreted as constructs, or has defined an explicit process for network hypotheses and the theories and models underlying them to be empirically tested and subsequently validated. One specific validation concept and process has been recently proposed—node validity (Bringmann et al., 2022)—which I interpret as a representative starting point for a three-stage validation process for NPV as network approaches more intently move from generating empirical phenomena in an exploratory manner to rigorously testing and subsequently validating network theories. NPV can be difficult to conceptualize without the proper terminology related to the network approach. Recent work by Freund and colleagues (2022) offers some helpful working definitions to clarify the general approach, which I now draw upon and offer a few extensions before providing a brief overview of the more explicit aspects of NPV.

A network model of mental disorders is a statistical model that represents features of a mental disorder as derived from a specific *network theory,* being an explanation of how the components in a network influence each other over time. A *network* is a representation of the relationships (formally called edges) between constituent variables (formally called nodes) within a system (Figure 3, Appendix). In psychopathology networks, the nodes represent various constituent elements of psychopathology within a data set, most often being individual symptoms, and are depicted by a circle. Nodes may also include other variables such as biomarkers, cognitive processes, and causally relevant components located outside of the network in the *external field*, an area that is outside of the network but which may causally intervene on the nodes or edges inside of the network. Edges represent conditional associations between nodes within a data set and are depicted by a line or a one or two-sided arrow (depending on the directionality of the relationship). A conditional association between two

nodes exists when the nodes are probabilistically dependent, conditional on all other variables in the data. If the association can be explained by other variables in the network such that their association would no longer persist once the other variables are controlled for, then the nodes are considered disconnected in the network. Two or more networks may be connected by what has been referred to as *bridge symptoms* (but could be thought of more generally as bridge nodes), defined as causal relations between nodes that form bridges across two or more networks. One commonly observed feature of mental disorders is their persistence in an individual despite the removal of the presumed originating cause(s). In a network model, the persistence of the network remaining in a dysfunctional state is referred to as *hysteresis*.

The formation and persistence of a mental disorder within a network model, conceptualized as a causal system of mutually reinforcing symptoms, may be characterized accordingly: some relevant biopsychosocial elements, including adverse life events, biological variables such as inflammation levels, and momentary social circumstances among the external field may trigger the activation of symptoms, typically represented by nodes, in a network (e.g., a "depressed" symptom network). These symptoms, when strongly interconnected via edges in the network, may lead to the increased activation of other "depressed" symptoms, attributable to the causality hypothesis (Borsboom, 2008; Cramer et al., 2010). Strongly connected nodes, based on the connectivity hypothesis then interact and reinforce one another (e.g., "fatigue" interacts with "difficulty concentrating" which might interact with "feelings of frustration") in a way where increased spreading activation in the network results in the network transitioning into a dysfunctional state (e.g., "depression"). The greater the connectivity among symptoms, the more quickly a network may form into and maintain a dysfunctional state. The more thorough the connections of the nodes in the network, the more likely the network is to remain in a

dysfunctional state even after removal of the original biopsychosocial variables (i.e., hysteresis), and may also reflect higher levels of severity in dysfunction (van de Leemput et al., 2014).

Under the network approach, this notion that a mental disorder develops and is sustained by the connectivity hypothesis is viewed as just one example of a *network hypothesis*—a testable and falsifiable hypothesis concerning the behavior of the network. In line with psychometric validity, creating the conditions necessary for appropriate testing of a hypothesis requires specifying and validating the *structure* of the network first. What is additionally important for validation within NPV, however, is not only validating the individual components of the network (i.e., the individual variables), but the dynamic relations between them, i.e., the network structure. The network structure is defined as the web of relations among elements in the network, which is hypothesized to causally affect the network's *state*, referring to the activation of the nodes in the network. Web of relations generally refers to the causal interactions among constituent variables that give rise to various properties assigned to individual aspects of the network and the network structure. Once the network structure has been estimated, it may be further analyzed in terms of its *properties*, generally defined as critically important features of the network. Thus, analysis of the network structure reveals both structural and relational features of a network, both of which are considered paramount to validate before deriving any meaningful empirical implications from the network model.

The validation process within NPV can be summarized as being comprised of three stages. The first stage is the validation of the individual components in the network, based in the concept of node validity, a model-specific validation process. Node validity is a two-step validation procedure that involves 1) *node selection,* referring to the adequacy of selecting appropriate variables as nodes in a network model, and 2) *node assessment*, referring to the

quality of the operationalizations used for selected variables. The second stage is the validation of the dynamical relations between the components, i.e., the validation of the network structure. The third and final stage is the empirical testing of network hypotheses, which, when successful, implicitly validates the underlying theory-derived network model. The reason the network approach came to adopt a validation process distinct from HiTOP's version of Structure-First Psychometric Validity and how the network approach to psychopathology informed the adoption of an explicit validation procedure is best understood in an examination of the origins of the network perspective and how the motivation for its creation was based in response to shortcomings in the data-driven psychometric models.

4.3 Origins of Network Psychometric Validity (NPV): The Exploratory Phase (2008–2018)

The origins of NPV reside broadly within the network perspective, which represents a family of models and methodologies that draw on the network approach to psychopathology, network theory, complex systems theory, and applications in network science to model and study mental disorders as "causal systems of mutually reinforcing symptoms" (Robinaugh et al., 2020, p. 353). The network perspective first gained traction in 2008 from psychometrician Denny Borsboom (Borsboom, 2008) and affiliated members of a psychometrics research group based in the Netherlands (Cramer et al. 2010), who first introduced the network approach to psychopathology as a novel approach toward explaining comorbidity among *DSM* diagnostic categories like Major Depressive Disorder and Generalized Anxiety Disorder. Over the past fifteen years, the network approach to psychopathology has expanded well beyond the Netherlands, with researchers across the United States and Europe exploring a variety of features of mental disorders from a network perspective, including specific *DSM*/*ICD* categories of

mental disorders (e.g., Bentall, 2014; Robinaugh et al., 2014; de Jonge at al., 2015; Levinson et al., 2017), psychological constructs (Heeren, Bernstein, & McNally, 2018), psychiatric diagnosis (van Os et al., 2013; Maung, 2016), symptomics (Armour et al., 2017; Fried, 2017), and the overall utility of the network perspective for clinical science (McNally, 2016; Hoffman, Curtiss, & McNally 2016; Contreras et al., 2019). Even as advocates for the network approach, these researchers were concerned by its inadequate selection and utilization of latent variable psychometric models. In the 2000s, Denny Borsboom and colleagues became worried about the general lack of theory motivating and informing such models (Borsboom, 2006) and others joined this line of research over the next decades (see Bonifay et. Al., 2017; Flake & Fried, 2020; Watts et al., 2023) For the network approach, a theory is conceived as a testable hypothesis that informs how a psychometric model represents a *target system,* "the particular parts of the real world and the relationships among them that give rise to the phenomena of interest" (Haslbeck et al., 2022, p. 931). Specifically, network proponents were frustrated by what they interpreted as the substandard methodological practice of setting up a data model, i.e., a model selected to represent a target system without any specific idea behind "how theoretical attributes are structured, how observables are related to them, or what the function form of that relation is" (Borsboom, 2006, p. 435); fitting the model to some data to demonstrate the model's plausibility; getting some model output, e.g., factor loadings that describe the extent to which a group of variables within a data set is associated with an underlying common factor derived from the model; and then in what amounts to a misguided eureka moment, go "aha! That's it—*that's* the structure of psychopathology."

The central issue is that without any theory motivating a particular psychometric model, the output of these data models only shows that certain factors and observable features have

something to do with one another, but not *how* or *why* they are related. In other words, such models do not support testable explanations of the phenomena.

Thus, the concerns of network proponents are twofold. First, there is the worry that absent any theory motivating the selection of specific psychometric models, choices by psychometric researchers to study and model specific features of psychopathology in a certain way (e.g., modeling for a common cause) are merely based on pragmatic and conventional grounds instead of substantive considerations related to theory. Second, the assumption made by psychometric researchers that the structure of psychopathological features "emerges from" or "is determined by" well-fitting latent variable models is to confuse the output of such models with a testable theory about such features. Thus, according to the network approach, psychologists working within a psychometric tradition who draw exclusively on data models to derive the hierarchical structure of psychopathology—e.g., the current overall strategy for HiTOP—are simply wrong in assuming that the output from their factor analytic techniques should be interpreted as the theory in which to further validate.

Given these concerns, network proponents sought to replace the data-driven models with more theory-based models of psychopathology. This meant selecting psychometric models based on theoretical reasons that, it is argued, provide the necessary rationale and testing conditions for why specific models ought to be this (and not that) way. In this regard, an initial explanatory model of mental disorders as networks began to take shape, with Borsboom and colleagues (Borsboom, 2008: Cramer et al., 2010) putting forward and testing an initial network hypothesis that symptoms may cluster together as syndromes due to the dynamic causal relationship between the symptoms themselves (Robinaugh et al, 2020). An important assumption of this hypothesis is that two observed symptoms such as difficulty sleeping and tiring easily can covary

without the need for a latent clinical disorder (e.g., Generalized Anxiety Disorder) or a latent dimensional construct (e.g., HiTOP's internalizing superspectra) to account for their correlation. That is, if one has difficulty sleeping, one will tire easily. (Borsboom et al., 2016).

With an initial theory-driven conception of mental disorders having shown promising results to guide their research, network proponents went to work in what could be described as a decade of exploratory empirical research between 2008–2018. This decade may further be characterized as moving from the look-and-see psychometrics of common factor models to try-and-see psychometrics with network models (the try-and-see referring to setting up a variety of network models motivated by some theory). During this time, researchers doing empirical work within the network approach primarily utilized exploratory network analyses, a collection of statistical techniques used to uncover data patterns, in which they would often take symptom criteria directly from major diagnostic manuals as a starting point. The characteristics most notably studied included network connectivity, which introduced the thesis that more densely connected networks confer a greater risk of psychopathology, and node centrality, which examined which symptoms were thought to be the most important or most central to a network. Simultaneously, a great deal of methodological articles appeared intending to work out how best to estimate a network's structure. A number of these articles featured both external (i.e., critics of the network approach aligned with some other approach) and internal (i.e., network proponents) methodological critiques regarding specific statistical analyses being used, with a common focus on the lack of statistical conclusion validity, being the degree to which conclusions drawn from the model are based on adequate statical methodology (Garcia-Perez, 2012). Of primary concern was the preponderance of network studies that used cross-sectional between-subjects data, i.e., data collected at a single point in time, to estimate a single network structure, of which several

methodological worries were discussed (Fried & Cramer, 2017, Forbes et al., 2017, DeYoung & Krueger, 2018). During this period, discussion did not center on the validity of the network models but instead was focused on the methodological criteria of the models within the tradition of complex systems. The most common criteria included *robustness*, related to the degree to which the empirical phenomena derived from the model can shape the development of network hypotheses for further testing, and *replicability*, whether the empirical phenomena could be successfully demonstrated with other data samples. How the methodological criteria of robustness and replicability were being discussed, however, appeared at times to go beyond strict methodological considerations and overlapped into a broader discussion regarding validation so that robustness and replicability were being conceived as validity-adjacent concepts.

In qualitative areas of research, there have been proposals to replace concepts such as internal validity with "trustworthiness" (Onwuegbuzie & Johnson, 2006) or "external validity" with "rigor" (Golafshani, 2003)—alternative concepts for which a systematized process of analysis may be available. While network proponents made no such direct comments regarding this, there is a general sense that what it means to promote robustness or produce replicable network structures is connected to establishing an implicit sense of validity—i.e., standards not explicitly stated but valued by the researchers—for the network approach. For example, a recent paper by Verwimp and colleagues on a network approach to dyslexia claims that their "results replicated previous findings, such as the dominant role of PA and RAN in reading (Allor, 2002; Araújo et al., 2015; Verhagen et al., 2008), thereby validating our new network approach." (Verwimp et al., 2023, p. 1022). For Verwimp et al., the term "validating" here is not to be interpreted in a narrow psychometric sense, but rather more broadly as a successful assessment of a specific desirable feature of their model.

As a result of discussions regarding the methodological criteria, two key iterations within the network approach took place. The first was the willingness of researchers to accommodate the inclusion of latent variables in some psychopathology network models with the introduction of hybrid models that combined network and latent variable modeling techniques. The earliest criticisms of network models of psychopathology were directed toward the question of why researchers ought to only include network models that purposefully exclude latent variables. To critics and some proponents of the network perspective, there wasn't anything that required the exclusion of the latent variable models—only a seeming dissatisfaction with an overvaluing and inadequate utilization of those methods, and a genuine interest in explaining certain features of psychopathology without needing to appeal to such variables. In response to these worries, network researchers Epskamp, Rhemtulla, and Borsboom (2017) demonstrated two possible generalizations of the network model to encompass latent variable structures, both of which addressed some of the shortcomings when using only a latent or network model and provided a best-of-both-worlds scenario.

The second key iteration was the development of the network perspective with principles of dynamical and complex systems theory. Not all network theorists think of psychopathology networks this way, but those who do (e.g., Cramer et al., 2016; Fried, 2022; Roefs et al., 2022) view a systems perspective for network methods as being better able to make sense of the components of a psychopathology network that may be inside the network (e.g., *DSM*-defined symptoms, biological features, cognitive processes) in relation to components that may be in the so-called external field (e.g., suicidal ideation in relation to Major Depression), which, while not a part of the network, nevertheless have causal influence. This shift may reflect a change in assumptions regarding the way variables in a network are thought to interact with one another

(e.g., in thinking about how external field variables influence a system of variables vs. only modeling the way a network of symptoms influences one another). Thus, the exploratory phase was marked not only by attempts to experiment with different network models but also by responses to criticisms and improvements in network methodology.

4.4 Origins of Network Psychometric Validity (NPV): The Confirmatory Phase (2018–)

Following the initial period of exploratory work, network proponents began to shift toward a confirmatory phase whereby their central aim changed from constructing network models to developing and testing network theories and thinking critically about what would be required to do that. One major requirement was the need to address the question, "What makes for a good theory?" That is, what qualities do network proponents desire in their network theories? According to Borsboom (2017), findings from network studies are "encouraging" when the empirical results align with "standing theory" and "clinical intuition" (p. 6). Other network advocates (e.g., McNally, 2021) further emphasize the importance of results cohering with some theory and well-established clinical observation.

Thus, a good network theory in the broadest sense supports what I interpret to be two implicit senses of validity, that of *theoretical validity* and of *clinical validity*. Theoretical validity refers to the degree to which a proposed theory is evidenced by the data from the theory-informed model, i.e., the degree to which "what you think you're talking about," i.e., your theory, ultimately corresponds with "what you end up talking about," i.e., responses from your models. As for clinical validity, this roughly refers to the degree to which attributes and features of some theory end up cohering with how clinicians think about and treat them in the world. The value of something like clinical validity for the network approach appears to be in helping to

ground as well as shape network ideas and theories in the real-world clinical settings to which they are applied. A second requirement is developing explicit standards for selecting and assessing the components in the network that would provide the conditions appropriate for testing network theories. Generally considered to be the most important network components for selection and assessment are the symptoms, i.e., the nodes within a network, that may be conceived of as reflecting the conceptually and causally distinct constructs to be validated. Given their unique causal importance to the network, these are "not 'exchangeable' with other components" (Cramer et al., 2012, p. 415).

While the network approach initially set up the nodes in a psychopathology network to be the symptoms that appear in diagnostic manuals (Borsboom, 2017), clinical psychologists Jones, Heeren, and McNally (2017) argued that this conception of nodes should be expanded beyond symptoms and include additional cognitive, biological, and sociological variables (i.e., non-symptoms) that vary at the level of the individual and have causal importance to the network. Non-symptom *elements* (McNally, 2012: Robinaugh et al., 2014) may include biomarkers (Santos et al., 2017), cognitive processes like resilience in depressive networks (Horelbeke et al., 2016), and those elements considered to be outside but causally relevant to the network itself, like stressful life events (Cramer et al., 2012). Insofar as such elements can have "direct, reciprocal relationships with multiple symptoms" (Jones, Heeren, & McNally, 2017, p. 2), nodes in psychopathology networks should not just correspond to symptoms. Upon this suggestion, the beginnings of a general framework for evaluating the selection of nodes in a network was presented, so that "adding or removing nodes should be argued on a case-by-case basis and should be accompanied by empirical support that the node in question plays an autonomous causal role in the relevant network." (Jones, Heeren, & McNally, 2017, p. 3).

Thus, following what could be interpreted as a sort of pre-node validity, Bringmann et al. (2022) formerly introduced the concept of node validity, being a two-step process for validating nodes in a network that serves as the first stage of NPV. The first step, *node selection*, refers to "the adequacy of selecting appropriate variables as nodes in a network model" (p. 3). In addition to a node reflecting a construct that varies at the level of the individual and has causal relevance to a network, node selection criteria are also based on a node being *minimally complete*, so that any single selected node will fit into a network structure that only includes the "nodes necessary to model the intended phenomena" (p. 3). Additionally, selected nodes should be sufficiently distinct so that they causally interact with one another, separately identifiable, and independently manipulable, i.e., able to be intervened upon without affecting other nodes. The process of node selection takes place not just in thinking of an isolated node, but of the node within a node set, as related to a given network hypothesis concerning how those nodes relate. The second step, *node assessment*, ultimately seeks to address the degree to which the select nodes embodied the selection criteria (e.g., to what degree is a node sufficiently distinct). Thus, if a node in a psychopathology network were to be successfully judged via the adequately selected evidence and assessed, the node would be validated or be viewed as a valid node in the network.

Node validity can be thought of as validating the individual components of the network, and the current standard in terms of what is primarily being explicitly validated within NPV and the network approach. The second stage in validation for NPV is validating the specified dynamical relations between the nodes i.e., how changes in one node in the network may affect the entire network. For example, in considering a network approach to understanding panic attacks, Robinaugh et.al. (2019) posited that the node "arousal schema" controlled for feedback effects from two other nodes, "arousal" and "perceived threat," in a network that may explain

how a panic attack is dynamically generated. Thus, the dynamic relation between these three nodes may be understood as a useful construct to be validated, and what network researchers see as a critical step in the process toward developing a network theory, although standards for this kind of validation are currently underspecified.

A third and final stage in network validation is to go beyond validating nodes of the network and their dynamic relations, and toward deriving empirical implications from the model to test (and subsequently validate) a network theory. At present, this stage is not discussed in terms of an explicit validation procedure, but rather in terms of *testability*, the ability to subject a theory to appropriate empirical testing conditions, and *falsifiability,* the ability of a theory to be rejected by empirical testing, of theories concerning the overall behavior of a network. Specifically, the network approach advocates for generating, developing, and testing formal theories of psychopathology—i.e., theories expressed in a mathematical or computational language—that allow for well-developed, testable, and falsifiable hypotheses regarding how the attributes and features of psychopathology tend to show up in the clinic and the world. Network proponents argue that a "more expansive use of formal theories…will equip theorists with tools for more rigorously generating and evaluating theories, laying the groundwork for accumulative advancement of psychological knowledge" (Haslbeck, et al., 2022, p. 931).

Thus, for the network approach, it is only once components of the network and the dynamic relations between them are identified that the empirical implications of such a model can be derived and a network hypothesis tested. For example, once Cramer et al.'s (2016) simulated network model of depression is assumed to meet both conditions, a testable hypothesis constructed into a formal theory concerning the behavior of the system may be explored. In this example, a reduction in one of the nodes in the network ("insomnia") will cause a reduction in

other nodes in the network (other depression symptom nodes). At present, testing may take the form of fitting network models statistically with the use of confirmatory network analysis techniques, but other methods are being explored. Whatever the testing method, it should be "confirmatory in the strictest sense of the term" (Haslbeck et al., 2022, p. 949) as per the recommendations for standards of confirmatory research made by Wagenmakers et al. (2012), which network proponents view as most useful for identifying threats to the validity of network theories. Resultingly, the entire validation process of NPV, notably absent of a systematic checklist of validators, is considered iterative and ongoing, with a strong willingness by network proponents to return to the drawing board and revise their models and theories as a result of rigorous testing.

This final stage of theory construction, and NPV, is reflective of Markus and Borsboom's (2013) notion of test validity, as depicted in their work, *Frontiers for Test Validity Theory: Measurement, Causation, and Meaning*. Test validity emphasizes the iterative nature of validation research as mirroring that of good scientific inquiry and a validation procedure that supports the testability of hypotheses. In their penultimate chapter, Markus and Borsboom (2013) introduce an Integrative View of Test Validity, which portrays a picture of a validation process with similarities to NPV. Most notably, the validation process for test validity focuses on 1) the specification of the construct-observation relation; 2) the derivation of hypotheses from that relation; and 3) the testing of the derived hypotheses. For test validity, as with NPV, a strong link is drawn between the concept of testability and validity, so that "testable implications…guide the collection of test validity evidence," and that "greater testability of the empirical claim yields greater amenability to validation and thus the support that comes from it" (Markus & Borsboom, 2013, p. 241).

4.7 Network Psychometric Validity: Its Relation to (and Departure from) Construct Validity Theory

Since the network approach is in the very beginning phases of sorting out how to think about its components, relations between components, and clusters of components as constructs, the degree to which its conception of validity and validation that is based in psychometric validity are reflections of contemporary CVT, and how it may be a distinct sense can be difficult to parse. In what follows, I describe some general sentiments regarding CVT that appear to connect with the conception of node validity, while also attempting to reconcile what may ultimately be a more general shift away from contemporary CVT to meet the network approach's focus of developing standards for what is referred to as *theory construction*.

One aspect where node validity is reflective of the contemporary understanding of CVT is in the emphasis on the identification of potential threats to validity. The initial emphasis on the selection of the appropriate nodes in a network is framed so that the selected nodes "will best illuminate the psychopathological processes" (Bringmann et al., 2022, p. 3) with the worry that including the wrong nodes will threaten further testability and validity claims of the components and relations of components in the model. Samuel Messick, in building upon ideas initially described by Campbell and Fiske (1959) and Cook and Campbell (1979), introduced two related concepts that parallel this threat to construct validity, *construct underrepresentation* and *construct-irrelevant variance*, later adopted as *construct-irrelevance.* Construct underrepresentation, described further in the *Standards* (AERA, APA, NCME, 2014), refers to the degree to which a measure fails to capture or represent important aspects of the construct, due to the content of the measure being too narrow and thus important features or dimensions are not included. Construct irrelevance refers to the degree to which measures of the construct are

impacted by features considered irrelevant or extraneous to the construct. Relating to node validity, the notion of a selected node set needing to be minimally complete follows a similar reasoning. The failure of a node set to not contain all nodes necessary to model the intended phenomena (a kind of construct underrepresentation) and the failure of a node set to exclude superfluous nodes (a kind of construct irrelevance) would be conceived by network proponents as potential threats to node validity, as well as future validation efforts. A second aspect of node validity's reflection of CVT is in its proposal for node assessment. A central concern for node validity is whether nodes in a network are sufficiently distinct, so that they are separately identifiable and independently manipulable from one another. To assess whether nodes in a network meet these criteria, network proponents point to "the classic psychometric criteria such as reliability and validity [which] form the fundamental basis for assessment (AERA, APA, & NCME, 2014)" (Bringmann et al., 2022, p. 4). Based on this recommendation, node validity, and thus the network approach, aligns in part with contemporary CVT when it comes to evaluating the selection and assessment of nodes in a network.

Despite network proponent's partial alignment with the contemporary understanding of CVT in the practice of node validation, the way the network approach relates indicator (i.e., observed) nodes in a network such as "fatigue" to the overall network, whereby "fatigue" is not reflective of an underlying construct such as "depression," may signal an incompatibility for NPV with CVT and as a result, a necessary departure. In discussing the use of constructs within network models, Markus and Borsboom (2013) discuss how a reconceptualization of constructs under a network approach may shift the central question for validity:

> For instance, rather than the indicators being measures of a construct like depression, they should be considered a part of that construct—in this respect the relation between indicators and constructs is similar to that in the theory of

> behavioral domains. The problem of validity, viewed now as involving the correct representation of the construct-indicator relation, is no longer primarily one of establishing the right measurement relations, but of incorporating the right variables and establishing appropriate structural relations between them. Measurement problems are involved in relating the variables in the network to their real-world counterparts (e.g., relating the answer to the question "Are you tired?" to fatigue), but at the level that psychologists normally ask the validity question (how well does the item "Are you tired?" measure depression?), a conceptualization of the problem in terms of measurement may be fundamentally misguided. How one should go about in conducting validation research for test items that relate their constructs as nodes relate to networks is, at present, an open question. (Markus & Borsboom, 2013, p. 133)

For Markus and Borsboom, a change in how we think about constructs within network models may necessitate a fundamental change in how we go about validating them. With CVT being the framework that centers validity squarely as a problem of the appropriate measurement relations between the indicator variables and the construct, a validation framework available to the researchers for "incorporating the right variables and establishing appropriate structural relations between them" (p. 133) is Markus and Borsboom's own causal, test-centered approach, which they refer to as test validity or test validity theory. Contrasted with the current understanding of CVT, test validity narrows the conception of validity such that measures of an attribute (which may also be a construct) are valid if (1) the attribute exists, and (2) variations in the attribute causally produce variation in the measurement outcome.

While an integrative view of test validity for scientific researchers has been outlined in detail in Markus and Borsboom's *Frontiers for Test Validity Theory: Measurement, Causation, and Meaning*, such a view has yet to explicitly feature within the validation procedures undertaken by the network approach. That is, while underlying ontological, causal, and meaning-related assumptions of test validity may be shared among some network proponents, the network literature has not gone on to articulate a validation process under a test validity framework. The

omission by network researchers to readily adopt a test validation framework as outlined by Markus and Borsboom may be in part due to that "the idea of network models as measurement models remains largely unexplored" (p. 184). New explicit conceptions and frameworks of how to think about validating components in a network thus may be interpreted as only temporarily aligning with more customary standards of the discipline (i.e., CVT) more familiar to the researchers as a sort of pragmatic validation phase during NPV's state of pre-validation.

Instead of spelling out an explicit test validation framework for the network approach, what network theorists have focused on instead is developing standards for theory construction. Borsboom and colleagues (2021) introduced what they term *theory construction methodology* (TCM), a five-step practical sequence designed "to facilitate the formation of explanatory theories" (p. 763) with a focus on formal modeling, using a mutualism model of intelligence as an example. Their sequence involves 1) identifying the relevant empirical phenomena to be explained, 2) constructing a prototheory, i.e., an initial explanatory model, 3) developing a formal model, conceived as a "thinking tool" (as opposed to data tools used for fitting model parameters to data) that permits the testing of a set of principles from a well-specified theory, 4) checking the adequacy of the formal theory so that it can explain the empirical phenomena, and 5) assessing the overall worth of the theory by appeal to some criteria. They propose as criteria that a theory might be subjected to include its predictive success, Kuhn's (1977) five properties of a good theory (accuracy, consistency, scope, simplicity, scope, and fruitfulness), or the theory's explanatory virtues, i.e., evaluation as inference to the best explanation (Thagard, 1992). Notably absent within the discussion of theory construction is the explicit mention of the concepts of valid, validity, and validation.

Relating TCM in terms of NPV's three-stage validation process, one can interpret TCM's five-sequence process as being consistent with NPV, whereby both are in a sense supportive of one another. The first three sequences of TCM align with the first two stages of NPV in that both support the selection and assessment of that which is to be validated and/or tested, being the components, the relation between the components, and the formal theory/model. It is only during these initial sequences/stages, however, that standards of contemporary CVT are explicitly drawn upon. Where NPV seemingly departs from contemporary CVT, and in a sense the difficult questions of validity and validation of network constructs in general, is in the fourth and fifth sequence of TCM, which aligns with the third and final stage of NPV, being the testing of the formal theory. Thus, it would appear that in adopting a multi-stage validation process in NPV, network theorists are employing just enough of an explicit validation procedure based in psychometric validity and CVT necessary to support the adequate testing of theories, while still ultimately wanting to base the worth of their approach in something separate from CVT. Network theories subjected to such rigorous standards for theory construction and testing arguably carry their own distinct and implicit sense of validity. Thus, for the network approach, a network theory meeting the standards for theory construction and testing is validation.

One final point is, that by incorporating aspects of CVT within the initial stages of NPV before theory testing, the network approach still faces an unresolved question: what counts as a construct? Under node validity, nodes in the network may be a symptom, a cognitive mechanism, or a biological mechanism, among other various elements. Should we take distinctive types of nodes, which may require assessment at different levels, to be the same type of construct? Network researchers also differ in terms of their approach to featuring constructs independent of or outside of the network. For example, some network researchers interpret nodes

that are not strongly connected in a given network to be less important constructs for selection and assessment (and ultimately treatment), whereas others see nodes independent of their overall connectivity but that have potential downstream consequences to be as just as important (e.g., "suicidal ideation" is still important even in a network where it is not strongly connected). Should nodes independent of (but still within) the network be conceptualized differently? Nodes outside of the network, such as "the context of daily life" can feature in the external field, as well as other broad causes through mixed network and latent variable models. Are those network constructs, too, and if so, should they adhere to the same criteria for node selection assessment and validation if they are not technically in the network? Network proponents such as Borsboom and colleagues have made a distinction between nodes as constructs, and the *focal construct*, referring to the overall characteristic of the network (Markus and Borsboom, 2013; Borsboom 2008b). What does it mean to have a construct represent the overall characteristic of a network, and how does validation of the focal construct relate to validation of a node? Should community structure, being the clustering of a densely connected group of nodes, be thought of as a sort of syndromal construct to be validated, and if so, what does that process look like? Various specifications and open questions for the types and relations of constructs in the network approach, and how those may be addressed in the first two stages of NPV remain.

4.8 Philosophical Underpinnings of the Network Approach and NPV

In addition to the conceptual issues regarding the status of constructs in the network approach, there may be contrasting or mixed positions in the philosophical assumptions of the overall network approach to psychopathology, and particular conceptions of NPV—which incorporates aspects of contemporary CVT in its early stages with node validity, while seemingly

departing from CVT during the stage associated with theory construction. An overview of these philosophical positions follows.

Scientific Realism and the Network Approach

Scientific realist positions within the network approach center around ontological and epistemological dimensions. In terms of ontology, the overarching thesis of the network approach—the network theory of mental disorders which posits that mental disorders develop and are sustained through direct dynamics between elements within a network, i.e., are complex systems—is seen by scientific realists within the network approach not simply as a way of conceptualizing mental disorders but the way they truly exist in the world. Not only does the network approach argue that the claim "mental disorders are complex systems" closely aligns with how clinicians conceptualize and treat mental disorders (Schipek, 2009), but their true nature as complex systems necessitate an anti-reductive, network approach. Network theories, when generated, developed, and tested appropriately, depict attributes of psychopathology that are independent of the formal or computationally specified model from which they are derived as they truly appear in nature: "Mental disorders are multifactorial in constitution, etiology, and causal background, which seems overwhelmingly plausible given the current scientific record" (Borsboom, 2017, p. 7). "Ontological distinctions are relevant (Borsboom, 2008, McNally, 2012). Both the network and latent variable perspectives are ontologically realist about symptoms as these have existential referents" (McNally, 2016, p. 101). In terms of epistemology, scientific realists assert that psychopathology networks are discovered via the analyses of network theories derived from their formal models: "Networks are empirically discovered, not formed by theorists who construct them to suit certain purposes. Indeed, the causal system

perspective is ontologically realist as it presupposes mind-independent phenomena discoverable via network analyses" (McNally et al., 2015, p. 839). Moreover, with a central aim for the network approach of deriving and testing theories concerning specific aspects of psychopathology, viz., how disorders develop, persist, weaken, strengthen, and other dynamically posited features, scientific realists within the network approach see network theories as capable of explaining psychopathology phenomena, such that their theories offer true accounts of psychopathological attributes:

> "Formalizing a network theory of panic disorder serves two key purposes. First, it equips us to evaluate how well our theory achieves its fundamental aim: the *explanation* of panic disorder-related phenomena (van Rooij & Blokpoel, 2020). To evaluate whether a theory can explain a phenomenon, it is necessary to evaluate whether the phenomenon indeed follows from the theory (van Dongen et al., 2022)." (Robinaugh et. al., 2019)

When considering scientific realist interpretations of network models and theories, some critics have pointed toward the idea that network models, which adhere to a theory of mutualism (direct causal relation between observed variables), are mathematically or statistically equivalent to latent variable models, which adhere a common cause theory (unobserved common causes explain relations between variables). Such an equivalence, it is argued, may threaten one's scientific realist leanings toward network models and network theories. In response, scientific realists within the network approach have argued that such statistical equivalence does not equate to network and latent variable models being equally plausible in referring to how features exist in the world, and in fact, such models are ontologically distinct (van Bork et al., 2021). In addition, further specification may be warranted as to just what scientific realists within the network approach are realists about. On the one hand, scientific realists in the network view the claim that the network approach is "just as realist at the latent variable approach" (McNally, 2016) with just a shift in focus of locating the causal features of the network elements among the

elements themselves. On the other hand, being a scientific realist about the relations between attributes may need a more appropriate distinction from being a scientific realist concerning underlying causal entities, as is the case with HiTOP.

Constructivist-Realist, Instrumentalism, and the Network Approach

While many network researchers could be said to be operating under realist assumptions with a central aim of the approach being to put forward theories that may explain psychopathological phenomena, others adopt a more pragmatic approach. For example, researchers such as Eiko Freid, who closer align to a systems perspective of mental disorders, see methodological criticisms of network models, such as estimating a network structure from cross-sectional data or the statistical equivalence of a model, as evidence that network models cannot and do not reflect the true world (van Bork et al., 2017). Just as network model researchers criticize HiTOP for suggesting the p-factor and other psychopathology dimensions may be "discovered" from the data via structural research, it may be equally problematic to suggest network researchers "discover" networks from their data. At the same time, researchers like Fried may maintain the belief that psychopathology attributes as they exist are more like networks in some real sense than not. Thus, some network proponents adopt a *constructivist-realist* position: one that views systems as our best, but fallible, constructed attempts to capture the nature of psychopathology as system-like, which has some reality independent of how we attempt to model it.

Generally, those who take a more pragmatic stance within the network approach will in turn tend to adopt some sense of instrumentalism about network models as well. For instrumentalists, the central aim of the network approach is not to explain psychopathological

phenomena as it truly is, but to serve as useful tools for understanding and predicting such phenomena. Van Loo and Romeijn (2019) argue that an instrumentalist view of network models will offer greater flexibility and utility to work in conjunction with latent modeling techniques they have been so strongly contrasted with. Indeed, network researchers have seemingly departed from the initial rollout of the network approach which viewed latent variable modeling as undesirable due in part to its realist leanings.

Constructivist-Realist, Instrumentalism, and Realism within Network Psychometric Validity

To what degree does NPV align with the network approach's realist, constructivist-realist, and instrumentalist underpinnings? First, if we accept that node-validity assumes the standards of the classic psychometric criteria from *Standards* (2014), then we can by extension attribute at least a portion of NPV's philosophical underpinning to the more constructivist-realist views of Samuel Messick (1989). The *Standards* (2014), which is consistent with a Messickian view, defines validity as "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (p. 225). That is, validity is not dependent on the reality of the construct or whether a measure actually measures some feature of the world, but is instead thought of as a property of the inferences made from the various categories of validating evidence. Messick further held that the constructivist-realist account supports an instrumentalist view of constructs.

> "…[it] is not meant to imply that for every construct there is a counterpart reality in the person, the situation, or the interaction. On the contrary, many useful constructs, especially higher order constructs…are employed within this framework as heuristic devices for organizing observed relationships with no necessary presumption of real entities underlying them. The presumption of real entities underlying such constructs…is similarly arguable…. [Within] the constructivist- realist position…some constructs…have real trait (or situation or

> interaction) counterparts…whereas other constructs do not. (Messick, 1981, p. 583)

On the other hand, network theorists such as Borsboom, Mellenbergh, & van Heerden (2003; 2004) and further discussed by Borsboom et al. (2009) have called for a complete overhaul of the *Standard's* version of validity in favor of a realist approach, which they deem essential. Following several lengthy discussions on what they take to be the unresolvable problems associated with construct validity (see Borsboom, Mellenbergh, & van Heerden (2004); Borsboom & Mellenbergh, 2007; Borsboom et al., 2009), an alternative view of validity is presented. The view is characterized as conservative and simple: the fundamental question for validity as a measurement concept is whether a test measures the attribute it is purported to measure. This view of validity resembles a return to the classical conception of test validity in psychometrics connected with Truman Kelley's (1927) claim, "the problem of validity is that of whether a test really measures what it purports to measure" (p. 14). Validity is not a property of the inferences made based on a testing instrument as the *Standards* suggests; instead, validity should (once again) be viewed as a property of the measurement instrument. Validation for Borsboom and colleagues is "the scientific process of researching whether the test has the property of validity" (Borsboom & Mellenbergh, 2007).

Under Borsboom and colleagues' conception of test validity, "a test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcome" (Borsboom, Mellenbergh, & Van Heerden, 2004, p. 1061). The view is considered by Borsboom and colleagues to hold "a realist interpretation of psychological attributes" (p.1065) unrelated to meaning (i.e., interpretation). Validity, in turn, requires reference to real attributes that exist independent of measurement. For

Borsboom et al., the question of whether some psychological measurement instrument is valid for assessing some psychological attribute entails that we assume the attribute exists apart from theory and can be measured by instruments for which outcomes are causally determined by the attribute (Colliver, Conlee, & Verhulst, 2012). Taking a causal interpretation of measurement, which requires the attribute to have a causal effect on the scores of the assessment, has two implications: 1) validity of this sort is a property of the measurement instrument, not of the interpretation of the measurement outcomes; and 2) the question of whether a test measures the attribute it is purported to measure is a question of truth, not evidence.

The emphasis on the use of the notion of a psychological *attribute* rather than construct is intentional. It is meant to avoid supposed ambiguities tied with the latter, and support an important feature of Borsoom et al.'s conception of test validity: that the use of psychometric theories, i.e., formal theories of test behavior, is central to validation:

> "On the one hand we have a psychological attribute that we hypothesize to exist in the world, and to cause variation in our measurement outcomes. That is the thing we want to measure. On the other hand, we have the theoretical term that we use in our theories and that, if we are lucky, in fact picks out the psychological attribute in question. If we can explain how the psychological attribute acts to cause variation in our measurement outcomes, we can truly say something about the validity of our measurement instrument. This requires us to investigate the structures and processes that make up the psychological properties we are interested in and to show that these properties are picked up by the test. In essence, this means that we need to construct a psychometric model that is *psycho*metric rather than psycho*metric*. Rather than substanceless models, preferred for their philosophical or statistical niceties, psychometric models should be formal theories of test behavior. The task of validation then comes down to testing these theories in whatever way necessary." (Borsboom et al., 2009, p. 164).

We may, therefore, interpret NPV as holding both constructivist-realist and perhaps instrumentalist positions regarding its focus on the validation of nodes in a network and a network structure (i.e., stages one and two of NPV). But when it comes to theory construction

(i.e., stage three of NPV), a scientific realist position that aligns with Borsboom and colleagues' commitments underlying test validity—one that is *inconsistent* with the Messickian view associated with contemporary CVT—seems predominant. While some such as Moss (1992) consider validity theory to be compatible with multiple philosophical positions, it may be the case that a future goal of NPV would be to develop a validation research framework that is scientific realist throughout.

Conclusion

In this chapter, I've argued that the network approach's validation practices, being a far cry from the *DSM*'s validation of common cause latent disease entities or HiTOP's validation of latent dimensional constructs, embody their own distinct sense of psychometric validity which I term Network Psychometric Validity (NPV). Motivated in part by what network theorists take to be methodological issues with data-driven psychometrics, NPV is currently comprised of a three-stage validation framework that begins with the validation of the components in the network and concludes with the development and testing of network theories. The three stages of NPV include 1) validation of nodes in a network, a two-step process termed node validity; 2) validation of the network structure, being the dynamical relation between the nodes; and 3) the testing of formal network theories of psychopathology. While the first two stages are based in standards of contemporary psychometrics (i.e., construct validity theory), the third stage may be interpreted as advocating for a departure from CVT in favor of a more implicit sense of validation based in either a distinct sense of test validity or theory construction methodology, both of which require a scientific realist framework. The central aim of validation within the

network approach may ultimately amount to the testing of formalized network hypotheses "in whatever way necessary" (Borsboom et al., 2009, p. 164).

Given the network approach's lack of emphasis on explicit standards of validity, future iterations of NPV may ultimately not resemble an explicit validation procedure. The network approach represents an example of a psychopathology research framework that is simply not too worried about achieving validity in a traditional psychometric sense, and instead trusting that the validity of its approach bears out from the development and testing of its network theories. That is, what matters most for the network approach is not what makes a valid network, but what makes a good theory. Validity as a result becomes less so a fixed feature or property of some category or construct as it has been previously conceived in psychiatric classification, and instead represents a much broader notion that stands for something desirable or good, and in this case, desirable and good features as they relate to network models and network theories. An outstanding question for the network approach is, will this broader notion of validity be sufficient for a future network-informed classification system? Or like with HiTOP, will the network approach eventually return to a validity framework that draws on the clinical validators of Robins and Guze?

In the penultimate chapter, we turn toward an integrative approach to validation that attempts to bring the disease model, psychometrics, dimensionality, and multicausality together with cognitive neuroscience in one unifying validity framework. Despite being integrative, such a framework is the first of the three approaches to assert that establishing the biological basis of psychopathology, being the establishment of *etio-pathophysiological validity*, is the key to resolving the box canyon problem.

References for Chapter 4

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aristodemou, M. E., Kievit, R. A., Murray, A. L., Eisner, M., Ribeaud, D., & Fried, E. I. (2023). Common cause versus dynamic mutualism: An empirical comparison of two theories of psychopathology in two large longitudinal cohorts. *Clinical psychological science*. https://doi.org/10.1177/21677026231162814.

Armour, C., Fried, E. I., & Olff, M. (2017). PTSD symptomics: Network analyses in the field of psychotraumatology. *European journal of psychotraumatology*, *8*(sup3), 1398003.

Beck A. T., Rush A. J., Shaw F. S., Emery G. (1979). *Cognitive therapy of depression*. Guilford Press.

Bentall, R. P. (2014). The search for elusive structure: A promiscuous realist case for researching specific psychotic experiences such as hallucinations. *Schizophrenia bulletin*, *40*(Suppl 4), S198–S201.

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical psychological science*, *5*(1), 184–186.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440.

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of clinical psychology*, *64*(9), 1089–1108.

Borsboom, D. (2017). A network theory of mental disorders. *World psychiatry*, *16*(1), 5–13.

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 85–115). Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological review*, *110*(2), 203.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological review*, *111*(4), 1061–1071.

Borsboom, D., Cramer, A. O., Kievit, R. A., Scholten, A. Z., & Franić, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications*. IAP Information Age Publishing.

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on psychological science*, *16*(4), 756–766.

Borsboom, D., Robinaugh, D. J., Group, T. P., Rhemtulla, M., & Cramer, A. O. (2018). Robustness and replicability of psychopathology networks. *World psychiatry*, *17*(2), 143–144.

Bringmann, L. F., Albers, C., Bockting, C., Borsboom, D., Ceulemans, E., Cramer, A., Epskamp, S., Eronen, M. I., Hamaker, E., Kuppens, P., Lutz, W., McNally, R., Molenaar, P., Tio, P., Voelkle, M. C., & Wichers, M. (2022). Psychopathological networks: Theory, methods and practice. *Behaviour research and therapy*, *149*, 104011.

Bringmann L. F., Vissers N., Wichers M., Geschwind N., Kuppens P., Peeters F., Borsboom, D., & Tuerlinckx F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE*, *8*(4), e60188.

Campbell, J. (2017). Validity and the causal structure of a disorder. *Philosophical issues in psychiatry*, *4*, 257–273.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81–105.

Caspi, A., & Moffitt, T. E. (2018). All for one and one for all: Mental disorders in one dimension. *American journal of psychiatry*. *175*(9), 831–844.

Colliver, J. A., Conlee, M. J., & Verhulst, S. J. (2012). From test validity to construct validity… and back?. *Medical education*, *46*(4), 366–371.

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

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (Vol. 351). Houghton Mifflin.

Cramer, A. O., Waldorp, L. J., Van Der Maas, H. L., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and brain sciences*, *33*(2–3), 137–150.

Cramer, A. O., Borsboom, D., Aggen, S. H., & Kendler, K. S. (2012). The pathoplasticity of dysphoric episodes: differential impact of stressful life events on the pattern of depressive symptom inter-correlations. *Psychological medicine*, *42*(5), 957–965.

Cramer, A. O., Van Borkulo, C. D., Giltay, E. J., Van Der Maas, H. L., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major depression as a complex dynamic system. *PloS one*, *11*(12), e0167490.

de Jonge, P., Wardenaar, K. J., & Wichers, M. (2015). What kind of thing is depression?. *Epidemiology and Psychiatric Sciences*, *24*(4), 312–314

DeYoung, C. G., & Krueger, R. F. (2018). Understanding psychopathology: Cybernetics and psychology on the boundary between order and chaos. *Psychological Inquiry*, *29*(3), 165–174.

Epskamp S., Borsboom D., & Fried E. I. (2017). Estimating psychological networks and their accuracy: A tutorial paper. *Behavioral Research Methods*, *50*, 195–212.

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, *82*, 904–927.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465.

Forbes, M. K., Wright, A. G., Markon, K. E., & Krueger, R. F. (2017). Evidence that psychopathology symptom networks have limited replicability. *Journal of abnormal psychology*, *126*(7), 969-988.

Freund, I. M., Arntz, A., Visser, R. M., & Kindt, M. (2022). Jumping back onto the giants' shoulders: Why emotional memory should be considered in a network perspective of psychopathology. *Behaviour research and therapy*, *156*, 104154.

Fried, E. I. (2017). Moving forward: How depression heterogeneity hinders progress in treatment and research. *Expert review of neurotherapeutics*, *17*(5), 423–425.

Fried, E. I. (2022). Studying mental health problems as systems, not syndromes. *Current directions in psychological science*, *31*(6), 500–508.

Fried, E. I., & Cramer, A. O. (2017). Moving forward: Challenges and directions for psychopathological network theory and methodology. *Perspectives on psychological science*, *12*(6), 999–1020.

García-Pérez, M. A. (2012). Statistical conclusion validity: Some common threats and simple remedies. *Frontiers in psychology*, *3*, 325.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The qualitative report*, *8*(4), 597–607.

Haeffel, G. J., Jeronimus, B. F., Kaiser, B. N., Weaver, L. J., Soyster, P. D., Fisher, A. J., Vargas, I., Goodson, J. T., & Lu, W. (2022). Folk classification and factor rotations: Whales, sharks, and the problems with the hierarchical taxonomy of psychopathology (HiTOP). *Clinical psychological science*, *10*(2), 259–278.

Haslbeck, J., & Waldorp, L. J. (2015). Structure estimation for mixed graphical models in high-dimensional data. *arXiv preprint arXiv:1510.05677*.

Haslbeck, J. M., & Waldorp, L. J. (2018). How well do network models predict observations? On the importance of predictability in network models. *Behavior research methods*, *50*, 853–861.

Haslbeck, J., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2022). Modeling psychopathology: From data models to formal theories. *Psychological methods*, *27*(6), 930-957.

Heeren, A., Bernstein, E. E., & McNally, R. J. (2018). Deconstructing trait anxiety: A network perspective. *Anxiety, stress, & coping*, *31*(3), 262–276.

Heininga, V. E., Dejonckheere, E., Houben, M., Obbels, J., Sienaert, P., Leroy, B., van Roy, J., & Kuppens, P. (2019). The dynamical signature of anhedonia in major depressive disorder: positive emotion dynamics, reactivity, and recovery. *BMC psychiatry*, *19*(1), 1–11.

Hofmann, S. G., Curtiss, J., & McNally, R. J. (2016). A complex network perspective on clinical science. *Perspectives on psychological science*, *11*(5), 597–605.

Hoorelbeke, K., Marchetti, I., De Schryver, M., & Koster, E. H. (2016). The interplay between cognitive risk and resilience factors in remitted depression: A network analysis. *Journal of Affective Disorders*, *195*, 96–104.

Hyman, S. E. (2021). Psychiatric disorders: Grounded in human biology but not natural kinds. *Perspectives in biology and medicine*, *64*(1), 6–28.

Jones, P. J., Heeren, A., & McNally, R. J. (2017). Commentary: A network theory of mental disorders. *Frontiers in psychology*, *8*, 1305, 14-16.

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book Company.

Kuhn, T. S. (1977). *The essential tension*. University of Chicago Press.

Levinson, C. A., Zerwas, S., Calebs, B., Forbush, K., Kordy, H., Watson, H., Hofmeier, S., Levine, M., Crosby, R. D., Peat, C., Runfola, C. D., Zimmer, B., Moesner, M., Marcus, M. D., & Bulik, C. M. (2017). The core symptoms of bulimia nervosa, anxiety, and depression: A network analysis. *Journal of abnormal psychology*, *126*(3), 340.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Taylor & Francis Group.

Maung, H. H. (2016). Diagnosis and causal explanation in psychiatry. *Studies in history and philosophy of biological and biomedical sciences*, *60*, 15–24.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological bulletin*, *89*, 575–588.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education; Macmillan.

McNally, R. J. (2016). Can network analysis transform psychopathology?. *Behaviour research and therapy*, *86*, 95–104.

McNally, R. J. (2021). Network analysis of psychopathology: Controversies and challenges. *Annual review of clinical psychology*, *17*, 31–53.

McNally, R. J., Robinaugh, D. J., Wu, G. W., Wang, L., Deserno, M. K., & Borsboom, D. (2015). Mental disorders as causal systems: A network approach to posttraumatic stress disorder. *Clinical psychological science*, *3*(6), 836–849.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of educational research*, *62*(3), 229–258.

Murphy, D. (2011). Conceptual foundations of biological psychiatry. In F. Gifford (Ed.), *Philosophy of medicine* (pp. 425–451). Elsevier.

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the Schools*, *13*(1), 48–63.

Rhemtulla, M., Fried, E. I., Aggen, S. H., Tuerlinckx, F., Kendler, K. S., & Borsboom, D. (2016). Network analysis of substance abuse and dependence symptoms. *Drug and alcohol dependence*, *161*, 230–237.

Robinaugh, D. J., LeBlanc, N. J., Vuletich, H. A., & McNally, R. J. (2014). Network analysis of persistent complex bereavement disorder in conjugally bereaved adults. *Journal of abnormal psychology*, *123*(3), 510-522.

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

Robinaugh, D. J., Hoekstra, R. H., Toner, E. R., & Borsboom, D. (2020). The network approach to psychopathology: A review of the literature 2008–2018 and an agenda for future research. *Psychological medicine*, *50*(3), 353–366.

Robinaugh, D. J., Haslbeck, J. M., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, *16*(4), 725–743.

Roefs, A., Fried, E. I., Kindt, M., Martijn, C., Elzinga, B., Evers, A. W., Wiers, R. W., Borsboom, D., & Jansen, A. (2022). A new science of mental disorders: Using personalised, transdiagnostic, dynamical systems to understand, model, diagnose and treat psychopathology. *Behaviour Research and Therapy*, *153*, 104096.

Santos Jr., H., Fried, E. I., Asafu-Adjei, J., & Ruiz, R. J. (2017). Network structure of perinatal depressive symptoms in Latinas: Relationship to stress and reproductive biomarkers. *Research in Nursing & Health*, *40*(3), 218–228.

Thagard, P. (1992). *Conceptual revolutions*. Princeton University Press.

van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & van der Maas, H. L. (2017). What is the p-factor of psychopathology? Some risks of general factor modeling. *Theory & psychology*, *27*(6), 759–773.

van Bork, R., Rhemtulla, M., Waldorp, L. J., Kruis, J., Rezvanifar, S., & Borsboom, D. (2021). Latent variable models and networks: Statistical equivalence and testability. *Multivariate behavioral research*, *56*(2), 175–198.

van de Leemput, I. A., Wichers, M., Cramer, A. O., Borsboom, D., Tuerlinckx, F., Kuppens, P., ... & Scheffer, M. (2014). Critical slowing down as early warning for the onset and

termination of depression. *Proceedings of the National Academy of Sciences*, *111*(1), 87–92.

van Loo, H. M., & Romeijn, J. W. (2019). What's in a model? Network models as tools instead of representations of what psychiatric disorders really are. *Behavioral and brain sciences*, *42*, e30.

van Os, J., Delespaul, P., Wigman, J., Myin-Germeys, I., & Wichers, M. (2013). Beyond *DSM* and *ICD*: introducing "precision diagnosis" for psychiatry using momentary assessment technology. *World psychiatry*, *12*(2), 113-117.

Verwimp, C., Tijms, J., Snellings, P., Haslbeck, J. M., & Wiers, R. W. (2023). A network approach to dyslexia: Mapping the reading network. *Development and psychopathology*, *35*(3), 1011–1025.

Watts, A. L., Poore, H. E., & Waldman, I. D. (2019). Riskier Tests of the Validity of the Bifactor Model of Psychopathology. *Clinical psychological science*, *5*(1), 3–13.

Watts, A. L., Greene, A. L., Bonifay, W., & Fried, E. I. (2023). *A critical evaluation of the p-factor literature*. https://doi.org/10.31234/osf.io/7yrnp.

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

Chapter 5: Etio-Pathophysiological Validity

5.1 Introduction

In the two preceding chapters, we were introduced to two alternative approaches to validation in psychiatric research and classification that, while both quite distinct, share a similar commitment to backing out of the validity box canyon and starting over as a means of informing a valid scientific nosology. Both approaches are motivated by what they view to be suboptimal standards associated with an overreliance on experts, poor study design, and inadequate theory development and testing in psychiatry, and instead are opting to forge their own validity paths with what they take to be the right starting point and set of guiding validity principles. Whereas HiTOP sets the "natural" hierarchical structure of psychopathology based in pure dimensional constructs derived via data models as its foundation, the network approach instead starts with theory-based network models to inform the development of formalized network theories it sees as thinking tools for the testing of network hypotheses. With each aligned with their own preferences for statistical techniques and divergent standards and procedures for validation, both claim their own approach to be the strongest candidate for dispelling psychiatry's crisis in validity.

One significantly underemphasized aspect of both approaches, however—one that lies at the heart of the original validity problem for psychiatry—is how to address the lack of etiological-pathophysiological mechanisms of psychiatric disorders. Despite advances in neuroscience and genetics research over the past forty years, psychiatric classifications remain very poorly linked to biomarkers, measurable indicators of some higher-level phenomena at a lower biological level. HiTOP suggests their own model, the HiTOP hierarchy, will better

facilitate the discovery of biomarkers such as neurobiological mechanisms beyond that of the *DSM* (DeYoung et al., 2023, p. 25). Similarly, the network approach, while strongly rejecting that psychiatric disorders can be explained solely based on their biology, holds that neural correlates may be better integrated and explained in relation to disordered systems via its theory-based network models. Furthermore, the *DSM* maintains that its updated continuous improvement model, with its elevated priority to its Biological Markers validator class, will in time contribute to the etiological and pathophysiological understanding of its diagnostic categories.

While each approach's innovations are commendable, a limited focus on spelling out precisely how each plans to successfully advance the underlying biological basis of psychiatric disorders in such a way that the underlying biology informs or becomes directly featured in its classification framework or future framework suggests that these approaches remain significantly underdeveloped in this respect. Any assertion from the *DSM*, HiTOP, or the network approach of achieving this particular sense of biological validity for psychiatry any time soon, so that we will in the near future hold confidence in our classifications as standing for some biologically based entity or system, seems no better than the claims made by the developers of the *DSM-III* in 1980 who thought the biological validation of its diagnostic categories was just around the corner.

Enter the third and final alternative approach to psychiatric research and classification: The National Institute of Mental Health's (NIMH) Research Domain Criteria (RDoC). The primary purpose of RDoC is to serve as a basic research framework "designed to integrate many levels of information (from genomics to self-report) to better understand the basic dimensions of functioning underlying the full range of human behavior, from normal to abnormal." (National Advisory Mental Health Council Workgroup on Tasks and Measures for Research Domain

Criteria, 2016). RDoC represents an alternative yet integrated approach in that it groups patients for clinical studies based on fundamental dimensions of behavior and neurobiological measures (genes, circuits, etc.). By adopting an approach that breaks up the *DSM*'s diagnostic categories like HiTOP and the network approach for researching attributes and features of psychopathology, yet unlike those approaches sets measuring the underlying biology of the phenomena as its foundation and right starting point, the NIMH hopes the RDoC will inform a more biologically valid classification system.

An important clarifying question to ask is, what does a "more valid" classification system mean for RDoC? After all, the RDoC framework does not (yet) include any diagnostic classifications, categorical or otherwise, intended to correspond to underlying disease entities or some other preferred conception of mental disorders—only organizing dimensions which RDoC conceptualizes uniquely as *concepts for investigation* and *exemplars*, and a means for doing basic research about them. Thus, the *DSM*'s application-specific method for achieving diagnostic validity of its diagnostic categories, HiTOP's Structure-First Psychometric Validity for the validation of its hierarchical constructs, and the network approach's Network Psychometric Validity and standards for theory development and testing of network hypotheses do not readily apply. RDoC's sense of validity within its current research framework, and its overall method for achieving it, are quite distinct in important respects and would benefit from additional explication. However, the NIMH changed leadership in 2015, resulting in a significant change in research priority and orientation. Thus, what RDoC meant by "more valid" and "validity" in general between 2009 and 2016, a period which arguably has had the greatest influence on the scientific and public perceptions of the RDoC program, appears to be distinct from RDoC's current validation process.

In this chapter, I provide an overview of RDoC's approach toward achieving *etio-pathophysiological validity* via its dual-track model of validation that 1) validates the tools of the RDoC framework, and 2) contributes to a future sense of syndromal validation. I then trace the development of RDoC's etio-pathophysiological validity beginning from prior cognitive neuroscience-based initiatives that informed RDoC's development through to its most recent changes with "RDoC 2.0," an unofficial term used internally among RDoC Unit members to denote the shift in approach. I further discuss RDoC's distinct sense of validity as it relates to construct validity theory (CVT), and conclude with a discussion on the philosophical underpinnings of RDoC constructs.

By offering an in-depth reconstruction of RDoC-ian validity, I hope to provide a more accurate description of RDoC's current validation process and what it considers to be the most scientific approach to psychiatric research and classification. I argue that while RDoC has long hailed itself as an integrative approach in that it combines multiple scientific disciplines, what is truly integrative about RDoC is its attempt to rebrand itself as a research framework that is open to integrating all the various senses of validity in psychiatry. RDoC includes aspects of diagnostic validity in RDoC's syndromal validity, psychometric validity in the validation of RDoC paradigms, construct validity at the level of an individual RDoC construct and construct hierarchies, expert curation in RDoC criteria and guiding principles, broader understandings of validity as a desirable quality or feature, and lastly, a willingness to allow researchers to employ their own standards of validity based on their aims and stage in research.

5.2 Overview of Etio-Pathophysiological Validity

The NIMH describes the goal of RDoC as "to understand the nature of mental health and illness in terms of varying degrees of dysfunction in fundamental psychological/biological systems" (NIMH, 2002, September 28). Within this framework, validity and validation exhibits a dual-track model of validation. The first track, which I term *biology-first function validity* (BFFV), is centered on the validation of the tools of the RDoC framework, referred to by RDoC as the concepts for investigation. These concepts include six major functional *Domains*, which represent the current understanding of the major systems of cognition, motivation, and social behavior, i.e., those systems which, when there is dysregulation and dysfunction within and across them, are thought to give rise to psychological and behavioral impairments. Each domain is accompanied by three to six *Construct*s*,* i.e., concepts summarizing data about a specified psychological or biological dimension of behavior, and recently defined as *empirical functions* (Table 7, Appendix). *Units of analysis* are the methods and instruments used to study the constructs from a "normal" to "abnormal" range of functioning. The selection of the domains in which RDoC constructs are assigned are primarily based in expert consensus, and the specific methods or tasks and paradigms of the Units of Analysis are validated mostly separately using psychometric validation. Consequently, the hallmark of validation within this first track—validation of that which is considered fundamental to the RDoC framework—is the biologically focused validation of the RDoC constructs. The second track, which is intended to be informed and supported by the first track, is the at-present open-ended and under-specified validation of a future diagnostic classification system, which I term *biology-first syndromal validity* (BFSV). While RDoC's primary aim is to guide psychopathology research, another longstanding goal of RDoC is that its research output may one day play a significant role in shaping a classification system of mental disorders. RDoC expects to contribute to future ideas of classification

(Cuthbert, 2022), such as a method for achieving syndromal validity that may be similar in respect to the method attributed to Robins and Guze and the *DSM*. Such a method, however, is not expected to be applied to a classification framework like the *DSM* with its top-down symptom-based diagnostic categories. Instead, BFSV would be applied to a classification system that can accommodate RDoC's specific hypothesis concerning mental disorders as representing broad and biologically-based heterogeneous syndromes as opposed to the *DSM*'s clinical description-based homogeneous syndromes. Taken together, both BFFV and BFSV validation tracks comprise an *RDoC-ian etio-pathophysiological validity*. Research conducted under the guiding principles of RDoC and validation of RDoC constructs using the first track (BFFV) can be expected to facilitate and support a particular sense of RDoC-ian syndromal validity afforded by the second track (BFSV).

The term biology-first applies to both tracks of validation and is in service of RDoC's primary aims of 1) developing an etiological and pathophysiological understanding of human systems of normal and abnormal functioning and 2) contributing to a future biologically based system of classification—two aims which RDoC believes the *DSM* is inherently incapable of achieving. In the first track, biology-first is evidenced by the specific criteria for RDoC constructs. For an RDoC construct to be initially selected and subsequently considered valid, it must include, among other requirements, evidence that a neural circuit or some biologically based system plays a role in implementing the function. So, while RDoC does maintain that it emphasizes the integration of validating evidence across units of analysis over any single type of evidence, a sticking point has been that biomarkers must play *some* role in the initial selection and later validation process. The same cannot be said for evidence from something like self-

report, in which Kozach and Cuthbert (2013) stated that "research that relies exclusively on self-report data would fall outside of the RDoC approach" (p. 933).

In the second track, biology-first reflects the notion that a future classification system to which RDoC will contribute will validate syndromes that have been shaped with an understanding of their biological basis. A leading frontrunner is the new classificatory concept of *biotypes*—transdiagnostic clusters defined by responses on measures across units of analyses that are "more biologically valid groupings than the diagnostic categories" (Cuthbert, 2020, p. 84). More "biologically valid" in this instance means that biotypes will yield more valid data regarding some underlying biological basis, i.e., biomarkers, of future classified syndromes. Additionally, such groupings are thought to be potentially more valid based on the standard validators of Robins and Guze such as in predicting illness, prognosis, and predicting treatment response.

While biotypes may be a leading model on which to base a future classification system, RDoC's most basic goal remains to encourage investigators to think about diagnoses in new ways—dimensional, categorical, or otherwise—and as such, is not (yet) committed to one conception of a future classification, nor an explicit validation process for syndromal validity. Using RDoC's guiding principles, researchers are expected and encouraged to propose new hypotheses and new approaches for the study and classification of mental disorders, as well as constructs not yet featured within RDoC. As a result, independent researchers will face different considerations regarding validity and validation depending on the stage of research trajectory and funding opportunity in which they are working, be it for the validation of a construct, or in the validation of behavioral paradigms for studying constructs. Given the flexibility in thinking about diagnoses as well as validation to some extent, RDoC, unlike the *DSM*, takes no official

stance toward giving priority to certain kinds of validating evidence over another (e.g., prognosis being prioritized over treatment response), aside from its stated emphasis on brain-based measures. The syndromal validity that RDoC may eventually adhere to or advocate for, if at all, may likely be proposed by independent researchers that have utilized the RDoC framework, while also being shaped by the priorities of the NIMH and the RDoC unit.

How did RDoC's overall approach to achieving etio-pathophysiological validity initially develop and how has it dramatically changed in just a short period? To address these questions, I now turn it its historical origins, beginning with two NIMH-sponsored predecessor programs on which it is based before returning to provide a more in-depth look at BFFV and BFSV.

5.3 Origins of Etio-Pathophysiological Validity: RDoC 1.0

To get a sense of RDoC's initial version of validity and process of validation, we may first look to the former NIMH-sponsored approaches: the Measurement and Treatment Research to Improve Cognition in Schizophrenia initiative (MATRICS), and the Cognitive Neuroscience Treatment Research to Improve Cognition in Schizophrenia initiative (CNTRICS). The initial focus of RDoC Project Workshop Proceedings of 2010 was designed to "build upon the knowledge base" of the CNTRICS project (National Institute of Mental Health, 2010). The first workshop was organized around the Working Memory Domain of Functioning (later updated to become Cognitive Systems) and was moderated by Cameron Carter and Deanna Barch, both former leaders of the CNTRICS project. The CNTRICS project had developed out of the New Approaches to Cognition (NAC) meeting through the MATRICS initiative with the goal of "developing measurement approaches from cognitive, social, and affective neuroscience so that

they may be implemented in efforts to develop treatments for impaired cognition in schizophrenia" (Carter & Barch, 2007, p. 7).

The rationale behind CNTRICS was that since medications targeted to treat the cognitive deficits of schizophrenia and the clinical science required to develop them was lacking, new methodological and translational approaches would increase the "precisions with which cognitive processes can be measured and related to underlying neural mechanisms" were needed (Cohen & Insel, 2008, p. 2). Moreover, recognizing a disconnect between 1) basic science on these targeted cognitive systems in cognitive psychology and cognitive neuroscience, and 2) non-neuroscience-based clinical standardized tasks being used in clinical research, CNTRICS encouraged the development and use of new tasks and new measures "that reflect the state of the art in cognitive neuroscience, and which would bridge animal models of higher cognitive function to studies using non-invasive imaging in humans" (CNTRICS, 2007, n.d.). The goal of CNTRICS, as stated by Thomas Insel, the former head of NIMH and RDoC spokesperson, respectively, was "to provide mechanisms for, incentives that motivate, and instructive examples of the translation of modern cognitive psychological and cognitive neuroscientific measures into clinically useful instruments" (Cohen & Insel, 2008, p. 3).

Building on the CNTRICS approach, RDoC set its sights on building a bigger, more expansive bridge between basic neuroscience research with clinical translational science that expanded on the work in cognitive systems. In a sense, RDoC is a generalized version of CNTRICS—one that focuses on understanding dysregulation and dysfunctions across all fundamental aspects of normal functioning behavior and potential impairments. During RDoC's initial phase of development, a total of five RDoC work groups convened between 2010 and 2011 to cover RDoC's five domains of functioning. Each working group contained participants

notably with a background in genetics, cellular and molecular neuroscience, and systems neuroscience. Each work group's responsibilities were focused on identifying and selecting dimensional constructs. Participants were tasked with the following: decide whether their group's three to six constructs should be included and, if so, how they might be revised; develop a working definition for each construct; begin to populate the elements of each unit of analysis for the constructs (e.g., genes, circuits, behaviors, etc.); and generate a list of research paradigms in the form of tasks that were thought to best characterize the component process of each construct to strengthen its validity (e.g., tasks that really measure Working Memory as opposed other similar constructs like Attention or Declarative Memory).

To move toward a more detailed understanding of psychiatric disease pathophysiology, RDoC took the approach, much like CNTRICS, of first selecting the normal behaviors, functions, and processes of interest in the form of dimensional RDoC constructs. The idea was that a more well-developed understanding of these constructs of normal functioning would establish a greater linkage when these constructs become perturbed in such a way that leads them to an abnormal state or dysfunction of the construct, and their underlying biological basis in relation to those abnormal states. For example, researchers may seek out detailed knowledge of RDoC constructs such as "reward learning" and "acute threat" on the hypothesis that when abnormally or dysfunctionally executed, such constructs contribute to something like drug abuse or post-traumatic stress. Understanding of the constructs and their biological basis in their functional capacity, it is assumed, will better aid in examining the biological basis of the constructs in a dysfunctional i.e., disordered state, resulting in the pathophysiological understanding of neurobehavioral features of psychopathology such as drug abuse and post-

traumatic stress. Such biologically based knowledge, it was hoped, would help to shape a future biologically based classification system.

Thus, RDoC in its initial phase was to serve as a kind of repository of systematized knowledge of dimensional constructs being related to underlying neural mechanisms and other biologically based measures. The *RDoC Matrix* was introduced as the organizing framework for this purpose, whereby domains of functioning and constructs were decided upon via consensus building during RDoC workgroup meetings. For a construct to be added to the RDoC matrix, there must be strong evidence "for the validity of the suggested construct itself [as a behavioral function," and "that the suggested construct maps onto a specific biological system, such as a brain circuit" (Cuthbert and Insel, 2013, p. 6).

Once the domains and constructs of the matrix were set, *elements* of the matrix, being measurements of the construct across various units of analysis, could be used to populate the matrix. The RDoC matrix, understood at the time as an "integrative scientific literature" (NIMH, 2011, September 9) could then serve as a means for orienting one's current research to fill in and contribute to a growing body of knowledge regarding RDoC constructs. While the matrix was never intended to serve as a classification system, the information populated into the matrix would eventually come to be reified in such a way that it would reflect a compendium of RDoC features. Whereas the *DSM* focuses on validating diagnostic categories and their diagnostic criteria using diagnostic validity, RDoC as it was first formed focused on validating the RDoC constructs assigned by the RDoC workgroups. In addition, the validation of measurement approaches often referred to as tasks used to study those constructs was of essential importance, as evidenced by the third aim of NIMH's Strategic Goal 1.4, "develop reliable and valid measures of these fundamental components of mental disorders for use in basic studies and in

more clinical setting" (Cuthbert & Insel, 2013, p. 4). The primary goal for RDoC research, however, was "to promote research employing clinical subjects from multiple diagnostic groups appropriate to the research question, to establish construct validity for some proposed organizing dimensions of psychopathology" (NIMH, 2011, September 9).

The establishment of construct validity of RDoC constructs was considered a top priority for RDoC, with several RDoC funding calls focused on the validation of RDoC constructs. Along the way, there were acknowledgments that what was considered most important about RDoC wasn't the matrix per se, but "the idea of freeing up investigators to pursue exciting translational research questions driven by neuroscience and behavioral science rather than by constraining sets of symptom clusters" (Cuthbert, 2014, p. 35), whereby translational research was understood to support turning basic science into deliverables in clinical medicine, i.e., from bench to bedside, most commonly through the utilization of animal models by which to study RDoC constructs. However, the populating of the matrix with constructs validated across multiple units of analysis was nevertheless still viewed as a metric of scientific progress achieved via research conducted through RDoC-ian principles (Table 8).

The RDoC construct validation process was initially characterized loosely. Validation of RDoC constructs was taken to mean establishing construct validity, which, for RDoC constructs, was defined as support from valid evidence from convergent measurement across multiple units of analysis. The criteria for evaluating RDoC constructs included that 1) constructs must be evaluated across at least two levels of analysis; 2) one of those units of analysis needed to be biological (i.e., genes, molecules, cells, circuits, physiology); and 3) the specific methods for assessing (i.e., measuring) the constructs should assume dimensionality of the constructs. While the various units of analysis were deemed of equal importance, the initial prioritization for RDoC

was on establishing a relationship between RDoC constructs and neural circuits that implemented those constructs. Levels of analysis proceeded up from measures at the level of neural circuitry functioning to the levels of behavioral and self-report measures of clinically relevant variation" or down to the lower levels of analysis such as to genetics and cellular and molecular mechanisms (Insel et al., 2010).

Empirical evaluations of construct validity of RDoC constructs primarily came from research focused on convergent validation, that multiple units of analysis of an RDoC construct would converge on that construct. Convergent validation for RDoC was assessed by examining relations among purported elements of the same RDoC construct to see if they are predictive of that construct. Consider an example referenced by Lilienfield & Treadway (2016): an RDoC study may demonstrate that the fear-potentiated startle (FPS), defined as a reflexive physiological reaction to a presented stimulus and an element featured within the Physiology units of analysis under the Acute Threat ("Fear") RDoC construct, is related to re-experiencing, avoidance, and hyperarousal symptoms, being elements across the Behavior and Self-Report units of analyses. Such findings of convergence across units of analysis would be considered to provide convergent validation that the fear-potentiated startle is an indicator of the RDoC construct Acute Threat ("Fear"), and thus, an establishment of construct validity of both the FPS as an element of Acute Threat ("Fear") and the RDoC construct itself. (Norrholm, 2015). Explicit standards for what would count as being predictive of, relating to, or across units of analysis were not provided, and instead it was left up to the researchers to employ their own standards. Thus, RDoC 1.0 initially espoused a version of etio-pathophysiological validity that centered exclusively on an underspecified convergent validation of its constructs in the RDoC to achieve what it considered to be construct validity. Under this sense of validation, biological

markers were heavily prioritized, as was the validation across multiple levels of analysis, both of which could be seen as foundational for its sense of validity and as desired characteristics or qualities for its validation process.

5.4 Origins of Etio-Pathophysiological Validity: RDoC 2.0

When RDoC first launched, it was proposed as a game-changing alternative to that of the *DSM*, but its strategic goals and overall mission began to change, most notably with the departure of then NIMH Director and Head of the RDoC Unit Thomas Insel in November of 2015. Insel stated that his two primary goals while director of the NIMH had been to "integrate neuroscience and psychiatry and create a new discipline of clinical neuroscience," (Moran, 2015, October 7). Bruce Cuthbert then became the NIMH interim director as well as the Head of the RDoC Unit. Joshua Gordon took on the directorship position in 2016, and Bruce Cuthbert remained Head of the RDoC Unit. An initial change preceding Insel's departure was a notable softening of RDoC's opposing relationship to the *DSM*. Whereas initial RDoC funding opportunities stated that "applications that focus primarily on validating a *DSM* diagnostic category will NOT be considered" (NIMH 2011, emphasis in original) and that RDoC research was intended to support the future development of an alternative classification system separate from the *DSM*, Insel and colleagues began to view RDoC as potentially working in conjunction with and even in service of the *DSM*. In a joint press release from the NIMH and APA issued in May 2013, Insel and then APA President-elect Jeffery Lieberman stated that "all medical disciplines advance through research progress in characterizing diseases and disorders. *DSM-5* and RDoC represent complementary, not competing, frameworks for this goal." In an interview with *Psychiatric News* in October 2015, just before his departure, Insel stated that "RDoC is a

guide to rethinking the way we do diagnosis and may inform *DSM-6* or -*7*, but for now clinicians should be using the *DSM* and *ICD*" (Psychiatric News, 2015, October 7).

With a change in leadership at the NIMH and RDoC, Bruce Cuthbert and colleagues took the opportunity to reflect on several criticisms of RDoC under Thomas Insel. One primary criticism centered around RDoC's overemphasis on the biological components of function and dysfunction at the level of the individual. Five of the seven units of analysis focused on biological indicators, and there was concern that the biological units of analysis were receiving excessive priority (Berenbaum, 2013). As a result, psychosocial variables such as the social, developmental, environmental, or cultural context were perceived as not being adequately represented (Hershenberg & Goldfried 2015, Lilienfeld 2014). In addition, the emphasis on the biological units of analysis was interpreted as inferring that those units were best suited for construct validation above non-biological measures. Measures that incorporated neural circuitry were seen *a priori* as inherently "more valid" measures of an RDoC construct than measures based in self-report and without a clear sense as to why this is the case. Establishing biomarkers may be a priority of RDoC, but, as per Lilinfield and Treadway (2016), "there is no inherent reason why self-report measures, which can readily capitalize on aggregation across indicators of behavior, cognition, and emotion across diverse situations, cannot provide equally—or more—construct-valid measures of biological systems relative to biological markers of these systems" (Lilienfield & Treadway, 2016, p. 450).

Apart from the concerns of RDoC being too reductionistic and neurocentric, i.e., assuming assessments based in neuroscience are more construct-valid and are thus of greater scientific value, members of the RDoC unit were worried about RDoC constructs and the matrix in general, and decided to prioritize those topics. Shortly after Insel left in 2016, RDoC

maintained that "the long-term goal is to develop a scientific base that can inform future neuroscience-based diagnostic systems of mental illness," (National Advisory Mental Health Council Workgroup on Tasks and Measures for Research Domain Criteria, 2016) thus reiterating the purpose of RDoC as serving as a systematic database. But more and more, this long-term goal was beginning to be seen as untenable. In reflecting on the shift from Thomas Insel's vision for RDoC to what has come to be described as "RDoC 2.0"—a term never officially used in publication but one that is used informally among members of the current RDoC Unit—Sarah Morris, Associate Head of the RDoC Unit commented on the way RDoC shifted away from thinking of the matrix as a kind of data repository:

> When RDoC was first launched and developed, it was focused on a grid framework, the RDoC matrix, in which we set the domains, constructs, and units of analysis. Then, in a series of workshops, we filled in the elements of the matrix. Very *quickly*, unfortunately. These workshops were only a day and a half long, and we ended up at the last-minute filling in elements in the matrix that we thought were reasonable. These elements then got reified to a degree that we really weren't comfortable with. We did some work attempting to curate those elements in the matrix in 2016, but there were still a lot of big gaps: Why should something be an element in the matrix versus not? How do you add an element? Is there a process for evaluating new elements for the matrix? We didn't have the bandwidth to maintain the matrix at that level of detail. So RDoC backed off from the matrix and attempted to communicate to the field that what's in the matrix should be only thought of as exemplars. RDoC is not meant to be a repository. (Morris, 2023)

The biggest shift in perspective came from the realization that 1) much of the RDoC matrix, while originally thought to have been empirically informed, had simply been considered good enough by expert consensus, and 2) an adequate validation process for evaluating RDoC constructs so that they could be considered to represent a kind of systemized knowledge base was beyond the capacity and scope of the RDoC Unit. To address the potential problem of reification of the matrix, the RDoC unit under Bruce Cuthbert formed The Changes to the RDoC

Matrix (CMAT) Council Workgroup in 2016. The purpose of the group was first to develop standardized methods for submitting proposed changes for RDoC's Domains and Constructs, and second, to develop and carry out a validation process for evaluating proposed changes, which CMAT would then submit to the ongoing National Advisory Mental Health Council (NAMHC) workgroup for approval.

There have been notable revisions to the RDoC matrix, such as the addition of a sixth sensorimotor domain, revisions to the Positive Valence Domain (NAMHC, 2018), and a conservative scaling back of the Genes units of analysis from the matrix. Yet, the RDoC Unit would ultimately take the initiative in 2019 to quietly abandon the matrix. RDoC's primary organizing tool was instead rebranded as the RDoC Framework, in which a new graphic of concentric circles would replace the rows and columns of the former matrix while still featuring all the original components from RDoC (Figure 4, Appendix). The RDoC Framework graphic communicates the new way of thinking about the purpose of RDoC in organizing psychiatric research. First, RDoC is not to be thought of as a repository—there is no table of rows and columns which to populate and fill in. Instead, RDoC is just the RDoC Framework, an organizational structure for doing psychiatric research, along with some guiding RDoC-ian principles that allow researchers to explore new ways of conceptualizing and developing psychiatric diagnoses. Consequently, RDoC has significantly scaled back its aims for directly demonstrating scientific progress, so much so that is now unclear how progress in psychiatric research under RDoC should be measured. Second, the three concentric circles of Units of Analysis, Environment, and Development across the lifespan symmetrically overlapping with the RDoC Domains is reflective of RDoC viewing all aspects of the analysis of domains of functioning as equally important. Prioritization on neural circuitry (and thus biology-first) would

remain, but the additional aspects that shape psychopathology such as psychosocial and environmental variables across the development of the lifespan were better articulated and emphasized in a way that was intended to communicate that RDoC is not to be thought of as only a neuroscientific endeavor.

5.5 Biology-First Function Validity (BFFV)

With RDoC 2.0 came RDoC's updated validation process, one I have interpreted as Biology-First Function Validity (BFFV), which we return to now. BFFV may be separated into two primary phases. The first phase, which I term *construct selection*, is the selection and justification of appropriate RDoC constructs based on a set of criteria. In addition to constructs still needing to demonstrate "evidence for a functional behavioral or psychological construct" and "evidence for a neural system or circuit that plays a major role in implementing the function" of the construct (NIMH, 2022b), RDoC 2.0 has adopted two new sets of criteria, as discussed below.

First, RDoC constructs must now show evidence for "a putative relationship to some clinically significant problem or symptom" (NIMH, 2022b), i.e., there must be evidence from prior studies that the extremes of functioning of a construct are related to psychological or behavioral impairment in some way (e.g., anhedonia, hallucinations, social cognition). Toward realizing the clinical impact of an RDoC construct, clinically relevant constructs must further demonstrate "translational potential" in that they can be studied in ways that are more meaningful and more applicable to benefit human health. Second, RDoC constructs must "demonstrate external generalizability to real-world behavioral assessments" (NIMH, 2022b). A longstanding criticism of RDoC was that despite its emphasis on its translational potential,

RDoC studies are too biologically-focused such that findings are not readily applicable to or inclusive of more clinically relevant measures (i.e., behavioral assessment). Thus, when reviewing RDoC recommended tasks and measures to nominate as a part of *RDoC's First Generation RDoC Measurement Elements*, RDoC Work Groups were instructed that "whenever possible, the measures should allow for behavioral assessment, as opposed to focused solely on biological signals (e.g., neuroimaging)" (NIMH, 2016). In current RDoC funding opportunities such as NIMH PAR-23-307: "Computational Approaches for Validating Dimensional Constructs of Relevance to Psychopathology," computational models derived from lab-based behavioral tasks also require additional testing for generalizability to behavioral data from real-world settings to achieve greater translational potential. The focus on real-world data is "to validate models developed from lab-based behavioral tasks" (NIMH, 2022b). Thus, the initial selection of an RDoC construct, based on now four specific criteria that must be evidenced by valid data, may be thought of as comprising the first phase of validation. Once a construct is selected, the second phase of BFFV, which RDoC refers to as *construct evaluation*, is an assessment of what RDoC interprets as the construct validity of an RDoC construct. This process must include both convergent and divergent validation using at least two levels of analysis, using the same levels (including at least one brain-based measure) across the construct(s) to be validated.

In RDoC 2.0, validation of RDoC constructs is no longer focused exclusively on convergent validation, i.e., demonstrating that certain elements across units of analysis are predictive of a specific RDoC construct, which was most notable in RDoC 1.0 (Shankman & Gorka, 2015) and had been criticized as insufficient (Lilinfield & Treadway, 2016). Instead, divergent validation, i.e., demonstrating that certain elements across units of analysis are not associated with non-hypothesized constructs, is considered equally as important. RDoC further

does not provide a specific checklist of validating evidence per se, but only initial guidelines by which researchers may propose and carry out validation research of RDoC constructs.

In conjunction with the two primary phases, two additional supporting kinds of validation are now encouraged by RDoC. The first is a type of *hierarchical* validation, whereby validation research is focused not on an individual construct, but on the "relationships between domains, constructs, and subordinate sub-constructs, both in terms of their correlational structure and their underlying neural circuitry" (NIMH, 2022b), i.e., their hierarchical structure. To assess such relationships, RDoC encourages what it refers to as "computational approaches for validating dimensional constructs of relevance" (NIMH, 2022b) in which researchers use quantitative, machine-learning approaches to achieve validation. For example, in validating the existing relationship between multiple constructs, RDoC researchers may:

> Perform unbiased data-driven validation of existing constructs that may involve merging, subdividing, or hierarchically organizing them by integrating data between and within constructs. The results of such studies may indicate that no changes to existing constructs and their organization are needed, but these studies will provide a better understanding of the relationships between constructs. (NIMH, 2022b)

The second type of validation research is a validation of specific measures and tasks within RDoC's units of analysis. RDoC's current focus with this line of validation research is on the validation of behavioral paradigms used for studying RDoC constructs. In the past five years, a focus on computationally informed behavioral paradigms that "have the potential for back-translation from humans to animals" (NIMH, 2022c) became of particular interest. "NIMH is interested in the development of a new set of computationally-informed behavioral paradigms and/or the deployment of novel computational models to existing paradigms to capture dimensional aspects of mental-health relevant behaviors" (NIMH, 2022c).

Computationally informed behavioral paradigms, which would fall within RDoC's Behavior units of analysis, are in the form of behavioral mathematical models that are developed to represent and predict behavioral outcomes related to RDoC constructs. Such models are thought to provide the computational rigor necessary to model the richness and variability in human behavior that other behavior measures typically lack, while also having a greater potential to generalize from lab-based behavioral measures to data collected in real-world settings. To best position the models for future back-translation studies from human animals to "advance novel therapeutic strategies" (NIMH, 2022c), model parameters are expected to be linked to underlying neurobiology as much as possible. The aim for this kind of validation research is to develop "a library of behavioral assays that can be used in both humans and animals" to assess RDoC constructs and "test hypotheses regarding neurobiological mechanisms," thus facilitating the ability of psychiatric researchers "to move research bi-directionally between humans and animals" (NIMH, 2022c).

In sum, Biology-First Function Validity (BFFV) can be thought of as primarily encompassing a two-phase process of construct selection and construct evaluation, that is supported by the validation of paradigms used to select and evaluate those paradigms, as well as the validation of the structural relationship between constructs. While psychometric validation of tasks may be characterized as separate from the overall validation process, it may be reasonable to see this as a part of BFFV insofar as the biology-first emphasis is maintained at the level of validating the measurement tool. What is unique regarding the psychometric validation of specific paradigms is that there exists a broader notion of validity in play that reflects the interest by RDoC in a particular methodology, as is the case of computationally informed behavioral

paradigms. Those tasks which are considered more valid are in part simply those which are more desired or prioritized.

5.6 Biology-First Syndromal Validity (BFFV)

RDoC 2.0's Biology-First Function Validity, which is used to validate RDoC constructs, may be thought of as informing what I term Biology-First Syndromal Validity (BFSV). A way to think about this kind of validity is in the form of an answer to the question, "If an RDoC-informed psychiatric classification system did exist, what kind of validation process would RDoC recommend be utilized?" The general idea around BFSV is that the empirical validation process for establishing syndromal validity of an RDoC-informed classification would be somewhat similar in principle to that of Robins and Guze's general formulation for achieving diagnostic validity, but one that is applied toward a biologically-based nosology. The hope is that RDoC-informed research will inform novel psychiatric categories or clusters that are not primarily shaped by clinical descriptions of signs and symptoms like *DSM* categories, but instead by research into well-validated RDoC constructs that have been strongly linked to brain-based measures. The new categories and clusters, developed via a biology-first validation approach, would be more informative in predicting prognosis and treatment response, and thus, show higher degrees of syndromal validity, making them more valid in this sense.

An application of a validation process on which BFSV may eventually come to be modeled—one which RDoC previously referenced as demonstrative of thoughtful work on validation in psychiatric research but has recently come to the fore in RDoC 2.0—comes from the Bipolar and Schizophrenia Network for Intermediate Phenotypes (B-SNIP) project. B-SNIP is a research consortium with five study sites across the US that began just before RDoC. The

specific aim of the B-SNIP consortium is to develop novel classifications of psychotic and mood disorders based on biological measurements that would amount to a "more valid scientific nosology for psychoses" (Keshavan et al., 2013). B-SNIP's four-stage process for developing such classifications as depicted in Keshavan et al. (2013) is as follows:

B-SNIP Process for Developing a Biologically Based Scientific Nosology

1. Agnostic deconstruction of disease dimensions, identifying disease markers and endophenotypes.
2. Mapping such markers across translational domains from behaviors to molecules
3. Reclustering cross-cutting bio-behavioral data using modern phenotypic and biometric approaches.
4. Validating such entities using etio-pathology, outcomes, and treatment-response measures.

Stages 1 and 2 of B-SNIP's process appear remarkably similar to the overall RDoC approach. Indeed, B-SNIP, while specific in application toward psychoses, may be thought of like CNTRICS as a specific research approach on which RDoC was partially based. Today, B-SNIP frames the way RDoC thinks about its relation to a future classification system, with Cuthbert (2020) adopting B-SNIP's recent terminology of biotypes in reference to "more biologically valid groupings" (p. 84) that he sees as eventually needing to be validated before entering the clinic.

Other than B-SNIP's far narrower focus in application on psychotic and mood disorders, the main difference between B-SNIP and RDoC is that B-SNIP uses its understanding of the dimensionality of specific disorders to take the next step in locating biologically similar clusters within a dimension and to draw on such clusters to develop biologically defined disease groups, referred to as biotypes (Stage 3). In terms of validating the biotypes (Stage 4), which would amount to a specific application of BFSV to achieve what might be termed *biotype validity*, leading B-SNIP researcher Carol Tamminga (2014) views this stage as particularly difficult:

> The development of validating characteristics for the clusters is the research challenge, namely a common systems understanding or a unifying molecular pathology for these biomarker clusters. The BSNIP approach begins dimensionally, using dense biomarker characterization, to form biologically common clusters, potentially useful as disease identifiers with biological targets. (Tamminga, 2014)

Interpreting this comment, validation of the biotypes for the B-SNIP approach may be distinct from Robins and Guze's method in its strong emphasis on establishing underlying etio-pathology of the biotypes as being the primary validating source for BFSV or biotype validity. On the other hand, Keshava et al. (2013) posit the following recommendations for the validation of biotypes that are much aligned with the traditional validators associated with diagnostic validity:

> A key step would thus be to validate agnostically derived clusters of disease characteristics by mapping them back to clinical features and "external" validators such as course, treatment response and etiological data (i.e., variation in genetic and environmental risk factors). In doing so, the clinical characteristics that map on to the biological characteristics (i.e., biotypes) may well not resemble conventional diagnostic categories but rather reflect cross-cutting dimensions (e.g., fear, impulsivity, and aberrant salience) (Keshava et al., 2013)

In sum, although a specific version of BFSV has not been explicitly described or endorsed by RDoC, one can interpret what the RDoC Unit views favorably in terms of syndromal validity as being reflective of the type of syndromal validity that could be endorsed if RDoC were to eventually inform a classification framework. One such model comes from the B-SNIP consortium's focus on the validation of what it refers to as biotypes, and this model shares some similarities with Robins and Guze's method for achieving diagnostic validity. While RDoC is favorable to the B-SNIP approach to classification, RDoC itself remains in a facilitating role of promoting a variety of different kinds of approaches from researchers who aim to develop and validate biologically informed disease groupings, and thus, their notion of BFSV may change considerably. Reflection on RDoC's conception of BFSV is additionally important since RDoC

is always attempting to set itself up for success. What RDoC interprets as necessary for the validation of biotypes may inform its approach to the validation of RDoC constructs.

5.7 Etio-Pathophysiological Validity: Relation to Construct Validity Theory (CVT)

The current focus of RDoC's validation process, which I characterized as Biology-First Function Validity (BFFV), centers on the validation of the dimensional RDoC Constructs, defined as empirical functions, that may further reflect "behavioral elements, processes, mechanisms, and responses" (Lobo et al., 2023, p. 1) which "comprise different aspects of the overall range of functions" (p. 5). With a focus on constructs, one might assume the validity of a construct may be understood straightforwardly as a kind of construct validity. But what kind of construct validity is this really referring to? After all, RDoC constructs are not tests, assessments, or even clinical instruments themselves in the traditional sense, and instead are thought of as dimensional empirical functions that also serve as research concepts for investigation. So, a notion of construct validity in a very technical sense as it would feature in something like psychometric validity does not readily apply. At the same time, the notion of RDoC constructs as exemplars, and the de-emphasis by RDoC of thinking of the RDoC constructs as part of a systematic base of scientific knowledge, makes RDoC constructs potentially fall more within the realm of a type of conceptual research instrument or heuristic device. Thus, there is a potential technical sense of construct validity that may be appropriate in some respects.

In attempting to excavate the use of RDoC constructs in practice and characterize them, Slaney (2017) groups the use of construct concept into three broad categories. The first is *conceptual* (constructed abstractions, hypotheses, models, theories), in which constructs are "tools used by researchers to designate and communicate about a specific domain of inquiry" (p.

115). The second, is *objective* (real, but unobservable or not directly measurable), in which constructs are portrayed as objective features of reality which researchers consider as the attributes or features to be "investigated," "explored", and "measured" (p. 115). The third, is *focal phenomenal*, in which constructs more broadly refer to the general subject matter of a research domain and, as with the first category, "the ontology of the construct is left fairly ambiguous" (p. 115). Constructs may fall into one or multiple categories, as it is often common practice for mixes of different uses of construct to be present within the same account.

Applying the categories of constructs from Slaney (2017) to RDoC constructs, it appears that RDoC constructs demonstrate a mixture of all three uses. First, RDoC constructs are very much defined in a conceptual sense, intending to serve as exemplars of the empirical functions rather than designated as the specific functions themselves. They are constructed abstractions, and additionally coincide with what Lovasz and Slaney (2013) found to be the two senses in which researchers think of *hypothetical* constructs, namely, "theoretical, constructed, heuristic" and "conjectural, provisional, open" (p. 116). RDoC constructs are additionally thought of as representing focal phenomena, in the sense that RDoC constructs are at times defined to reflect a more general domain or area of study. For example, the Loss construct, featured within the Negative Valence Systems Domain, is defined as "a state of deprivation of a motivationally significant con-specific, object, or situation" (NIMH, 2014) where "Loss" in a more general sense covers a broad variety of types of loss (e.g., loss of shelter, behavioral control, loved ones) as well as response to loss (e.g., grief, sustained loss). Lastly, RDoC constructs are discussed in an objective and real sense as evidenced by attempts to not only measure and validate such constructs via biological measures but at times define such constructs in terms of their underlying biology. For example, the RDoC construct Reward Learning, defined by RDoC as the

"process by which organisms acquire information about stimuli, actions, and contexts that predict positive outcomes" (NIMH, 2014), is at times equated with the biological and neural mechanisms that are hypothesized to implement the function. That is, not only is it assumed that RDoC constructs like Reward Learning refer to some real process or attribute that can be investigated and measured, but Reward Learning is often equated with its referent, i.e., the thing that the construct is meant to represent, being the mechanisms. The constructs in which there is a considerable amount of prior neuroscientific work concerning that construct tend to be those that are characterized in this sense.

In terms of BFFV's relation to contemporary construct validity theory (CVT), RDoC appears to have taken a similar approach to many psychological researchers who, in utilizing construct validity, ultimately pick and choose specific aspects from contemporary CVT to serve their specific validation aims. There are six principles in which BFFV reflects contemporary CVT as featured in *Standards for Educational and Psychological Testing* (2014). First, BFFV takes a unified conception of validity, from Loevinger (1967) who first articulated that "construct validity is the whole of validity from a scientific point of view" (p. 636) and Messick (1998) who states, "all validity is of one kind, namely, construct validity" (p. 37). Second, BFFV sees validity as a matter of a degree. Third, BFFV views validity as not a property of the construct itself but, as per Cronbach (1971), an interpretation of the data from a specific measurement procedure. Fourth, BFFV does not distinguish between kinds of validity and instead understands validity as coming from differing sources of validity evidence. Fifth, BFFV sees the process of validation as involving the gathering and evaluating of evidence to support an intended inference of the data, weighing both favorable validity evidence and that which is disconfirming. Sixth, the validation process is considered an ongoing endeavor.

RDoC's two primary sources of validity evidence within BFFV are thought to come from what they refer to as convergent and divergent sources of validity evidence, which essentially amounts to mapping RDoC constructs onto and across various units of analysis that are hypothesized to be associated with that construct (convergent), while also showing that RDoC constructs, as predicted, do not map onto certain units of analysis hypothesized to not be associated with that construct (divergent). This source of validity evidence closely mirrors Convergent and Discriminant Evidence featured in the *Standards* (2014) within the Evidence Based on Relations to Other Variables category of evidence, also commonly referred to as evidence for the external validity of a construct in psychological research. Unlike CVT which features four other sources of validating evidence (Evidence Based on Test Content, Evidence Based on Response Processes, Evidence Based on Internal Structure, Evidence Based on Consequences of Testing), BFFV chooses only to focus on Evidence Based on Relations to Other Variables, except in the instance of validating the hierarchical relationship between constructs, in which Evidence Based on Internal Structure is invoked. Additionally, BFFV deviates from contemporary CVT in that the degree in which data from biological measures are integrate" from multiple studies and sources of evidence appears to be considered a source of validating evidence. Something like Evidence Based on the Integration of Biological Variables could be thought of as its own unique RDoC-ian category for a source of validating evidence within BFFV.

Another source for establishing validity within BFFV that is not typically conceived as a source of validating evidence, but for BFFV may be appropriate to think of it as such, is evidence based on establishing the construct validity (in the psychometric sense) of the RDoC measurement instrument itself, be it a biological measure, behavioral paradigm, behavioral task,

or self-report measure. The idea that validity may be established regarding the measurement instrument and simultaneously in the construct of interest has been discussed in the initial *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (1954):

> One tends to ask regarding construct validity just what is being validated—the test or the underlying hypothesis? The answer is, both, simultaneously. If one predicts an empirical relation by supposing a certain personality organization, the verification of this prediction tends to confirm both the component suppositions that gave rise to it. (APA, AERA, NCM, 1954, p. 15)

For RDoC and BFFV, RDoC tests, instruments, and assessments that are evidenced to provide valid measures of RDoC constructs are considered essential. After all, what good is validity evidence if that evidence comes from a poorly validated measurement instrument to begin with? In a very general sense, more valid and standardized measures in principle can always be assumed to provide better sources of validity evidence, i.e., valid measurement instruments support the validation of constructs. One specific RDoC unit of analysis where the validation of the instrument itself may be thought to provide additional sources of evidence for the validity of RDoC constructs is in the development of validated behavioral paradigms and tasks. For RDoC such tasks (behavioral, affective, and cognitive) are used often in RDoC research and are expected to meet relatively standard psychometric and generalizability criteria. Additionally, tasks are not required but encouraged to 1) be capable of being used with methods to interrogate brain circuitry (e.g., functional magnetic resonance imaging [fMRI] and EEG, and 2) demonstrate known relationships between task performance and neural signal(s). The thought is, if we include in RDoC tasks' validating criteria that tasks are shown to more readily interrelate with other biological measures, then the validation of those tasks may in turn be thought to serve as a source of additional validating evidence of RDoC constructs within BFFV. In other words, constructs measured using tasks that could more readily relate to other biological

measures, as opposed to tasks that do not, could be judged to have a higher degree of construct validity. This would amount to a source of validity evidence in the form of "Evidence Based on Biologically Relevant Behavioral Tasks."

At the same time, RDoC tasks are not necessarily selected and validated with the primary aim of helping RDoC constructs achieve the highest degree of construct validity. An additional not required but encouraged criteria for RDoC tasks is whether the task can be used in non-human animals, or whether a non-human animal version of the task is available. Thus, certain types of tasks end up being selected and prioritized not due to their overall construct validity and/or relation to biological measures, but how well they can be utilized in animal studies. "Although workgroup members acknowledged the importance of self-report measures of PVS constructs, performance-based or behavioral tasks were prioritized to maximize potential translation to and back-translation from preclinical (animal) models" (NAMHC, 2016). One area for further evaluation is to what degree RDoC units of analysis are selected and prioritized not based on their overall validity, but for serving other aims such as being amenable to specific types of experiments and studies with non-human animals. We would like to think that validation of RDoC constructs is achieved by selecting the right task for the job (construct validation). But if the task has been selected for another purpose in mind (back-translation), where a broader notion of validity is being used that reflects those which are desirable qualities or features, how might that impact RDoC researchers' validation of an RDoC construct?

5.8 Philosophical Underpinnings of RDoC and Etio-Pathophysiological Validity

The majority of philosophical criticism toward the RDoC approach centers on RDoC when it initially appeared to have adopted a model of explanatory reduction (Sarkar, 1992), the

notion that RDoC's constructs may be sufficiently explained by the underlying neurobiological processes at the lower levels of analysis. Statements made by former NIMH director Thomas Insel who aimed to reconceptualize psychiatry as clinical neuroscience seemed to warrant such criticisms by those who saw RDoC as unjustifiably prioritizing certain kinds of biological evidence and explanations over those at the psychological, behavioral, developmental, and environmental levels. In response, RDoC has pushed back, claiming that its biology-first approach does not necessitate such overly reductionistic commitments and that it is only promoting a more thoughtful and precise integration of biological measures with other useful measures of psychiatric phenomena. An indication of this latter commitment is the shift from RDoC 1.0 of Thomas Insel to RDoC 2.0, which reflects additional changes in the philosophical status of RDoC constructs. In what follows, I touch on how RDoC's shift in approach has resulted in a change in some of its underlying philosophical positions, and how that might coincide with its validation process and overall goals of one day informing a biologically valid psychiatric classification system.

RDoC Constructs and Scientific Realism

With RDoC 1.0 (2009–2016), the RDoC matrix was intended to serve as a scientific repository of knowledge accumulated from the construct validation of its RDoC constructs, which represented organizing dimensions of psychiatric phenomena. As a result, RDoC held an ontologically realist position regarding their constructs. For example, consider the Loss construct, described as "a state of deprivation of a motivationally significant con-specific object or situation" (NIMH, 2014). Under RDoC 1.0, the Loss construct, although initially selected via expert consensus, was considered a real or true feature of the world. The inclusion of constructs

in the matrix wasn't simply instrumental—rather, the constructs featured in the matrix were selected for their assumed standing of real aspects of functioning and dysfunction thought to be related to real psychiatric phenomena. Furthermore, RDoC Domains that organized constructs and their relation to one another were also considered real in this sense.

Additionally, RDoC 1.0 was epistemically scientific realist regarding the constructs, so that claims interpreted literally about their constructs, when produced with the appropriate epistemic criteria, would establish knowledge of psychiatric phenomena to be added to the matrix as it exists in nature. In contrast to the quantitative empirical methods and structural validity that HiTOP utilizes to discover the true nature of psychopathology, however, what mattered for establishing the reality of RDoC constructs was relating them to underlying neurobiological measures through convergent validation across multiple units of analyses. While operationalized, the existence of RDoC constructs as real attributes or features ultimately justified their inclusion in the matrix insofar as there was "strong evidence that the suggested construct maps onto a specific biological system, such as a brain circuit" (Cuthbert & Insel, 2013, p. 6) which made them real. Appropriately validated RDoC constructs could then ideally be drawn upon for further study, so that more and more elements could be populated into the matrix that as evidence of the existence of the constructs. This process would place RDoC in a better position to conceptualize a future classification system based on accumulated knowledge around its constructs.

RDoC 2.0 (2016–), however, did away with the concept of the matrix and dropped the idea of RDoC as serving as a kind of scientific repository. RDoC has since replaced its scientific ontological realist commitments around its constructs with a constructivist-realist stance. Constructs as concepts for investigation and exemplars essentially denote that they are very

much *constructed* and intended to serve as tools for researchers to utilize rather than as things that necessarily exist. This shift in underlying philosophy, however, may ultimately be more of a pragmatic methodological move that attempts to circumvent any further reification by researchers who utilized RDoC constructs initially selected via expedient expert consensus. That is, there is still a sense that RDoC wants it both ways. It wants to maintain that its constructs, when understood as "empirical functions" are real—e.g., that there really is something that is Loss out there and that responses to it may be dysfunctional. At the same time, it wants us to be scientific realists about future classifications built on this research, too. After all, why be biology-first, an approach that assumes the reality of the phenomena is evidenced by the relation to their underlying biology about the constructs if, at the end of the day, the constructs are only heuristic devices?

At present, however, RDoC is extra cautious given psychiatry's history of reification of its concepts (e.g., *DSM* diagnoses). RDoC has thus additionally replaced its epistemic scientific realism about its constructs with epistemic instrumentalism (Lockard, 2013), so that its constructs are understood as tools or instruments that are selected and revised to achieve certain scientific aims. In this case, RDoC's central aim is now to serve as a conceptual tool repository for researchers to draw upon to explore the "varying degrees of dysfunction in fundamental psychological/biological systems" (NIMH, 2022a), whereby RDoC can provide the right investigative concepts (RDoC constructs) and tools for measuring them (RDoC tasks and paradigms) for researchers to start with and guidelines for how to appropriately use them.

Etio-Pathophysiological Validity and Constructivist-Realism

Per Markus and Borsboom (2013), a change in how we think about constructs may result in how we go about validating them. For RDoC, this change has both benefits and drawbacks. The main benefit is that the underlying philosophy of RDoC's sense of its BFFV track of etio-pathophysiological validity, which centers on the validation of its constructs as validation of the tools or instruments of the RDoC Framework, is now better aligned with the RDoC 2.0 approach overall. Both BFFV, which draws on select aspects of contemporary standards of CVT, and RDoC 2.0 maintain a constructivist-realist stance in which Messick (1981) sees validated constructs as "heuristic devices for organizing observed relationships with no necessary presumption of real entities underlying them" (p. 583). RDoC's more conservative approach thus now permits greater consistency between its validation process, centered on the convergent and divergent validation of its constructs via construct validity, and its overall scientific approach. If we set aside the persistent conceptual ambiguity regarding what it really means to think of constructs as exemplars vs. empirical functions and just think of all of the facets of RDoC (Domains, Constructs, or tasks and paradigms within the Units of Analysis) as simply instruments or tools to be validated via construct validity, then we may have greater confidence that a straightforward program in contemporary construct validity, BFFV, is appropriate.

The main drawback with taking a more conservative and instrumental approach to one's constructs, however, is that RDoC may be further away from informing a future biologically-based classification system for which it may apply its BFSV track of etio-pathophysiological validity. With RDoC 1.0, there was at least the hope that with the RDoC matrix, RDoC was building toward something with its validated constructs. With RDoC constructs conceived and validated as tools or instruments, however, it's now arguably less clear how RDoC constructs will inform or feature in a classification system at all, or if that is still a goal. For example, with

an approach such as B-SNIP, it's unclear how one can go from taking evidence of RDoC constructs understood only as exemplars to reclustering phenomena from such evidence into "cross-cutting bio-behavioral data using modern phenotypic and biometric approaches" (p. 10), i.e., its biotypes, to apply a BFSV validation process. That is, it may be more difficult for BFSV to be applied toward constructs that are not understood nor validated as "real" constructs that necessarily stand from some "real" psychiatric phenomena.

One might argue that RDoC, in becoming a more adaptable framework that has relaxed its scientific realistic commitments and reductionistic assumptions, may better facilitate those outside research groups to figure things out on their own. This seems to be the general sentiment among former RDoC critics who consider RDoC as having moved away from its neurocentric approach and toward a focus in providing tools such as computational methods. They see this step as making RDoC more promising in its efforts to solve the validity problem. On the other hand, there is a counter-sentiment from those in clinical neuroscience who bought into RDoC 1.0's initial vision that RDoC 2.0 essentially is ineffective. The biology-first of RDoC 2.0, which now sets neuroscience more evenly with other higher levels of analysis is insufficient, and by relaxing its previous neurocentric stance and validation standards, it has distanced those who originally saw it as promising for this reason alone. Moreover, neuroscience-based research programs that are encouraged to develop their own validity standards rather than have them be provided for them might ultimately face trouble with being compatible with a more instrumentalist RDoC framework.

Lastly, it's not clear that if a research group draws on RDoC constructs to develop their own classification system, to what degree should we attribute their success to RDoC? For example, what if someone proposed something that's not biologically based—will it be an RDoC

project, or something else? Fortunately, RDoC is an adaptable and responsive research framework that can change and modify as it continues, so that we can expect some positive iterations by the RDoC Unit as more researchers attempt to utilize RDoC to develop biologically-based classification systems. At present, however, there remain some outstanding questions regarding the underlying philosophy of its approach and validation. But instead of reductionism, it's instrumentalism that may be an issue.

Conclusion

In this chapter, I have argued that RDoC employs a dual-track model of validation, with Biology-First Function Validity as a form of construct validity used to validate its RDoC constructs, and Biology-First Syndromal Validity as that which will contribute to a future sense of syndromal validation. By providing a historical overview of RDoC's change in research orientation from RDoC 1.0 to RDoC 2.0, I further clarify the shift in the conceptualization of its constructs, its relation to construct validity theory, and its changing philosophical positions. In doing so, I have demonstrated that RDoC has positioned itself as an alternative framework that integrates a variety of different senses of validity, while also being open to researchers utilizing its tools to bring about their own ideas and standards for validation.

Unlike other alternative frameworks, which have more deliberately selected a specific path by which to venture out of the box canyon, RDoC's proposed solution may be thought of as sending out a group of scouts at the base of the canyon to search for different paths to take. There isn't an assumption that there's one correct path—instead, there's an opportunity for multiple paths, in which different groups may report back on their respective paths and which may ultimately be promising. The question is, will a lack of explicit standards be facilitative enough,

or will the many paths separate the group too much so that they end up in completely different places, left wishing they had stuck together? Having reconstructed the fourth and final sense of validity within the Holy Quadrinity, we now turn to the final chapter to assess the implication of the four distinct senses of validity scientific psychiatry and how we may address the problem of disparate validation.

References for Chapter 5

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Berenbaum, H. (2013). Classification and psychopathology research. *Journal of abnormal psychology*, *122*(3), 894-901.

Borsboom, D., Cramer, A. O., & Kalis, A. (2019). Brain disorders? Not really: Why network structures block reductionism in psychopathology research. *Behavioral and brain sciences*, *42*, e2.

Casey, B. J., Craddock, N., Cuthbert, B. N., Hyman, S. E., Lee, F. S., & Ressler, K. J. (2013). *DSM-5* and RDoC: Progress in psychiatry research?. *Nature reviews neuroscience*, *14*(11), 810–814.

Carter, C., Barch, D. (2007). *Cognitive Neuroscience Treatment Research to Improve Cognition in Schizophrenia*. CNTRICS. https://cntrics.ucdavis.edu/

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.

Cohen, J. D., & Insel, T. R. (2008). Cognitive neuroscience and schizophrenia: Translational research in need of a translator. *Biological psychiatry*, *64*(1), 2–3.

Cronbach, L. J. (1971). Test validation. In R. Thorndike (Ed.), *Educational measurement* (2nd ed., p. 443). Washington DC: American Council on Education.

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC medicine*, *11*(1), 1–8.

Cuthbert, B. N. (2014). The RDoC framework: Facilitating transition from *ICD*/*DSM* to dimensional approaches that integrate neuroscience and psychopathology. *World psychiatry*, *13*(1), 28–35.

Cuthbert, B. N. (2020). The role of RDoC in future classification of mental disorders. *Dialogues in clinical neuroscience*, *22*(1), 81–85.

DeYoung, C. G., Blain, S. D., Latzman, R. D., Grazioplene, R., Haltigan, J. D., Kotov, R., ... & Tobin, K. E. (2023, May 5). The hierarchical taxonomy of psychopathology (HiTOP) and the search for neurobiological substrates of mental illness: A systematic review and roadmap for future research. https://doi.org/10.31234/osf.io/yatw7.

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.

Insel, T. R. (2014). The NIMH research domain criteria (RDoC) project: Precision medicine for psychiatry. *American journal of psychiatry*, *171*(4), 395–397.

Ivleva, E. I., Bidesi, A. S., Keshavan, M. S., Pearlson, G. D., Meda, S. A., Dodig, D., ... & Tamminga, C. A. (2013). Gray matter volume as an intermediate phenotype for psychosis: Bipolar-schizophrenia network on intermediate phenotypes (B-SNIP). *American journal of psychiatry*, *170*(11), 1285–1296.

Keshavan, M. S., Clementz, B. A., Pearlson, G. D., Sweeney, J. A., & Tamminga, C. A. (2013). Reimagining psychoses: An agnostic approach to diagnosis. *Schizophrenia research*, *146*(1–3), 10–16.

Lilienfeld, S. O., & Treadway, M. T. (2016). Clashing diagnostic approaches: *DSM-ICD* versus RDoC. *Annual review of clinical psychology*, *12*, 435–463.

Lobo, R. P., Bottenhorn, K. L., Riedel, M. C., Toma, A. I., Hare, M. M., Smith, D. D., ... & Laird, A. R. (2023). Neural systems underlying RDoC social constructs: An activation likelihood estimation meta-analysis. *Neuroscience & Biobehavioral Reviews*, *144*, 104971.

Lockard, M. (2013). Epistemic instrumentalism. *Synthese*, *190*(9), 1701–1718

Lovasz, N., & Slaney, K. L. (2013). What makes a hypothetical construct "hypothetical"? Tracing the origins and uses of the 'hypothetical construct'concept in psychological science. *New ideas in psychology*, *31*(1), 22–31.

Moran, M. (2015, October 7). *Insel to step down as director of NIMH*. Psychiatric Times. https://psychnews.psychiatryonline.org/doi/10.1176/appi.pn.2015.10b2

Morris, S. E., Sanislow, C. A., Pacheco, J., Vaidyanathan, U., Gordon, J. A., & Cuthbert, B. N. (2022). Revisiting the seven pillars of RDoC. *BMC medicine*, *20*(1), 220.

National Advisory Mental Health Council Workgroup on Tasks and Measures for Research Domain Criteria. (2016). Behavioral assessment methods for RDoC constructs. https://www.nimh.nih.gov/sites/default/files/documents/about/advisory-boards-and-groups/namhc/reports/RDoC_council_workgroup_report.pdf

National Advisory Mental Health Council Workgroup. (2018a, September 20). RDoC changes to the matrix (CMAT) workgroup update: Addition of the sensorimotor domain. https://www.nimh.nih.gov/sites/default/files/documents/about/advisory-boards-and-groups/namhc/reports/cmat-sensorimotordomainreport-508.pdf

National Advisory Mental Health Council Workgroup on Changes to the Research Domain Criteria Matrix. (2018b, May 17). RDoC changes to the matrix (CMAT) workgroup update: Proposed positive valence domain revisions.

https://www.nimh.nih.gov/sites/default/files/documents/about/advisory-boards-and-groups/namhc/reports/cmat-pvs-report-508.pdf

National Institute of Mental Health (2011, September 9). *Dimensional approaches to research classification in psychiatric disorders (R01)*. NIMH. https://grants.nih.gov/grants/guide/rfa-files/RFA-MH-12-100.html

National Institute of Mental Health. (2014). Development and definitions of the RDoC domains and constructs.

National Institute of Mental Health (2022a, September 28). *NIMH research domain criteria roundtable - Data-driven refinement of psychopathology: Toward precision diagnostics*. NIMH. https://www.nimh.nih.gov/news/events/2022/nimh-research-domain-criteria-roundtable-data-driven-refinement-of-psychopathology-toward-precision-diagnostics

National Institute of Mental Health (2022b, August 31). *Computational approaches for validating dimensional constructs of relevance to psychopathology (R01)*. NIMH. https://grants.nih.gov/grants/guide/pa-files/PAR-23-307.html

National Institute of Mental Health (2022c, August 31). *Computationally defined behaviors in clinical psychiatry (R21 clinical trial option)*. NIMH. https://grants.nih.gov/grants/guide/pa-files/PAR-23-305.html

Sarkar, S. (1992). Models of reduction and categories of reductionism. *Synthese*, *91*, 167–194.

Shankman, S. A., & Gorka, S. M. (2015). Psychopathology research in the RDoC era: Unanswered questions and the importance of the psychophysiological unit of analysis. *International journal of psychophysiology*, *98*(2), 330–337.

Tamminga, C. A. (2014). Approaching human neuroscience for disease understanding. *World Psychiatry*, *13*(1), 41.

Chapter 6: Psychiatry's Second Validity Crisis and Unrecognized Plurality

6.1 Introduction

Over the past forty years, scientific psychiatry has faced a crisis of confidence in the validity of its psychiatric classifications (Philips, 2013). In response, three alternative research frameworks, the Hierarchical Taxonomy of Psychopathology (HiTOP), the Network Approach to Psychopathology, and the Research Domain Criteria (RDoC) have emerged with the shared goal of studying and classifying psychiatric disorders in new ways. By approaching psychiatry's validity problem in ways that are unbound by the constraints of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*), the hope is this will bolster the validity of future psychiatric classifications, resolving psychiatry's longstanding validity problem. In addition, the *DSM* itself has recently been updated with a new continuous improvement revision model (APA, 2023) to address concerns regarding the validity of its diagnostic categories and contribute to a more evidence-based scientific nosology.

A yet unexplored aspect of psychiatry's "validity crisis" is related to disagreements regarding the *standards* of validity. Disagreements regarding standards of validity that amount to multiple distinct senses of validity point to a thornier methodological problem for psychiatry that I term "the problem of disparate validation." This two-part problem can be summarized as follows: scientific psychiatry aims at achieving empirically informed classifications that demonstrate validity in the sense that they correspond to real attributes or features of psychopathology. To achieve this, alternative research frameworks are now approaching the conceptualization, testing, organizing, and validation of features of psychopathology by their own standards in the hopes of one day informing more valid systems of psychiatric

classification. The first problem is, given a system of classification, by whose standard of validity should such a system be validated? Second, when we attempt to validate classifications informed by differing standards of validity, will any such validation process be capable of assessing a unified fundamental sense of validity that exists across the various frameworks, or will each approach only be valid in its own narrow sense?

In this final chapter, I offer an assessment of the problem of disparate validation through faithful reconstructions of what I term the Holy Quadrinity of distinct senses of validity in scientific psychiatry and psychological testing. By evaluating psychiatry's distinct senses of validity, I argue that despite the appearance of a shared goal of informing valid classifications, the existence of multiple frameworks in which each employs their own standards of validity is problematic methodologically speaking for trying to do any kind of unified validation work. At its core, differing standards that inspire fundamental disagreements concerning 1) the underlying phenomenon that researchers are attempting to make inferences about, 2) the sources of validating evidence, and 3) the very nature of validity and validation move each framework further and further into a state of non-complementary *unrecognized plurality*, being that we have yet to fully realize to what extent these frameworks are really not at all talking about the same thing when it comes to validity and as a result are presently engaged in very different projects with different aims. The consequence is a second validity crisis for psychiatry.

I conclude with a positive program that suggests in what ways such different frameworks with distinct validation procedures can achieve validity in their own specific sense while also coming to inform one another through a kind of complementary pluralism. I offer general recommendations as to how the frameworks may *stabilize* their own principles and procedures of validity and validation. Given the inability to establish a unified sense of validity, I advocate for

the development of convergent standards of *utility*, which may help facilitate comparing the usefulness within and between current and future systems of classification.

6.2 Psychiatry's Second Validity Crisis: It's Not for a Lack of Evidence

Before going into the different senses of validation, I would like to clarify what I mean by this notion of psychiatry's second validity crisis and distinguish it from the initial crisis of confidence in validity. As discussed in chapter 1, the initial crisis of confidence in validity, with validity understood to be the degree to which the *DSM*'s diagnostic categories (e.g., Major Depressive Disorder) may be said to stand for real underlying clinical syndromes, is such that after decades of continued research into the *DSM*'s diagnostic categories and despite significant advances in the psychological and brain sciences, the extraordinary lack of evidence for establishing validity suggests that our psychiatric classifications do not stand for real psychiatric disorders. With a central aim of psychiatry as aligned with mainstream medicine being to classify, diagnose, and treat real disorders as they actually exist, the worry that the *DSM*'s diagnostic categories are lacking in evidence to support their validity is of pressing concern.

Several explanations have been offered for psychiatry's longstanding validity problem. Some claim the hypothesis that psychiatric disorders are discrete syndromes is incorrect. Others say the *DSM* serves too many masters in its competing scientific, professional, and practical aims. A third explanation points not to a single issue but instead suggests there is just so much wrong with the *DSM*'s overall approach for developing valid classifications that psychiatry as a whole needs a do-over. In this regard, psychiatry's status in achieving validity for psychiatric disorders has been likened to being stuck in a "validity box canyon" (Kendler & Parnas, 2012). For some, the only way out is to retrace our steps back the way we came, at which point we may

forge new paths toward achieving validity by approaching the validity problem in the right way and provide improved opportunities to collect and evaluate validating evidence of a higher quantity and quality.

The problem is psychiatry's original crisis of confidence in validity isn't solely based on the lack or quality of validating evidence. Based on separate evaluations of the same set of validating evidence, some say the *DSM* maintains a sufficient degree of validity, while others claim the *DSM* has "0% validity" (Lynch, 2018, p. 5). These differences suggest that the initial crisis in validity is in part due to disagreements based in the interpretation and meaning of the validating evidence, i.e., disagreements concerning the *standards* of validity, with debates centered around what is validity or sufficient validity and what counts as validating evidence. That there exist such disagreements in this regard is problematic for psychiatry, but so long as they persist within a single validation framework (e.g., the *DSM*'s method of achieving diagnostic validity), the problem may be relatively contained and successfully addressed.

A second validity crisis arises once psychiatry introduces one or more additional approaches. Each approach is currently situated so that their distinct standards have them working on the validity *problem within their own framing of what counts as validity,* compounding the original problem by producing results whereby what is deemed valid for one approach may be invalid for another. This notion must be reiterated, for until now it was assumed that while these approaches knowingly adopt different governing conceptions of mental disorders and distinct methods for researching different aspects of psychopathology that amount to widespread *methodological* pluralism, progress in achieving validity is attainable through a plurality of approaches because all the approaches share the same scientific aim—develop valid psychiatric classifications. However, scientific psychiatry has yet to recognize the implications

of differing standards and what this means for establishing validity since overall, validity and validation within each framework have remained underspecified and at times inconsistently defined. As a result, the approaches do not share the same aims since when it comes to the goal of achieving validity, they end up not discussing the same issue. Without addressing these disagreements in standards of validity and their implications, psychiatry will continue to remain at the bottom of the validity box canyon. To get a clearer sense of how disparate these approaches' senses of validity really are and the implications of their divisions, I now introduce the four main approaches in psychiatry and an overview of what I interpret to be their distinct senses of validity, followed by an analysis of the commonalities between approaches as well as their differences.

6.3 The Holy Quadrinity of Validity in Psychiatry: An Overview

The Holy Quadrinity plays off the so-called trinitarian doctrine of criterion validity, content validity, and construct validity, and instead represents four distinct senses of validity of each primary psychiatric research framework. They include Diagnostic Validity (*DSM*), Structure-First Psychometric Validity (HiTOP), Network Psychometric Validity (the Network Approach), and Etio-Pathophysiological Validity (RDoC).

Diagnostic Validity

Validity in the *DSM*, referred to as diagnostic validity, is the specific sense of validation first articulated by Eli Robins (1921–1994) and Samuel Guze (1925–1996) in 1970 and subsequently updated alongside the various iterations of the *DSM*'s categorical classification system in the context of clinical medicine. Robins and Guze's method was motivated by an

interest in achieving psychiatrist Emil Kraepelin's (1856–1926) big idea: that psychiatric disorders may be shown to be discrete disease entities that can be accurately identified through clinical observation of their signs and symptoms, direct observation of their pathological anatomy and underlying physiology, or through the study of their etiology. Their overall strategy was simple: 1) identify discrete and homogenous diagnostic groups based on what we know, being observable signs and symptoms, and 2) further study those groups that may serve as steppingstones to discovering their underlying and corresponding pathological processes.

Diagnostic validity in its present form is the extent that a *DSM*-based *diagnostic category*, comprised of a set of operationalized *diagnostic criteria* intended to represent the observable signs and symptoms of an underlying clinical syndrome (Table 3, Appendix), is supported by a specific set of *validators*, understood as acceptable sources of validating evidence for a diagnostic category (Table 2, Appendix). The central question is "whether we have any confidence in the validity of this syndrome based on the set of validators" (Kendler, 2009, p. 8). Diagnostic validity is based in evaluations made by expert-led committees over twenty disorder groupings of "the overall strength of evidence across all validators" and an evaluation of the "strength of evidence for each of the validators" (p. 3).

A diagnostic category's set of diagnostic criteria are infallible indicators of the underlying clinical syndrome—i.e., they shouldn't be taken literally. The symptoms only index the disease, they don't constitute the underlying syndrome. This distinction between the diagnostic criteria (i.e., the symptoms) and the underlying clinical syndrome is important, as it supports what I take to be two overlapping yet distinct senses of diagnostic validity. The first sense, which I interpret as *small-V diagnostic validity*, relates to judgments regarding whether a diagnostic category can accurately identify, measure, or refer to a disorder based on its set of

diagnostic criteria. In this more instrumental or pragmatic sense of diagnostic validity, diagnostic criteria, although not a measurement instrument in the traditional sense, are treated as measurement instruments as such to predict the presence of a disorder. Essentially, small-V diagnostic validity amounts to judgments about how well a diagnostic criteria set is at identifying, differentiating, and predicting homogeneous diagnostic groupings. The aim of small-V diagnostic validity amounts to the diagnostic categories achieving predictive utility. The second sense of diagnostic validity, which I interpret as *big-V diagnostic validity*, relates to whether the diagnostic category stands for some real underlying clinical syndrome. In this more scientific realist sense of diagnostic validity, the aim is to provide new information about the *focal phenomena*, being the underlying clinical syndrome, that the diagnostic criteria actually represent to establish the existence of the psychiatric disorder.

Structure-First Psychometric Validity

HiTOP's specific sense of validity, which I interpret as *Structure-First Psychometric Validity* (SPFV), developed independently from diagnostic validity in the context of psychometrics, a quantitative scientific research tradition traced to the introduction of the common factor model of general intelligence introduced by Charles Spearman (1904) "that concerns itself with the study of measurement and human behavior" (Wijsen, 2021). HiTOP's psychometric approach rejects the *DSM*'s notion of a psychiatric disorder, being the result of underlying discrete clinical syndrome, and instead views psychopathology as arising from coherent and distinct latent (unobserved) dimensions as represented by its HiTOP constructs that hold a shared causal influence on a set of indicator (observed) variables, being the symptom groupings of psychopathology. HiTOP's data-driven strategy, described as evidence over

experts, uses a statistical technique called factor analysis to derive the most common factors of psychopathology, i.e., HiTOP *Spectra*, which it refers to as the structure of psychopathology, from psychological testing data. HiTOP then organizes the "empirically-derived" structure from broadest (e.g., the general factor of psychopathology) to narrowest (e.g., homogeneous symptom components/maladaptive traits) (Figure 2, Appendix). HiTOP ultimately seeks to validate the entire data-driven structure in the hopes of informing a future transdiagnostic classification system that may replace the *DSM*.

SPFV falls within *psychometric validity*, broadly understood as the degree to which a test, being a response to a standardized situation devised to measure a psychological construct has the desired psychometric features (e.g., various senses of validity, reliability, utility, etc.). Unlike diagnostic validity, SPFV is based in *construct validity*—the unified sense of validity that encompasses all psychometric validity within the context of psychological testing. SPFV is an application-specific psychometric sense of validity centered on supporting the development of psychiatric classifications that carry both *scientific accuracy*, understood as the degree to which HiTOP's constructs represent true features of psychopathology, and *clinical utility*, being the degree to which classifications are considered practically and pragmatically useful in the clinic.

Whereas the *DSM* understands utility as being one of several reasons for adding or subtracting specific criteria from a diagnostic category and thus views utility as being a part of a diagnostic category, HiTOP rejects this conceptualization. HiTOP instead maintains that utility should not be part of the definition of a HiTOP construct, and instead holds that utility follows from the correct definition, i.e., the scientific accuracy, of the construct. Thus, scientific accuracy is not only heavily prioritized by HiTOP but seemingly equated with the very concept of validity, so that "establishing the scientific accuracy of psychiatric classification systems is essentially the

task of establishing their validity" (Forbes et al., 2023, p. 12). Thus, for SVPF, validity can be thought of as the degree to which validating evidence supports the empirical quantitative measurement of a psychiatric classification to establish scientific accuracy.

SFPV is an ongoing three-stage validation process. The first and most important stage is the evaluation of structural validity (hence, structure-first), being the degree to which a particular construct accounts for the empirically observed *covariance* (i.e., the direct relationship) between different signs and symptoms of psychopathology. *Structural evidence* is based in exploratory and confirmatory factor-analytic research with a preference for continuous latent variable models. Such models produce factor loadings, i.e., standardized correlations between the original variables and an underlying factor (the construct), which are assessed based on the goodness of fit of the model to the data. The second stage is an evaluation of *external validity*, referring to the degree to which evidence for a certain HiTOP construct correlates with other (relevant) indicators of that construct. HiTOP distinguishes external validity into five distinct types (Table 6, Appendix). The third stage is an evaluation of *reliability*, the extent to which all items on a test measure the same construct, as well as clinical utility, and predictive utility, the degree to which a construct is helpful in differentially predicting outcomes of interest. Through SPFV, HiTOP aims to validate its constructs as well as the broader system of multiple, hierarchically organized constructs, i.e., the entire HiTOP hierarchical model itself.

3. Network Psychometric Validity

Network Psychometric Validity (NPV), also based in psychometric validity, operates within the network perspective, which represents a family of models and methodologies that draw on the network approach to psychopathology, network theory, complex systems theory, and

applications in network science to model and study mental disorders as "causal systems of mutually reinforcing symptoms" (Robinaugh et al., 2020, p. 353). For example, under the network approach, a "depressive episode" is hypothesized to arise from the causal interaction between symptoms such as depressed mood, anhedonia, and others (e.g., insomnia, fatigue). As a result, symptoms are not conceived as fallible indicators of some underlying common cause such as the *DSM*'s Major Depressive Disorder or HiTOP's Internalizing Spectra. Instead, it is the mutual interaction between symptoms that constitutes "depression" itself.

Unlike the *DSM* or HiTOP, the network approach to psychopathology, which first gained traction in 2008 from psychometrician Denny Borsboom (Borsboom, 2008) and affiliated members of a psychometrics research group based in the Netherlands (Cramer et al. 2010), is not affiliated with any single governing body or research organization, nor does it offer a system of psychiatric classification or organizing research framework. The overall strategy and aim of the network approach is to replace the data-driven latent variable models typical of the HiTOP approach, which the network approach claims do not support testable explanations of the phenomena, with theory-based network models of psychopathology. This means selecting network models based on theoretical reasons which, it is argued, provide the necessary rationale for developing and testing network theories.

A *network model* of mental disorders is a statistical model that represents features of a mental disorder as derived from a specific *network hypothesis,* being a testable and falsifiable hypothesis of how the components in a network influence each other over time. A *network* is a representation of the relationships (formally called edges) between constituent variables (formally called nodes) within a system (Figure 3, Appendix). In psychopathology networks, the nodes represent various constituent elements of psychopathology (e.g., symptoms, biomarkers,

cognitive processes) and are depicted by a circle. *Edges* represent conditional associations between nodes are depicted by a line or a one or two-sided arrow (depending on the directionality of the relationship). Two or more networks may be connected by what has been referred to as *bridge symptoms* (or bridge nodes). The *external field* is an area outside of the network whose components (e.g., "stress") may causally intervene on the nodes or edges inside of the network. In a network model, the more thorough the connections of the nodes in the network, the more likely the network is to remain in a dysfunctional state even after removal of the original biopsychosocial variables (i.e., hysteresis), and may also reflect higher levels of severity in dysfunction (van de Leemput et al., 2014).

NPV can be summarized as being comprised of three stages to support the development and testing of network theories. The first stage is the validation of the individual components in the network, based in the concept of *node validity*, a model-specific validation process (Bringmann et al., 2022). Node validity is a two-step validation procedure that involves 1) *node selection,* referring to the adequacy of selecting appropriate variables as nodes in a network model, and 2) *node assessment*, referring to the quality of the operationalizations used for selected variables. The second stage is the validation of the dynamical relations between the components, i.e., the validation of the *network structure*, whereby the dynamic relation between specific nodes may be understood as a kind of useful construct to be validated. The third and final stage is the empirical testing of network hypotheses via confirmatory techniques based in standards of confirmatory research (e.g., Wagenmakers et al., 2012) and *theory construction methodology* (e.g., Borsboom et al., 2021), which, when successful, implicitly validate the theory-derived network model and network theory.

1. Etio-Pathophysiological validity

Etio-pathophysiological validity is the particular sense of validity of RDoC, an alternative psychiatric research framework of the National Institute of Mental Health (NIMH). Following its initial launch in 2009, in 2015 the RDoC program changed leadership following Thomas Insel's departure as director of the NIMH, resulting in a significant change in research priority and orientation—an RDoC 2.0—that would eventually lead to the retirement of the original RDoC matrix. RDoC in its present form represents an integrative approach to validation in that it attempts to bring the syndrome model, psychometrics, dimensionality, and multicausality together with cognitive neuroscience in one unifying validity framework. RDoC's overall strategy is to group patients for clinical studies based on fundamental dimensions of behavior and neurobiological measures (genes, circuits, etc.). By adopting an approach that breaks up the *DSM*'s diagnostic categories like HiTOP and the network approach, yet unlike those approaches fixed measuring the underlying biology of the phenomena as its foundation and right starting point, RDoC aims to inform a more valid classification system in the future that is based firmly in biology, behavior, and context.

The RDoC approach employs what I interpret as a dual-track model of validation. The first track, which I term *biology-first function validity* (BFFV), is centered on the validation of the tools of the RDoC Framework, referred to by RDoC as the concepts for investigation, with a primary focus on RDoC constructs (Figure 4, Appendix). These concepts include six major functional *Domains*, which represent the current understanding of the major systems of cognition, motivation, and social behavior, i.e., those systems which, when there is dysregulation and dysfunction within/across them, are thought to give rise to psychological and behavioral impairments. Each domain is accompanied by three to six *Construct*s, i.e., concepts summarizing

data about a specified psychological/biological dimension of behavior, and recently defined as empirical functions (Table 3, Appendix). *Units of analysis* are the methods and instruments used to study the constructs from a normal to abnormal range of functioning.

The second track, which is intended to be informed and supported by the first track, is the at-present open-ended and under-specified validation of a future diagnostic classification system, which I term *biology-first syndromal validity* (BFSV). BFSV would be expected to be applied to a classification system that can accommodate RDoC's specific hypothesis concerning mental disorders as representing broad and biologically heterogeneous syndromes as opposed to discrete clinical syndromes. BFSV would require validating such a classification system using etiology, pathophysiology, prognosis, and treatment response measures.

The term biology-first applies to both tracks of validation and is in service of RDoC's primary aims of 1) developing an etiological and pathophysiological understanding of human systems of normal and abnormal functioning and 2) contributing to a future biologically based system of classification. In the first track, biology-first is evidenced by the specific criteria for RDoC constructs. For an RDoC construct to be initially selected and subsequently considered valid, it must include, among other requirements, evidence that a neural circuit or some biologically based system plays a role in implementing the function. The second track, biology-first reflects the notion that a future classification system to which RDoC will contribute will be validating syndromes that have been shaped via an understanding of their biological basis. A leading frontrunner is the new classificatory concept of *biotypes*, being transdiagnostic clusters defined by responses on measures across units of analyses that are "more biologically valid groupings than the diagnostic categories" (Cuthbert, 2020, p. 84). While biotypes may be a leading model on which to base a future classification system, RDoC's most basic goal remains

to encourage investigators to think about diagnoses in new ways—dimensional, categorical, or otherwise—and as such, is not (yet) committed to one conception of a future classification, nor an explicit validation process for syndromal validity (Table 7, Appendix).

6.4 Similarities Across the Disparate Senses of Validity

Despite the four approaches each having developed distinct senses of validity from their differing standards of validity, there exists several commonalities across them that, before having faithfully reconstructed each framework's sense of validity, have gone almost entirely unnoticed. The three most significant commonalities between the approaches are discussed below.

A Return to the Original Validators of Robins and Guze (1970)

One of the most surprising commonalities learned upon analyzing the three alternative frameworks and the *DSM*'s approach is that at some point in the validation process for those frameworks seeking to inform a future psychiatric classification system, i.e., HiTOP and RDoC, there is a return to some of the original validators of Robins and Guze (1970) that were adopted and modified within the *DSM*'s application-specific method of establishing diagnostic validity. The original validators include sources of validating evidence such as etiology (i.e., family history), prognosis (i.e., the likely course or outcome of the illness), and treatment response. This is unexpected for two reasons. First, the primary motivation of the alternative frameworks was to go beyond the validating methods of the *DSM*, and thus to do something fundamentally different. Second, both HiTOP and RDoC have been so outwardly critical of the *DSM*'s standards for establishing validity of the diagnostic categories that a return to what they deem problematic seems counterintuitive to their aims.

To clarify, HiTOP and RDoC do explore other validity avenues first before returning to a portion of the original validators to evaluate things like prognosis, biomarkers, and etiology of their newly formed constructs. For the *DSM*, the strategy has been to begin with the clinical description of the signs and symptoms of psychopathology, being the initial original validating class, then proceed to the other validators. HiTOP first develops structural evidence of the HiTOP spectra, and then proceeds to validate HiTOP constructs against some of these validators. They are utilized to strengthen an overall claim to the validity of a construct if there is validating evidence, or if there isn't, indicate that a return to the structural evidence is needed. For RDoC, following a focus on biology-first functional validity for establishing construct validity of RDoC constructs, validation of a future RDoC-informed classification system via biology-first syndromal validity depends on how successful the biologically-based classifications predict prognosis and treatment response.

Expert Curation

A second unpredicted commonality is that all the alternative frameworks employ expert curation, meaning that decisions as to what is ultimately included in their model(s) or classification system(s) are based on compromises between experts. These experts, who are key opinion leaders and most often hold positions within an evaluation or oversight committee, are tasked with assessing and judging the validating evidence in relation to other various epistemic and non-epistemic aims of the approach, e.g., compromising between the truth of the classification and the degree it is considered to be useful in clinical practice. This is also surprising for the fact that one of the primary criticisms of the *DSM*'s overall approach is that it is too expert-based and is thus overtly biased and subjective or beholden to outside interests.

When alternative frameworks are promoted, the perception they all put forward is that their "data-driven," "theory-driven," or "biologically driven" approach removes the overreliance on experts and permits their own version of an empirically-based process that "determines" or "discovers" its classifications in an "objective" and "scientific" manner without the use of experts.

To their credit, all the frameworks, including the *DSM*, appear to try to either limit the reliance on experts or make the expert-led decisions more "scientific" or "objective," although what each considers to be scientific or objective differs depending on the approach. The *DSM* still utilizes committees to evaluate and judge various sources of evidence for revising its classification system and is arguably still the most expert-driven approach, but it has over the years sought to systematize and standardize this process in a way that permits the quality and strength of the empirical evidence to have more weight in the revision process. HiTOP distinguishes various features in its overarching model, such as the separation between Somatic and Internalizing Spectra, or that of the higher-order general factor of psychology or p-factor based on expert decisions as opposed to those features being empirically determined solely by factor analysis. HiTOP combines different analyses of factor analytic studies into a single model, meaning that it's the experts, not the evidence, who select what goes into HiTOP. At the same time, HiTOP implements the evidence-based GRADE rating system in their revision decisions to make the use of experts more objective. The Network Approach utilizes a partially expert-led process of node selection and node assessment, whereby the "clinical or theoretical hypothesis of a clinician or clinical researcher, often formulated together with the patient, plays a role in the choice of the set of variables or nodes in the network" (Bringmann et al., 2022, p. 3). In turn, they maintain that rigorous empirical testing is the ultimate arbiter. RDoC developed the original

RDoC Matrix during a series of two-day workshops through a process of good enough expert consensus, whereby experts hurriedly populated RDoC constructs into the Matrix that they took to be the most reasonable. RDoC still adopts specific tasks and experimental paradigms that experts deem important, but their strategy has been to consider the constructs in the now updated RDoC Framework as more exemplars to be tested, thus reducing the strength of their claims regarding their expert-selected constructs.

Validity is Broadly Understood as That Which Is Considered "Good" or "Desirable"

Despite the elaborate presentation of validity being empirically or scientifically based, accompanied by long lists of validators or intricate sequencing of how such and such evidence should be evaluated, a third commonality is validity for each approach in the broadest of sense boils down to mean that which is considered to be "good" or "desirable" for that approach. That validity accounts in psychiatry may be interpreted so broadly is not inherently a criticism. In the 2010s, the "great validity debate" within educational and psychological testing centered specifically on how and whether to define validity in a narrow or very broad sense (Newton & Baird, 2016). The question was whether validity should be defined in the narrow, traditional sense of determining whether a test is actually measuring the thing you want to measure (Markus, 2016)—which is far more difficult to establish—versus a broad sense so that validity is more flexible and may come to mean anything to do with whether an assessment procedure is "good" or "bad" (Newton & Shaw, 2016). How to define validity was essentially a debate concerning how the term validity should be used, while the more difficult philosophical questions as to whether validity should establish a link between the construct one is attempting to

measure and the measuring instrument, or whether validity should be thought of as demonstrating that the thing being measured exists, did not drive the debates.

Evidence to suggest validity may at times be understood more broadly in psychiatry is the fact that no current account of validation in any of the four approaches engages in the difficult measurement questions in terms of what we are *really* doing when we're measuring and subsequently validating, or what is really required of psychological measurement so that we may be confident that the thing underlying the construct (e.g., a clinical syndrome, the general factor of psychopathology) exists and thus may be considered valid. There is a heightened focus on conceptualizing the phenomena, thinking about how to study it, deciding which types of evidence are important, and developing and testing specific hypotheses, but like what has been previously observed by Borsboom (2005) in psychological measurement, scientific psychiatry has seldom if at all engaged with the more difficult and rigorous problems concerning measurement and validation.

6.5 Differences in Validity and Validation Across Frameworks

Despite some unanticipated commonalities among the approaches, findings related to the differences in validity between frameworks are even more surprising. These differences result from specific validity standards that inspire fundamental disagreements concerning 1) the phenomenon that researchers are attempting to make inferences about; 2) standards of validating evidence; and 3) the nature of validity and validation.

What Is Being Studied? The Nature of the Phenomenon

As each framework sets out to improve upon the validity of psychiatric disorders, each adopts a different conception as to the nature of the phenomenon being studied as described in Section 6.3. Maintaining different conceptions of the phenomenon to be studied is common in psychology and science in general. That there are different ideas as to what is the fundamental nature of psychiatric disorders isn't the problem. The issue is that each approach finds the governing conception of the underlying phenomenon of competing approaches to be incapable of being validated. HiTOP views the *DSM*'s diagnostic categories which are meant to stand for some real underlying clinical syndrome as having no basis in reality and which do much more harm than good. The Network Approach takes both the *DSM* and HiTOP's common cause approach, and in particular, HiTOP's assumption that HiTOP's higher order dimensions such as the p-factor as broad underlying causes that exist in the brain, as getting it flat wrong in contrast to their conception of psychopathology arising from the interaction between elements of a system. RDoC sees all the approaches as not taking the biological reality of function and dysfunction seriously, maintaining that any account of disorder that has not been built bottom-up beginning with the underlying neurobiological mechanisms is completely devoid of being valid. Lastly, the *DSM* sees the attempts of a paradigm shift by the alternative approaches in terms of how the underlying phenomenon goes against, and in some cases are unwarranted, by what they take to be promising evidence of their conception of psychiatric disorders they are currently iterating across.

What Counts as a Source of Evidence? The Nature of the Validating Evidence

Each framework has taken time to publish specific criteria as to what they deem to be their own sources of validating evidence. Such criteria include relevant validity concepts and

terms, a list of acceptable sources of validating evidence, how such evidence is to be evaluated, interpreted, weighted, and/or integrated, the specific process or procedure for assessing or establishing validity, how validity should be related to the concepts of reliability and utility, and how judgments of validity inform curation decisions by experts. Implicitly, such standards of validity for each approach tend to align with what each takes to be scientific, and have also been informed of that which is to be evaluated, being the underlying phenomenon and/or category, dimension, or construct meant to stand for the phenomenon.

Each approach maintains very distinct standards, in part due to how they all differently conceive the nature of the phenomena and how they expect it be validated. As a result, they all find one another's sources of validating evidence to be unacceptable. The *DSM* maintains its own collection of validators to be the best path forward for validating underlying clinical syndromes, and view other approaches as either underspecified or doing something entirely different. HiTOP views the *DSM*'s validators to be insufficient, especially in relation to its non-reliance on psychometric data and structural validity on which HiTOP constructs are based. The Network Approach views HiTOP as confusing the output of the data models used to discover its constructs with theory, and insists that HiTOP is not "discovering" but constructing its constructs. The Network Approach instead insists upon letting theory inform its models and emphasizes standards for testability and falsifiability of network hypotheses. RDoC would say all three approaches fail to emphasize neurobiological sources of validating evidence sufficiently to validate that mental disorders are brain-based disorders. Given distinct standards of validity, the implication is that what counts as a source of validating evidence for one simply won't count for the other.

What is Validity? The Nature of Validity and Validation

The nature of validity and validation can be understood as what each approach takes the notion of validity and the process of validation to amount to fundamentally. While one could interpret these aspects as being a part of the standards of validity, I interpret these differences to refer to some more significant underlying conceptualization, philosophy, or theory for validity. Once you invoke specific differences in the fundamental nature of validity, it is at this stage that we really start to see just how different these senses of validity in psychiatry truly are. The most fundamental difference among these approaches has to do with how each views validity, how each's conception of validity interacts with and relates or relates to construct validity theory (CVT), and how each maintains distinct engagements with scientific realism.

For the *DSM*, small-V diagnostic validity is, at its core, evaluations of predictive power and clinical utility of its diagnostic categories, whereas big-V diagnostic validity are judgments concerning the ability of a diagnostic category to stand for, i.e., represent the focal phenomena, the underlying clinical syndrome. For HiTOP, validity is only one thing, the scientific accuracy of its constructs, which is overwhelming based in structural validity and thus construct validity. For the Network Approach, validity is achieved implicitly when a network theory meets the standards for theory construction and testing. RDoC's biology-first function validity views establishing linkages between constructs and some biological or cognitive processes at an underlying biological level as what establishes a valid construct, while its biology-first syndromal validity is the degree to which a future RDoC-informed classification system successfully stands for broad and biologically-based heterogeneous syndromes.

In terms of their relation to CVT, the *DSM*'s diagnostic validity, while not based in CVT, maintains some general overlap with surface-level features, in that both view validity as a matter

of degree and as a unitary concept that is supported by various sources of validity evidence. HiTOP, being based in psychometric validity and construct validity, at times appears as a mishmash of disparate senses of CVT—a practice also common in experimental psychology as noted by Slaney (2017), who argues such differences may imply "a multiplicity of ontological and epistemological stances, some of which are incompatible, which itself carries nontrivial implications for both theory and practice" (p. 6). The network approach also draws on contemporary CVT with node validity, but only just enough to support the conditions appropriate for the testing of network hypotheses. With RDoC, specific interpretations of convergent and divergent forms of construct validity are invoked that differ in emphasis and application from both HiTOP and the Network Approach, with RDoC seeking to validate their RDoC constructs both as tools and that which may stand for some underlying biological or cognitive function.

When it comes to the underlying philosophical differences, while each approach maintains scientific realist aspirations in terms of establishing their constructs or future classifications as standing for that which really exists, the underlying philosophy of their distinct senses of validity may include mixed or even contrasting philosophical positions. For example, the *DSM*'s diagnostic validity features a mix of operationalism, by which concepts are stipulated in terms of their operations to establish their existence for small-V diagnostic validity, and scientific realism for big-V diagnostic validity. HiTOP's SPFV may (unintentionally) rely on what others have characterized as a positivist characterization of CVT, whereby hypothetical constructs are conceived as being without reference or meaning and at best useful fictions (something the HiTOP framework would not agree with), while such appeals may also be interpreted as a "methodological move" for permitting the testing of hypotheses based in a form of scientific realism. The network approach's first two stages of NPV, based in part on

contemporary CVT, maintain what is referred to as a constructivist-realist view associated with validity theorist Samuel Messick (1989) whereby validity is not dependent on the reality of the construct but is instead a property of the inferences made. In contrast, the third stage of network theory testing within NPV maintains a realist interpretation of psychological attributes. Lastly, RDoC's BFFV also maintains constructivist-realist stance in relation to the validation of RDoC constructs, whereas BFSV holds scientific realist underpinnings.

6.6 Unrecognized Plurality: Implications of Psychiatry's Second Validity Crisis

From examining the key differences, we begin to see how when one framework conceptualizes that which is to be validated in one way and thus pursues a specific validation procedure, another framework may reject this in part or entirely, thinking of the underlying phenomenon in a different way and thus going about on their own validation path. Similar concepts and terms such as validity, validation, validating evidence, reliability, utility, and construct, are used in service of seemingly similar aims of establishing validity or informing a valid classification system, but only presently within their own framing and thus only contributing to their own distinct sense of validation. The current situation can be summarized such that each approach rejects the very idea of a *complementary pluralism*, whereby multiple approaches may relate to one another in different ways and jointly contribute toward achieving validity, because the attitude of each approach is that theirs is the best and only way to do it.

Where we may start to see the implications of this situation is when we consider the space of potential solutions for scientific psychiatry in establishing valid psychiatric classifications. Embracing and encouraging the proliferation of alternative approaches distinct

from the *DSM* is based on at least the possibility of a complementary pluralism and assumes one of the following five general scenarios will result:

a) All of the alternative approaches fail in establishing validity, at which point we'll be no worse or ever so slightly worse off than we were before.

b) One of the alternative approaches succeeds in establishing validity so that psychiatry abandons the *DSM* and adopts a single new approach.

c) Two or more approaches succeed in establishing validity. The two or more successful approaches are evaluated, and psychiatry adopts a single approach for practical, convenient, and/or conventional reasons.

d) Two or more approaches partially succeed in establishing validity. Psychiatry coordinates and integrates between the approaches, permitting a plurality of approaches that contribute to one another's establishment of validity.

e) Two or more approaches partially succeed in establishing validity. Psychiatry coordinates and integrates between the approaches, contributing to a single, unified sense of validity.

Unfortunately, the approaches are currently in what I take to be a state of non-complementary *unrecognized plurality*, meaning that while each approach fully understands the other frameworks to be doing psychiatric research and classification in distinct ways, scientific psychiatry has yet to fully recognize the implications in the form of specific methodological difficulties resulting from uncompromising standards that amount to distinct senses of validity. Consequently, none of the above scenarios may at present be achieved due to three key

methodological difficulties: difficulties in evaluating between frameworks, difficulties integrating and coordinating between frameworks, and difficulties in establishing a unified sense of validity.

Difficulties Evaluating Between Frameworks

One of the most immediate implications of disagreements regarding standards of validity that amount to multiple distinct senses of validity is the difficulty in comparing frameworks in terms of which is more valid. Consider a recent debate between critics of the HiTOP model and HiTOP proponents in the journal *Clinical Psychological Science*. In their paper, "Folk Classification and Factor Rotations: Whales, Sharks, and the Problems with the Hierarchical Taxonomy of Psychopathology (HiTOP)," psychologists Haeffel et al. (2022a) challenged the notion that the HiTOP approach significantly improves upon traditional taxonomies such as the *DSM*. In critiquing HiTOP, they point to what they judge to be two major weaknesses that limit its potential toward achieving validity and thus scientific progress, in that its "data-driven" approach is atheoretical and, as a result, is unfalsifiable, meaning it is not suitable for theory-building. They use an example of the HiTOP approach being applied to a biological classification system of non-human animals and suggest that HiTOP's data-driven approach when used in this way would incorrectly classify whales and sharks together based on the statistical tendencies among their shared features. They further criticize HiTOP's simple-structure factor-analytic approach and its use of the degree of model fit as an indicator of validity (i.e., structural validity), claiming that "…HiTOP's hierarchical approach is not valid" (p. 262) nor does it hold the potential for achieving validity. They conclude by suggesting that both the *DSM* and RDoC have greater potential in this regard.

In their response, "Answering Questions about the Hierarchical Taxonomy of Psychopathology (HiTOP): Analogies to Whales and Sharks Miss the Boat," HiTOP proponents DeYoung et al. (2022) clarify HiTOP's validation procedure, arguing that its atheoretical, data-driven approach "maximizes coherence of constructs and distinctiveness between them" (p. 280) and is thus beneficial and valid, while also suggesting their approach is capable of hypothesis testing "according to their fit to the data" (p. 281). In response to the response, "The Hierarchical Taxonomy of Psychopathology (HiTOP) is Not an Improvement Over the *DSM*," Haeffel et al. (2022B) assert that DeYoung et al. (2022) fail to meet their initial criticisms, affirming that "decisions to change or replace a classification system should be based on the results of scientific competition (e.g., tests of incremental) validity" (p. 288) which they judge HiTOP to be incapable of producing.

Setting aside the fact that neither of these groups of authors takes the time to define nor provide an account of validity and validation, this version of a back and forth in debates concerning the validity (or in this case, the potential for achieving validity) among psychiatric frameworks is incredibly common and formulaic. Approach X will say of Approach Y, that "Because Approach Y does not lend itself to a, nor can it account for b, c, or d, this approach is ultimately less valid than Approach X." The approach in question responds by saying, "Actually, Approach Y does lend itself to a, just in this other more empirically valid way, and it can account for b, c, and d, unlike Approach X, which cannot account for a through d." Then the original critics from Approach X will respond, "no—you haven't responded sufficiently to our concerns, and are failing to see how Approach Y is fundamentally flawed and thus cannot be successful in achieving validity." All the while, a proponent from a third approach, Approach Z, who doesn't so much care for Approach X but who really doesn't like Approach Y is emailing their like-

minded colleagues, saying, “have you read the paper critiquing Approach Y? It’s so good, you must read it!”

The above deconstruction is meant to illustrate that an evaluation of either framework’s validity or potential for achieving validity in the future is compromised by the fact that all parties are operating on entirely different conceptions and standards for what counts as validity. Now, one might object and say neither approach agrees because all have failed to provide charitable interpretations of one another’s frameworks aside from their own. If they don’t create strawmen, there is potential to objectively evaluate and make a judgment between approaches as to which is the most valid. To this, I ask, according to whose conception and standards of validity? Given these debates come down to that which is good or desirable for their approach, how can we presently adjudicate between approaches and conclude as Haeffel et al. (2022) do that one approach is necessarily less valid than the other?

Difficulties Coordinating and Integrating Between Frameworks

A second implication is the prevention of coordinating and integrating differing approaches that may contribute to one another’s path to achieving validity. Despite each approach maintaining that their own framework is the best, there have nevertheless been several recent proposals to suggest how certain approaches may be successfully linked so that some aspects of one framework may be combined or informative to another framework. For example, the well-cited article by Michelini et al. (2021) “Linking RDoC and HiTOP: A New Interface for Advancing Psychiatric Nosology and Neuroscience” suggests that since RDoC and HiTOP both advocate for a dimensional understanding of psychopathology but maintain different approaches, they may still be informative to one another (Figure 5, Appendix). RDoC, with its biologically-

based focus, may provide helpful tools for "elucidating the underpinnings of the clinical problems in HiTOP" (p. 4), whereas HiTOP may motivate RDoC studies "by providing psychometrically robust clinical targets" (p. 4).

Any kind of coordinative and integrative interface between HiTOP and RDoC assumes the potential for achieving *construct stabilization*. Construct stabilization as defined by Sullivan (2016a) is the process by which there is active coordination across researchers situated in the same and different areas of science to come to agreements regarding (1) how to generally define terms designating constructs, (2) the best experimental paradigms for studying a given construct, and (3) the conditions under which two experimental paradigms can be said to measure the same construct. Coordination and integration between frameworks are thought to be contingent on the ability to stabilize constructs. A recommendation by Sullivan (2016b) specifically for RDoC is to organize researchers from disparate research programs to come together to discuss "what the relevant constructs are, how to investigate them, how to stabilize them, and related issues" (p. 313).

The problem is that given the significant disagreements between approaches that amount to distinct senses of validity, conditions 2 and 3 of construct stabilization at present cannot be met. Two cannot be met because both HiTOP (structure-first) and RDoC (biology-first) maintain that their own approach informs the best experimental paradigms and, given these approaches are so disparate, they ultimately will not coalesce on specific paradigms. Three cannot be met since differing standards for what is required for validation dictate diverging standards as to acceptable ways to measure a HiTOP or RDoC construct.

Perhaps the best argument against coordination and integration between HiTOP and RDoC is that RDoC and HiTOP constructs are not the same thing and do not even refer to the

same thing. HiTOP constructs are understood as *pure* constructs, i.e., data-driven factors based on statistical output from factor analysis of psychological testing data which are thought to refer to broad underlying dimensional factors of psychopathology (e.g., internalizing spectra). For RDoC, constructs are empirical functions selected via a process of expert curation based on whether there is "solid evidence [for the constructs] to serve as a platform for ongoing research" using biologically-driven experimental paradigms such as specific animal models (Morris et al., 2022, p. 6), and are thought to refer to specific biological and cognitive processes (e.g., Reward Learning). Thus, even if researchers from either approach were to come together to discuss what they mean by "construct," without dramatically overhauling their entire approach, construct stabilization isn't currently achievable. If these approaches cannot stabilize on what a construct is or the nature of what it is thought to refer to, then they will face methodological difficulties to be coordinated or integrated.

Difficulties in Establishing a Unified Sense of Validity

A third implication is that the approaches will be unable to establish validity in any unified sense. The difficulty begins with how each framework respectively and distinctly conceptualizes and represents the phenomena it wishes to validate. Take the broad concept of depression for which researchers wish to develop valid psychiatric classifications. The *DSM* maintains depression is a clinical syndrome that underlies the diagnostic category that is Major Depressive Disorder. HiTOP understands depression not as a clinical syndrome, but as existing across multiple transdiagnostic dimensions. For example, depression for HiTOP may amount to existing across dimensions of positive emotionality, negative emotionality, and physiological arousal. The very idea of turning depression into such a simplified category in which a patient

has it or doesn't in HiTOP goes against what depression actually is. In contrast to both the *DSM* and HiTOP, the Network Approach to Psychopathology thinks depression is not a result of some underlying common cause, but is more like a causal system of mutually reinforcing symptoms. The symptoms interact and self-sustain the disordered state, meaning depression is a system, not a syndrome. Lastly, RDoC may view depression as a collection of biotypes that draw on both structural and functional measures of biomarkers to drive the grouping of individuals with depression into neurobiologically distinctive and biologically meaningful clusters.

Unfortunately, the presence of disparate conceptions of phenomena amounts to a troubling scenario for trying to do any kind of unified validation. The central issue is that the approaches don't agree on the phenomenon that they're trying to make inferences about. If they don't agree on what the features or attributes are, this makes progress in achieving a unified sense of validity for psychiatry impossible. Those who view the phenomenon in one way will not just want but need to pursue one sense of validation, with specific standards and underlying conceptions of what amounts to validity, whereas another approach, finding the original approach fundamentally misguided, will reject it completely and insist on pursuing it in a different manner. How each approach conceptualizes and represents the phenomenon of interest necessitates how its researchers will study, interpret, and make inferences concerning validating evidence, ultimately informing their validation standards and shaping any sense of validity that is achieved.

6.7 Resolving Psychiatry's Second Validity Crisis

Initially, there may have been some hope that a plurality of approaches would lead us closer to achieving a unified sense of validity in psychiatry. But to borrow an old metaphor

offered by validity theorist Gregory Cizek, you can try to mix oil and water, but ultimately, you don't have a solution. You can shake them up, and they'll appear at times to *sort of* come together. But upon closer examination, what you have is not truly a solution, but multiple sources that end up separating because they cannot be combined.

This is the case with the Holy Quadrinity of validity in psychiatry and the problem of disparate validation. The way each approach understands validity simply cannot be synthesized with the way another approach understands validity. Each approach may only pursue its own path to validation. Perhaps it is only a matter of time, and evidence will suggest which of those paths is likely to produce the most progress in achieving validation in the most fundamental sense, i.e., in the development of psychiatric classifications that stand for real collections of attributes or features of psychopathology. But given each approach takes its own path to adopting its own sense of validity, the best approach is not something at present that evidence alone can tell us. We are left with the realization that there does not exist a single validation approach, or a combination thereof, nor will any single approach be capable of achieving a unified sense of validity across frameworks. The best psychiatry can hope for is that each approach develops valid classifications in its own specific sense.

To move the field forward, I offer a few recommendations. First, we should abandon hope to develop a unified sense of validity across disparate approaches. Scientific psychiatry instead should embrace and permit a plurality of distinct approaches with a plurality of non-complementary distinct sense of validity. This is separate from current attitudes which suggest either these disparate approaches may eventually work together through a division of labor toward validating psychiatric phenomena in a single sense, or that the correct approach will present itself based on specific findings related to some unified standard of validating evidence.

Second, despite the inability of the approaches to contribute to a unified sense of validity, the approaches are not so incompatible that they could not eventually move toward a complementary pluralism by which they inform and/or relate to one another. We can think of distinct senses of validity as if they are different languages. While languages differ in many ways, people can still be taught and learn to speak more than one. To this end, I recommend that each framework stabilize its own validation principles and procedures, which at present remain underspecified and at times inconsistent. Stabilizing should go beyond what Sullivan (2016a) recommends as defining the terms, the best experimental paradigms, and the conditions for measurement. It should also consider such aspects as stabilizing the scientific aims of the validation approach, standards for assuring consistency in the application of validation procedures, epistemic and non-epistemic values that may inform the decision process in drawing inferences from and making decisions based on the available validating evidence, and a willingness on the part of each approach to engage with more fundamental questions of measurement. To this last point, Maul (2017) recommends the ways validation methods in psychiatry may more readily be demonstrated to connect with the constructs they are trying to measure.

Third, I recommend developing shared standards of utility, understood broadly as the degree to which a psychiatric classification may predict course, outcome, and likely treatment response (Jablensky, 2016). Specifically, I recommend a moderate convergentist approach, which implies that we may reduce utility down to a shared sense that focuses on shared similarities of conceptions of utility between frameworks. Since most of the frameworks return to the original validators of Robins and Guze, which Solomon (2022) has characterized as *utilitators* for their ability to serve double-duty as sources of evidence to establish both validity

and utility, I suggest them as a valuable starting point for shared standards of utility. A convergent set of standards may help facilitate comparing the usefulness within and between current and future systems of classification.

By focusing on the development of shared standards of utility, I recommend that scientific psychiatry still pursue valid psychiatric classifications in terms of its overarching goal of discovering real psychiatric phenomena, but that it strongly reconsiders the high value its frameworks continue to place on using it as a tool for helping to reach classification decisions. Only once we've begun to stabilize validity within the individual frameworks themselves—getting a hold on what we are even doing when we carry out our validation procedures—should validity have its special status returned.

References for Chapter 6

American Psychiatric Association (2023). *Submit proposals for making changes to DSM-5-TR*. https://www.psychiatry.org/psychiatrists/practice/dsm/submit-proposals

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Borsboom, D. (2017). A network theory of mental disorders. *World psychiatry*, *16*(1), 5–13

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in education: Principles, policy & practice*, *23*(2), 212–225.

DeYoung, C. G., & Krueger, R. F. (2018). A cybernetic theory of psychopathology. *Psychological Inquiry*, *29*(3), 117-138.

First, M. B., Regier, D. A., & Kupfer, D. J. (2002). *A research agenda for DSM-V*. American psychiatric pub.

Haeffel, G. J., Jeronimus, B. F., Kaiser, B. N., Weaver, L. J., Soyster, P. D., Fisher, A. J., Vargas, I., Goodson, J. T., & Lu, W. (2022). Folk classification and factor rotations: Whales, sharks, and the problems with the hierarchical taxonomy of psychopathology (HiTOP). *Clinical psychological science*, *10*(2), 259–278.

DeYoung, C. G., Kotov, R., Krueger, R. F., Cicero, D. C., Conway, C. C., Eaton, N. R., ... & Wright, A. G. (2022). Answering questions about the hierarchical taxonomy of psychopathology (HiTOP): Analogies to whales and sharks miss the boat. *Clinical psychological science*, *10*(2), 279–284.

Haeffel, G. J., Jeronimus, B. F., Kaiser, B. N., Weaver, L. J., Soyster, P. D., Fisher, A. J., Vargas, I., Goodson, J. T., & Lu, W. (2022). The hierarchical taxonomy of

psychopathology (HiTOP) is not an improvement over the *DSM*. *Clinical psychological science*, *10*(2), 285–290.

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.

Jablensky, A. (2016). Psychiatric classifications: Validity and utility. *World psychiatry*, *15*(1), 26–31.

Kendler, K. S., & Parnas, J. (2012). Epistemic iteration as a historical model for psychiatric nosology: Promises and limitations. *Philosophical issues in psychiatry II: Nosology*, 305–322.

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., ... & Zimmerman, M. (2017). The hierarchical taxonomy of psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of abnormal psychology*, *126*(4), 454-477.

Lee, C. T., Kelley, S. W., Palacios, J., Richards, D., & Gillan, C. M. (2023). Estimating the prognostic value of cross-sectional network connectivity for treatment response in depression. *Psychological medicine*, 1–10.

Michelini, G., Palumbo, I. M., DeYoung, C. G., Latzman, R. D., & Kotov, R. (2021). Linking RDoC and HiTOP: A new interface for advancing psychiatric nosology and neuroscience. *Clinical psychology review*, *86*, 102025.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary research and perspectives*, *15*(2), 51–69.

Markus, K. A. (2016). Alternative vocabularies in the test validity literature. *Assessment in education: Principles, policy & practice*, *23*(2), 252–267.

Morris, S. E., Sanislow, C. A., Pacheco, J., Vaidyanathan, U., Gordon, J. A., & Cuthbert, B. N. (2022). Revisiting the seven pillars of RDoC. *BMC medicine*, *20*(1), 220-230.

Newton, P. E., & Baird, J. A. (2016). The great validity debate. *Assessment in education: Principles, policy & practice*, *23*(2), 173–177.

Newton, P. E., & Shaw, S. D. (2016). Disagreement over the best way to use the word 'validity'and options for reaching consensus. *Assessment in education: Principles, policy & practice*, *23*(2), 178–197.

Phillips, J. (2013). The conceptual status of *DSM-5* diagnoses. In *Making the DSM-5: Concepts and controversies* (pp. 143–157). Springer.

Robins, E., & Guze, S. B. (1970). Establishment of diagnostic validity in psychiatric illness: Its application to schizophrenia. *American journal of psychiatry*, *126*(7), 983–987.

Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. Springer.

Solomon, M. (2022). On validators for psychiatric categories. *Philosophy of medicine*, *3*(1), 1–23.

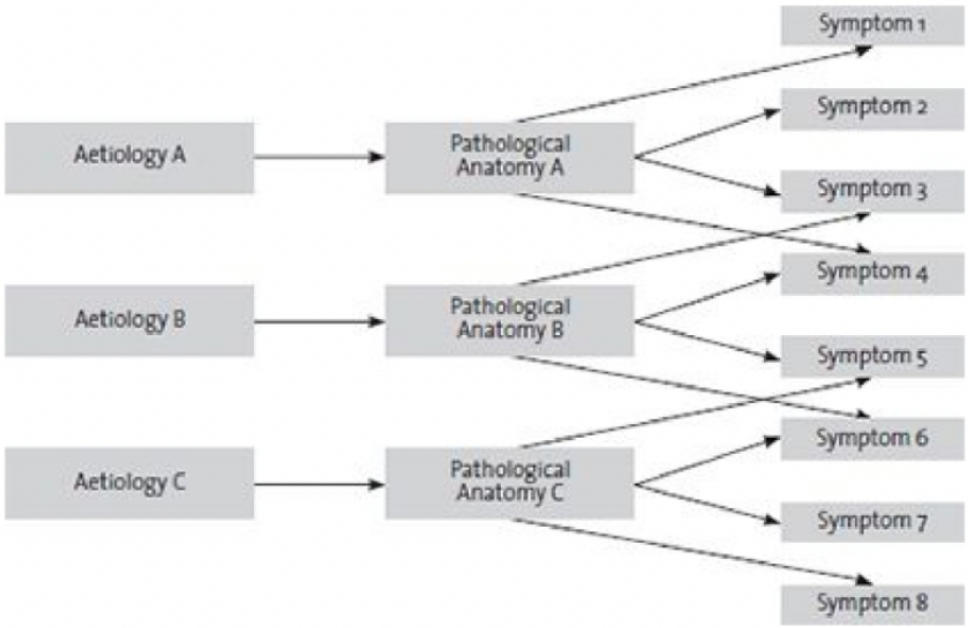
Sullivan, J. A. (2016a). Stabilizing constructs through collaboration across different research fields as a way to foster the integrative approach of the research domain criteria (RDoC) project. *Frontiers in human neuroscience*, *10*, 309-313.

Sullivan, J. A. (2016b). Construct stabilization and the unity of the mind-brain sciences. *Philosophy of science*, *83*(5), 662–673.

APPENDIX

Figure 1

Emil Kraeplein's big idea.



Reprinted from Bentall, R. P. (2003). *Madness explained: Psychosis and human nature*. Penguin UK.

Table 1
Cloninger's interpretation of validity types included in Robins and Guze's five phases.

Phase	Psychometric Validity Type
(1) Clinical Description	Content Validity
(2) Laboratory Studies	Concurrent Validity (Criterion-Oriented Validity)
(3) Delimitation from Other Disorders	Discriminant Validity (Construct Validity)
(4) Follow-Up Study	Predictive Validity (Criterion-Oriented Validity)
(5) Family Study	Criterion-Oriented Validity

Reprinted from Cloninger, C. R. (1989). Establishment of diagnostic validity in psychiatric illness: Robins and Guze's method revisited. *The validity of psychiatric diagnosis*, 9-18.

Table 2
List of validators of the DSM-5-TR.

Antecedent Validators
a. *Familial aggregation and/or co-aggregation (i.e., family, twin, or adoption studies)
b. Socio-demographic and cultural factors
c. Environmental risk factors
d. Prior psychiatric history

Concurrent Validators
a. Cognitive, emotional, temperament, and personality correlates (unrelated to the diagnostic criteria)
b. *Biological markers, e.g., molecular genetics, neural substrates
c. Patterns of comorbidity
d. *Degree or nature of functional impairment

Predictive Validators
a. *Diagnostic stability
b. *Course of illness
c. *Response to treatment

Antecedent Validators*:* Evidence that can be accumulated from past clinical data and research. Types include a) familial aggregation, b) sociodemographic and cultural factors, c) environmental risk factors, and d) prior psychiatric history.

Familial Aggregation*: Evidence describing genetic and non-genetic influences posited to underly a diagnostic category from family, twin, and adoption studies. Evidence of genetic correlates supports a diagnostic category as representative of a distinctive disorder.

Sociodemographic and Cultural Factors: Evidence describing sociodemographic-specific and cultural-specific risk factors associated with a diagnostic category from community and population (observational) studies. Sociodemographic and cultural risk factors are thought to support more accurate diagnostic categories that may feature more culturally specific criteria, e.g., alternative symptom expressions, variations in the boundaries of disorders in relation to specific groups, i.e., differing diagnostic cut-offs, etc.

Environmental Risk Factors: Evidence describing environmental risk factors associated with a diagnostic category from observational studies. Analysis of specific environmental factors is thought to help in determining boundaries between disorders that may share common risk factors (e.g., early childhood adversity or trauma-related loss) as well as highlight those risk factors that are disorder-specific.

Prior Psychiatric History: Evidence that describes the relationship between past psychiatric assessment and the prevalence of meeting specific diagnostic criteria for

particular disorder based on observational studies. Such evidence is thought to help distinguish between subgroups within a diagnostic category.

Concurrent Validators: Evidence that can be accumulated from clinical studies. Types include a) cognitive, emotional, temperamental, and personality correlates, b) biological markers, c) patterns of comorbidity, and d) degree or nature of functional impairment.

Cognitive, emotional, temperamental, and personality correlates: Evidence that describes correlations between specific cognitive, emotional, temperamental, and personality factors and a particular diagnostic category based in observational studies. Such evidence is thought to provide external support to diagnostic groupings in addition to providing potential predictive utility.

Biological Markers*: Evidence that describes the pathophysiological correlates of a diagnostic category from studies such as pathological findings, blood tests, genetic tests, and neuroimaging studies. Such studies are thought to provide external support to a particular grouping of a diagnostic category.

Patterns of Comorbidity: Evidence that describes the frequency of co-occurrence between two or more disorders from observational studies. Evidence of comorbidity may highlight a lack of distinctiveness of a particular diagnostic category from another.

Degree of nature of functional impairment*: Evidence describing the categorical and/or dimensional features (e.g., levels of severity) of a diagnostic category.

Predictive Validators: Evidence that is accumulated from follow-up studies. Types include diagnostic stability, course of illness, and response to treatment.

Diagnostic Stability*: Evidence that describes whether a diagnostic category may evolve over time. Evidence of a lack of change in a diagnostic category is supportive of that category being distinct from other categories.

Course of Illness: Evidence of prognostic information for a diagnostic category's predicted duration. Consistent and predictive course of illness is considered to support the accuracy and distinctiveness of diagnostic criteria.

Response to treatment: Evidence of response to both pharmacological and psychotherapeutic interventions. Ability to predict specific and consistent responses to treatment provides support for diagnostic criteria.

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.). https://doi.org/10.1176/appi.books.9780890425787

Table 3
Diagnostic criteria A for Major Depressive Disorder.

A. Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure.
Note: Do not include symptoms that are clearly attributable to another medical condition.

1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad, empty, hopeless) or observation made by others (e.g., appears tearful). (**Note:** In children and adolescents, can be irritable mood.)
2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation).
3. Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month) or decrease or increase in appetite nearly every day. (**Note:** In children, consider failure to make expected weight gain.)
4. Insomnia or hypersomnia nearly every day.
5. Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).
6. Fatigue or loss of energy nearly every day.
7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
8. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).
9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.).
https://doi.org/10.1176/appi.books.9780890425787

Table 4
Meets Criteria for a Mental (Psychiatric) Diagnosis.

Features

- a behavioral or psychological syndrome or pattern that occurs in an individual
- the consequences of which are clinically significant distress (e.g., a painful symptom) or disability (i.e., impairment in one or more important areas of functioning)
- must not be merely an expectable and culturally sanctioned response to a particular event, for example, trance states in religious rituals
- that reflects an underlying psychobiological disturbance
- that is not solely a result of social deviance or conflicts with society
- that has diagnostic validity using one or more sets of diagnostic validators (e.g., prognostic significance, psychobiological disruption, response to treatment)
- that has clinical utility (e.g., contributes to better conceptualization of diagnoses, or to better assessment and treatment)

Other Considerations

- no definition adequately specifies precise boundaries for the concept of either "medical diagnosis" or "mental/psychiatric diagnosis"
- diagnostic validators and clinical utility should help to differentiate a diagnosis from diagnostic nearest neighbors
- in adding/deleting entities from the nomenclature, potential benefits (e.g., provide better patient care, stimulate new research) should outweigh potential harms (e.g., hurt particular individuals, be subject to misuse)

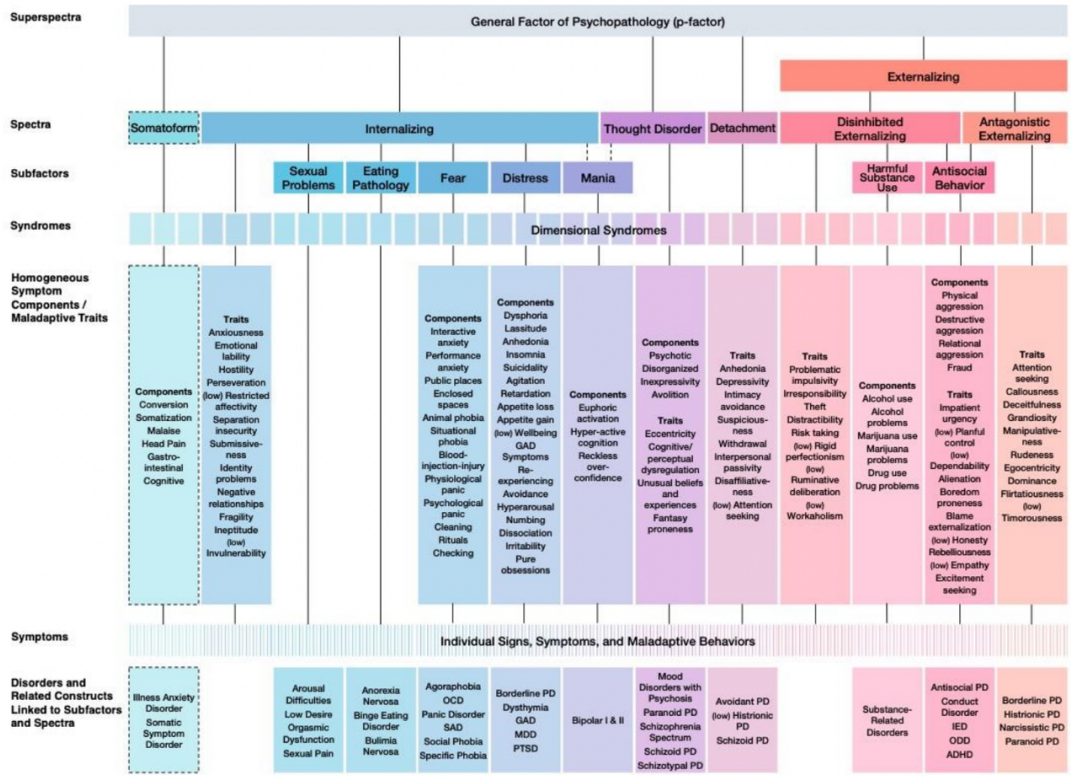
Kendler, K.S., Kupfer, D., Narrow, W., Phillips, K., & Fawcett, J. (2009). Guidelines for making changes to *DSM-V*. Unpublished manuscript.

Table 5
The Five Phases.

1. Clinical Description: Describe the overall "clinical picture" of the disorder. May include signs and symptoms, as well as other relevant features such as race, sex, age at onset, precipitating factors, etc.
2. Laboratory Studies: Relate the clinical description of the disorder with biological (e.g., chemical, physiological, radiological) and anatomical (e.g., biopsy and autopsy) findings, as well as psychological tests when appropriate.
3. Delimitation from other disorders: Specify exclusion criteria to ensure that the group being studied is as homogenous as possible.
4. Follow-Up Study: Determine whether the original patients are suffering from some other disorder that could better account for the original clinical description.
5. Family Study: Identify the increased prevalence of the same disorder among close relatives of the original patients.

Robins, E., & Guze, S. B. (1970). Establishment of diagnostic validity in psychiatric illness: Its application to schizophrenia. American journal of psychiatry, 126(7), 983–987.

Figure 2.
The HiTOP Model.



HiTOP constructs center around *Spectra,* the main common factors underlying psychopathology informed by 1st order factor analytic studies. Constructs "lower" in the hierarchy are chosen from empirical measures of symptom data "not designed for structural research" (Kotov et al., 2017, p. 15), whereas constructs one level below or one (to two) levels above the spectra are informed by 2nd (or 3rd) order factor analytic studies.

Reprinted from Forbes, M. K., Ringwald, W. R., Allen, T., Cicero, D. C., Clark, L. A., DeYoung, C. G., Eaton, N., Kotov, R., Krueger, R. F., Latzman, R. D., Martin, E. A., Naragon-Gainey, K., Ruggero, C. J., Waldman, I. D., Brandes, C., Fried, E. I., Goghari, V. M., Hankin, B., Sperry, S., . . . Wright, A. G. C. (2024). Principles and procedures for revising the hierarchical taxonomy of psychopathology. *Journal of Psychopathology and Clinical Science, 133*(1), 4–19.

Table 6
List of Five Types of External Validity for HiTOP.

Convergent validity: the degree to which there is a relationship between different measures used to assess the same or similar construct.

Discriminant validity: the degree to which there is a lack of a relationship between different measures used to assess two or more independent constructs.

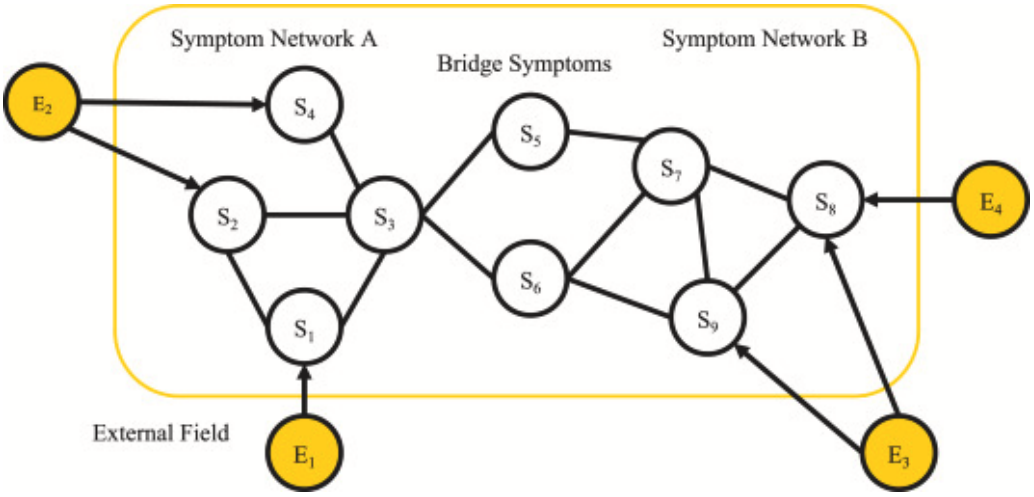
Concurrent validity: the degree to which measures of a construct correlate with other indicators of that construct measured at the same time.

Aetiological validity: assesses the degree to which measures of a construct correlate with etiological factors (e.g., biological, environmental) associated with the same or similar construct.

Prospective validity: the degree to which measures of a construct may correlate with other indicators of that construct measured at a future time.

Forbes, M. K., Ringwald, W. R., Allen, T., Cicero, D. C., Clark, L. A., DeYoung, C. G., Eaton, N., Kotov, R., Krueger, R. F., Latzman, R. D., Martin, E. A., Naragon-Gainey, K., Ruggero, C. J., Waldman, I. D., Brandes, C., Fried, E. I., Goghari, V. M., Hankin, B., Sperry, S., . . . Wright, A. G. C. (2024). Principles and procedures for revising the hierarchical taxonomy of psychopathology. Journal of Psychopathology and Clinical Science, 133(1), 4–19.

Figure 3
Two Connected Symptom Networks



Adapted from Freund, I. M., Arntz, A., Visser, R. M., & Kindt, M. (2022). Jumping back onto the giants' shoulders: Why emotional memory should be considered in a network perspective of psychopathology. Behaviour research and therapy, 156, 104154.

Table 7
RDoC entry for the Loss construct.

Domain: Negative Valence Systems

Description: A state of deprivation of a motivationally significant con-specific, object, or situation. Loss may be social or non-social and may include permanent or sustained loss of shelter, behavioral control, status, loved ones, or relationships. The response to loss may be episodic (e.g., grief) or sustained.

Molecules	Androgens, CRH, Estrogens, Glucocorticoid receptors, Inflammatory Molecules, Oxytocin, Vasopressin
Circuits	Amygdala, Default mode network, Dorsolateral Prefrontal Cortex, Habit systems (Striatum/caudate/accumbens), Hippocampus, Insular, Orbitofrontal cortex, Parietal cortex, Posterior Cingulate Gyrus, PVN, Reward circuity, vmPFC
Physiology	ANS, HPA, neuroimmune, Prolonged psychophysiological activity

Adapted from National Institute of Mental Health. (2024) *RDoC Constructs: Loss. https://www.nimh.nih.gov/research/research-funded-by-nimh/RDoC/constructs/loss*.

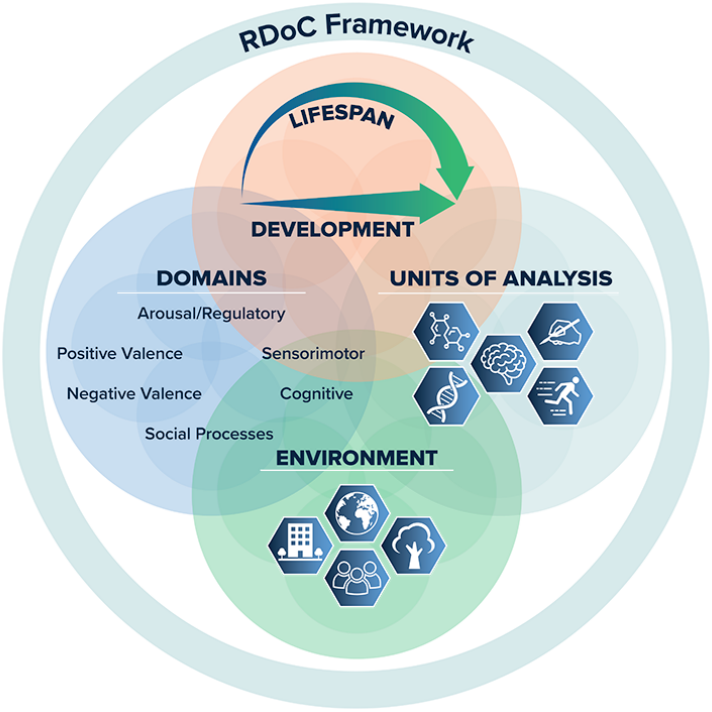
Table 8
Original conceptualization of the RDoC matrix.

			UNITS OF ANALYSIS					
DOMAINS/CONSTRUCTS	**Genes**	**Molecules**	**Cells**	**Circuits**	**Physiology**	**Behavior**	**Self-Reports**	**Paradigms**
Negative Valence Systems								
Acute threat ("fear")								
Potential threat ("anxiety")								
Sustained threat								
Loss								
Frustrative nonreward								
Positive Valence Systems								
Approach motivation								
Initial responsiveness to reward								
Sustained responsiveness to reward								
Reward learning								
Habit								
Cognitive Systems								
Attention								
Perception								
Working memory								
Declarative memory								
Language behavior								
Cognitive (effortful) control								
Systems for Social Processes								
Affiliation/attachment								
Social communication								
Perception/understanding of self								
Perception/understanding of others								
Arousal/Modulatory Systems								
Arousal								
Biological rhythms								
Sleep-wake								

Rows represent various constructs grouped hierarchically into broad domains of function. The columns of the matrix denote different levels of analysis, from genetic, molecular, and cellular levels, proceeding to the circuit level, being the focal element of the RDoC organization (Insel et al., 2010).

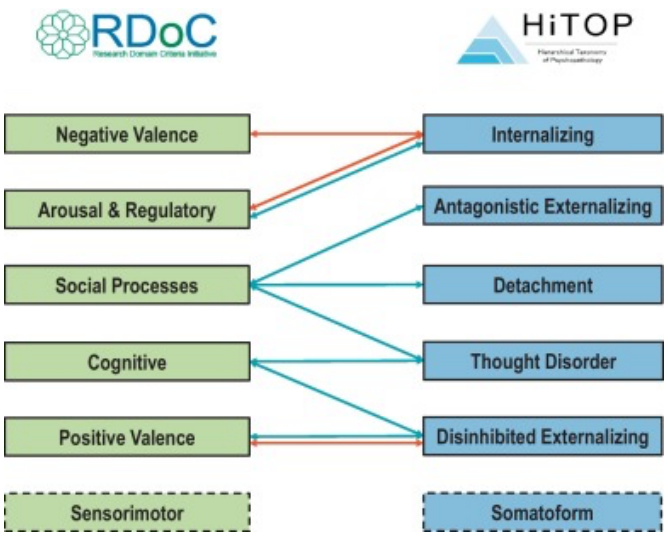
Reprinted from Cuthbert, B. N. (2014). The RDoC framework: Facilitating transition from *ICD*/*DSM* to dimensional approaches that integrate neuroscience and psychopathology. *World psychiatry*, *13*(1), 28–35.

Figure 4
The RDoC Research Framework.



Concentric circles represent 1) Domains of Functioning; 2) Units of Analysis; 3) Environment; and 4) Development across the lifespan. Reprinted from Morris, S. E., Sanislow, C. A., Pacheco, J., Vaidyanathan, U., Gordon, J. A., & Cuthbert, B. N. (2022). Revisiting the seven pillars of RDoC. *BMC medicine*, *20*(1), 220.

Figure 5
Overview of suggested links between the Research Domain Criteria (RDoC) and the Hierarchical Taxonomy of Psychopathology (HiTOP)



Arrows show such links between RDoC domains and HiTOP spectra. Red = positive association; Blue = negative association. Reprinted from Michelini, G., Palumbo, I. M., DeYoung, C. G., Latzman, R. D., & Kotov, R. (2021). Linking RDoC and HiTOP: A new interface for advancing psychiatric nosology and neuroscience. *Clinical psychology review*, *86*, 10202.

Nicholas Zautra

Cognitive Science Program
Indiana University Bloomington
1001 E. 10th St.
Bloomington, IN 47405

Email:
nicholaszautra@gmail.com

Area of Specialty Cognitive Science

Area of Concentration Philosophy of Science

Education

PhD in Cognitive Science — May, 2024
Minor in History and Philosophy of Science, Minor in Philosophy
Indiana University Bloomington
Dissertation: "Psychiatry's Second Validity Crisis: The Problem of Disparate Validation"

MA in History and Philosophy of Science — December, 2022
Indiana University Bloomington
Research areas: Philosophy of Psychiatry

MA in Applied Ethics and the Professions: Biomedical and Health Ethics — August, 2013
Arizona State University
Research areas: Bioethics; Clinical Ethics

BA (Hons.) in Psychology — May 2012
Arizona State University; Barrett, the Honors College
Research areas: Behavioral Neuroscience, Cognitive Psychology

Refereed Publications

- Zautra, N. (2015). Embodiment, Interaction, & Experience: Toward a Comprehensive Model in Addiction Science. *Philosophy of Science,* 82(5).
- Zautra, N. (2015). Rethinking the Conceptual History of the Term 'Cognitive'. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.) *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Kufahl, P. R., Watterson, L. R., Nemirovsky, N. E., Hood, L. E., Villa, A., Halstengard, C., Zautra, N., & Foster Olive, M. (2013). Attenuation of methamphetamine seeking by the mGluR2/3 agonist LY379268 in rats with histories of restricted and escalated self-administration. *Neuropharmacology, 66,* 290-301.
- Kufahl, P. R., Nemirovsky, N. E., Watterson, L. R., Zautra, N., & Olive, M. F. (2013). Positive or negative allosteric modulation of metabotropic glutamate receptor 5 (mGluR5) does not alter expression of behavioral sensitization to methamphetamine. *F1000Research*, 2.

Ongoing Research

• Zautra, N. (2024). Validity and Validation for a New Scientific Psychiatry: The Heirarchical Taxonomy of Psychopathology (HiTOP) vs. Research Domain Criteria (RDoC) (In preparation).

Presentations

- Zautra, N. (2024, May). Psychiatry's Second Validity Crisis? The Problem of Disparate Validation. Paper presented at the AAPP meeting, New York, NY.
- Zautra, N. (2019, March). The Philosopher of Science is by necessity a social animal. Paper presented at the 2019 Mid/South Philosophy of Science Meeting, Lexington, KY
- Zautra, N. (2018, November). The greatest challenge facing philosophy of science today (according to philosophers of science). Poster presented at the 2018 Philosophy of Science Association Meeting, Seattle, WA.
- Zautra, N. (2017). RDoC: Toward a Historical Understanding of National Institute of Mental Health's Brain-Centric Research Program. Paper presented at the Cognitive Lunch Colloquium meeting at Indiana University, Bloomington, IN.
- Zautra, N. (2016, May). From DSM to RDoC: On the Beginning of the End of Animal Models of Mental Illness. Poster presented at the Society for Philosophy and Psychology 42nd Annual Meeting, University of Texas at Austin, Austin, TX.
- Zautra, N. & Robert, J. (2014, October). Core Competencies for "Benchside" Research Ethics Consultation Services. Paper presented at the American Society for Bioethics and Humanities 16th Annual Meeting, San Diego, CA.
- Zautra, N. & Robert, J. (2013, July). Humanizing Animals: The Selection and Justification of the Prairie as an Animal Model for Autism Spectrum Disorders. Paper presented at the ISHPSSB meeting, Montpellier, France.

Honors and Awards

MAGS Excellence in Teaching Award (Nominated)	2017
Cognitive Science Supplemental Research Fellowship	2017
IU Lieber Memorial Teaching Associate Award	2017
Cognitive Science Supplemental Research Fellowship	2015
Poynter Center Jesse Fine Teaching Fellowship	2014
National Science Foundation Travel Grant	2014
Arizona State University GPSA Travel Award	2013
School of Life Sciences Travel Award	2013
Center for Biology & Society Travel Award	2013
ASU Graduate Humor Conference Competition Honorable	2013
Mention Center for Biology & Society Travel Award	2012
Not My Kid Speaker Scholarship	2012
Barrett, The Honors College Honors Project Funding Award	2011

Service Activities

Teacher Panelist. *D. Ames Shuel Academic Center*	2017
Reviewer. *Philosophical Psychology*	2016
Judge. *Indiana University Ethics Bowl*	2016
Reviewer. Indiana Unveristy HPS Graduate Student Conference	2016
Reviewer. Southwest Graduate Student Philosophy Conference	2013
Judge. Central Arizona Chapter of Association for Women in Science Awards of Excellence, Arizona Science & Engineering Fair.	2013