Running title: Precision phenotyping for psychopathology

Precision behavioral phenotyping as a strategy for uncovering the biological correlates of psychopathology

Jeggan Tiego[1*], Elizabeth Martin[2], Colin G. DeYoung[3], Kelsey Hagan[4], Samuel E. Cooper[5], Rita Pasion[6], Liam Satchell [8], Alexander J. Shackman[9], Mark A. Bellgrove, and Alex Fornito[1] and the HiTOP Neurobiological Foundations Work Group

[1] Turner Institute for Brain and Mental Health and School of Psychological Sciences, Monash University, Melbourne, Victoria, Australia

[2] Department of Psychological Science, University of California, Irvine, Irvine, CA, USA

[3] Department of Psychology, University of Minnesota, Minneapolis, MN, USA

[4] Department of Psychiatry, Columbia University Irving Medical Center, New York, NY, USA

[5] Department of Psychiatry and Behavioral Sciences, University of Texas at Austin

[6] Lusófona University, HEI-LAB, Portugal

[8] Department of Psychology, University of Winchester, UK

[9] Department of Psychology, University of Maryland, College Park, MD, USA

*Correspondence:

Jeggan Tiego

Monash Biomedical Imaging

770 Blackburn Road, Clayton VIC 3168, Australia

Reference count: 100

Abstract word count: 148

Glossary word count: 614

Word count: 5,310

Number of tables: 1

Number of figures: 3

Abstract

Our capacity to measure diverse aspects of human biology has developed rapidly in the past decades, but the rate at which these techniques have generated insights into the biological correlates of psychopathology has lagged far behind. Slow progress is partly due to the poor sensitivity, specificity, and replicability of many findings in the literature, which have in turn been attributed to small effect sizes, small sample sizes, and inadequate statistical power. A commonly proposed solution is to focus on large, consortia-sized samples Yet it is abundantly clear that increasing sample sizes will have a limited impact unless a more fundamental issue is addressed: the precision with which target behavioral phenotypes are measured. We discuss challenges and outline several ways forward and provide worked examples to demonstrate key problems and potential solutions. Arguably, a precision phenotyping approach can enhance the discovery and replicability of associations between biology and psychopathology.

Keywords: *psychopathology; psychiatric neuroimaging; psychiatric genetics; biomarkers; psychometrics.*

Glossary

Attenuation bias – the attenuation (reduction) of a statistical effect (e.g., correlation) due to measurement error (i.e., imperfect reliability)

Berkson's bias – a group of individuals is sampled from a subpopulation rather than a representative population resulting in spurious associations between the target health condition and other measured characteristics, such as risk factors

Bias – systematic divergence of statistical parameter estimates from true effects

Biomarker – a biological measure that can indicate susceptibility or risk for disease, presence of disease, or response to an intervention

Clinician's illusion – a clinical sample biased towards those individuals that experience the condition over a longer duration, who are therefore more likely to present with the condition at the time of sampling, versus those individuals that will ever present with the condition (i.e., prevalence versus incidence)

Comorbidity problem – psychiatric disorders co-occur in the same individuals more often than would be expected for independent entities, suggesting shared phenomenology and etiology

Construct homogeneity – the assumption or evidence that a construct reflects variance in a single phenotype (i.e., unidimensionality)

Criterion variable – a variable being predicted by a study; an outcome or dependent variable

Deep phenotyping – comprehensive assessment of one or more phenotypes. Contrasts with minimal phenotyping, which uses relatively few or superficial measures of the target phenotypes

Effect size – the strength of a statistical effect (e.g., mean difference or correlation) in standardized units, permitting direct comparison across approaches and samples

Efficiency – the size of the standard error of a statistical parameter estimate. Efficiency is inversely proportional to the size of the standard error

Factor Indicators – Measured or "manifest" variables that provide the observed indicators of a latent variable within a common factor model

Genotype – the genetic makeup of an individual, comprised of the full set of genetic variants that shape phenotypic variability

Heterogeneity problem – the grouping of cases with divergent symptom presentations (e.g., too much or too little sleep) into the same diagnostic category (e.g., major depression), or the grouping of symptoms with divergent etiology, pathophysiology, course and/or treatment response

Homogeneity assumption – the assumption that different people with the same psychiatric diagnosis (e.g., posttraumatic stress disorder) are phenotypically similar

Imaging-derived phenotype – a phenotype measured using neuroimaging techniques (e.g., MRI)

Latent variable/trait – an unobserved construct that is related to measured variables through mathematical models (e.g., covariance between measured variables within a common factor model)

Measurement invariance – the structural association between indicators and a construct is equivalent between groups and over time

Method bias – sources of systematic measurement error stemming from the measurement process (i.e., method effects) for constructs (e.g., self-report, interview, etc.)

Phenotype – an observable characteristic or trait. Used here in reference to psychopathology-related behavioural characteristics, unless otherwise specified.

Phenotypic complexity – the extent to which a phenotype reflects multiple sources of variance, including from one or more hierarchical levels extending from broad generality to narrow specificity

Phenotypic resolution – the reliability or precision of measurement of a phenotype along the full spectrum of the latent trait continuum

Polythetic-categorical constructs / polythetic diagnoses – diagnoses defined by an established minimum number of criteria, not all of which are required for diagnosis (e.g., any 4 of 8 symptoms). Polythetic diagnostic criteria contribute to the heterogeneity problem (see above)

Replicability – the extent to which findings replicate from one study and one sample to another

Signal-to-noise ratio – the ratio of meaningful variance to error variance

Statistical power – the probability of rejecting the null hypothesis when the alternative hypothesis is true; increases with greater measurement reliability and bigger sample sizes

Unidimensionality – the covariance amongst a homogenous item set is captured by one factor or latent variable, as opposed to two or more factors in the case of multidimensionality

Introduction

A comprehensive understanding of psychopathology requires a systematic investigation of functioning at multiple levels of analysis, from genes to brain to behavior[1,2]. The development and widespread use of technologies—including magnetic resonance imaging (MRI) and inexpensive genetic assays—promised to transform our understanding of psychiatric disorders[3] and lead to **biomarkers** that would enhance diagnosis, prognosis, and treatment[4]. However, increasing technological advances and sophistication in the acquisition and analysis of these data have generally failed to produce consistent research findings with broad and significant clinical relevance to the diagnosis and treatment of mental disorders[5]. Biology-psychopathology associations are typically small[6], often fail to replicate[7], and generally lack diagnostic specificity[8-10]. In short, despite decades of work, thousands of studies, and hundreds of millions of research dollars, modern neuroimaging and genetic tools have largely failed to uncover clinically actionable insights into psychopathology[11,12].

Modest effects and poor replicability have promoted calls to establish consortia-sized samples to identify reproducible biology-psychopathology associations[7], with theoretical and empirical studies indicating that problems of low power and replicability can be addressed with sample sizes ranging from the thousands to tens of thousands[6,7]. This approach has become standard in molecular genetics and has yielded reliable genetic "hits" for several psychiatric disorders[12]. Recent analyses suggest a similar approach may be necessary for neuroimaging studies[6]. Other investigators have focused on improving the validity and accuracy of neuroimaging measures, through the use of sophisticated data acquisition techniques[13], improved denoising techniques[14], and individually-tailored analyses[15]. Similarly, in genetics, growing interest in moving beyond common genetic variation to study high-effect rare variants mandates an order of magnitude increase in sample size[16].

In this review, we suggest that such attempts will have limited success unless we develop more precise or statically optimized psychiatric **phenotypes**. We begin by briefly summarizing the adverse consequences of phenotypic imprecision for discovering reproducible biology-psychopathology associations and highlight some of the most common types of imprecision. We then provide concrete recommendations for **precision phenotyping** that will help overcome these challenges. Throughout the review, we provide worked examples of key concepts, using genetic data obtained at the baseline wave (n = 2,218) and behavioral data obtained from the 2-year follow-up wave (N = 5,820) of the Adolescent Brain Cognitive Development (ABCD) study (behavioral data – release 3.0, genetic data – release 2.0)[17].These examples support the conclusion that phenotypic imprecision can thwart the consistent detection of potentially important biology-psychopathology associations. In each case, we describe countermeasures that can be deployed to bolster precision and reliability. Taken together, these strands of psychometric theory and empirical data suggest that the systematic adoption of precision phenotyping has the potential to substantially accelerate efforts to understand the neurogenetic correlates of psychopathology and, ultimately, set the stage for developing more effective clinical tools.

Note that we focus on mental health measures in our manuscript because: (a) the limitations of such measures are rarely discussed in comparison with the extensive literature devoted to improving biological measures; (b) prevalent practices to measure behavior are sub-optimal; and (c) addressing these sub-optimal practices is arguably the most cost-effective and quickest way of improving current methodologies. It also merits comment that, while this review is centered on psychiatric phenotypes, we note that biological measures are also prone to error and may equally contribute to the problems of weak signal in biology-psychopathology association studies[18]. Thus, our proposals parallel considerable efforts

devoted to improving the validity and accuracy of **imaging-derived phenotypes** [13-15], which is sometimes also called precision phenotyping.

The impact of measurement imprecision on detecting and replicating associations between biology and psychopathology

An important step in understanding and treating psychiatric disorders is the identification of pathophysiological mechanisms. Doing so requires the discovery of robust associations between biology and psychiatric phenotypes, an endeavor that is fundamentally constrained by the validity and reliability of the measured phenotypes. *Validity* concerns the correspondence between a psychological measure and the construct it is designed to measure. If a psychological measure fails to measure a real entity or changes in the state of that entity fail to produce systematic variations in the psychological measure, any analyses that rely on the psychological measure will be inaccurate. *Reliability* refers to the consistency of a measure across items, scales, occasions, or raters; and is the inverse of measurement error. Lower reliability (higher error) contributes to noisy estimates and decreased accuracy of rank-ordering of individuals when measuring biology-psychopathology associations[19]. In fact, reliability imposes an upper limit on the magnitude of linear associations that can be detected (i.e., observed biology-psychopathology associations are inversely proportional to measurement reliability), mandating larger and more expensive samples for adequate power and reproducibility[20] (see **Box 1**). In sum, adequate validity and reliability are necessary for identifying robust and meaningful biology-psychopathology associations[20,21]. It is noteworthy that phenotypic precision is a necessary, but not sufficient, condition for uncovering biology-behavior associations. For example, measurement of human intelligence is well-developed psychometrically and yet our understanding of the neurobiology and genetics of intelligence is not yet complete. The validity and reliability of psychiatric phenotypes can be compromised by a variety of factors, which we collectively refer to as *phenotypic*

imprecision. In this section, we highlight common and pernicious causes of phenotypic imprecision.

1. Sampling biases

Different research aims demand specific sampling strategies. For studies seeking to identify biology-psychopathology associations, it is important to have samples that are representative of the population of interest and that maximize **statistical power** for this research design. Sampling biases, non-representative samples, and generalizability issues have been broadly discussed in the literature[22], but several specific aspects of sampling bias are particularly relevant to the measurement of psychiatric phenotypes in biological association studies. As a primary example, most psychiatric neuroimaging and genetic research has focused on examining case-control differences defined by traditional diagnostic frameworks, such as the *Diagnostic and Statistical Manual for Mental Disorders* (*DSM-5*) and the *International Classification of Diseases (ICD-11)*. These frameworks have questionable reliability and validity[23], and likely show a limited correspondence with biological correlates (**Box 2**). Indeed, there is ample evidence that psychiatric phenotypes are dimensional[23], indicating that distinctions between cases and controls based on arbitrary clinical cut-points can artificially reduce statistical power for detecting associations with biological measures; the so-called 'curse of the clinical cut-off''[24] (but see Fisher et al.[25]). The approach may also complicate attempts to identify at-risk individuals with subclinical/subthreshold symptomatology[26] and may result in only a subpopulation of the most severely affected individuals being sampled, leading to problems such as **Berkson's bias** and the **Clinician's Illusion**.

A further complication arises with the recruitment of appropriate control groups. Researchers often exclude controls who endorse past or current *DSM-5* or *ICD-11* diagnoses

or other signs of morbidity, resulting in an unrepresentative 'super control' group. When compared to a group of patients meeting diagnostic threshold, the resulting study design embodies an extreme-groups approach rather than a simple dichotomization of a dimensional variable. Such designs, when applied to the study of dimensional phenomena, are known to confer biased effect estimates[27]. We acknowledge that traditional approaches to clinical description and diagnosis of mental disorders have clinical utility[26]. However, in this article, we explore the application and implications of refined approaches to studying the biological correlates of psychopathology in research rather than clinical contexts. The importance of ethnic and demographic diversity with respect to representativeness, ethnic matching of biological measures, and generalizability of predictions of behaviour from biology, has also been discussed in the literature[28,29]. Crucially, some cross-cultural initiatives in population neuroscience and genetics have been developed to meet this need[29-31].

2. Minimal and inconsistent phenotyping

The sheer cost and practical challenges of large-scale recruitment and testing often mean that the time and resources available for psychiatric phenotyping are limited[32]. Minimal or 'shallow' phenotyping, is one of the more commonly encountered causes of phenotypic imprecision in biological studies of psychopathology[32]. Minimal phenotyping is one-shot assessment using single, and sometimes abbreviated, scales. This will increase the proportion of occasion-specific state variance (error) compared to averaging across two or more occasions, and attenuate biology-psychopathology associations. Furthermore, minimal phenotyping may fail to capture important aspects of psychopathology that are associated with biological measures.

Aggregation of data in consortia is further complicated by substantive differences in phenotypic assessment across sites. Numerous scales and questionnaires are available for

assessing common psychiatric conditions (e.g., depression) and these measures vary greatly in their inclusion and emphasis of symptoms[33]. Minimal phenotyping exacerbates the **heterogeneity problem**[34], because superficially similar cases—for instance, individuals self-reporting a lifetime history of depression in response to a single self-report probe—likely diverge on important, but unmeasured characteristics, dampening effect sizes and power. For example, Schnack and Kahn[35] have demonstrated that increasing sample sizes for neuroimaging research of schizophrenia may result in samples that are more heterogeneous, which can lead to lower prediction accuracy in machine learning analyses. This aligns with evidence that people diagnosed with schizophrenia and other disorders often show considerable heterogeneity in biological phenotypes[36]. Similarly, large clinical cohorts forming the reference samples for genome-wide association studies (GWAS) may also be heterogeneous in terms of clinical phenomenology, which is not revealed by minimal phenotyping[37]. Thus, despite the advantages of large samples, counterintuitively, increasing sample sizes through consortia-like data pooling may result in decreased, rather than increased, **signal-to-noise ratio**. Therefore, the quest for ever-larger sample sizes, without consideration of precision phenotyping, is neither efficient nor economical, and will not, on its own, ensure the discovery and replicability of biology-psychopathology associations [38].

3. Phenotypic complexity

The use of raw behavioral scores in simple bivariate correlational (or related) analyses with biological variables assumes a unifactorial and non-hierarchical structure of the target phenotype. However, psychiatric phenotypes often have a multidimensional and hierarchical structure (i.e., **phenotypic complexity**). Collapsing complex, multidimensional psychiatric phenotypes (e.g., depression) into unitary scores has the potential to obscure biologically and clinically important sources of variance (e.g., anhedonia vs. guilt)[39]. Binary diagnostic labels create similar problems. Apart from multidimensionality, psychiatric phenotypes may also

exhibit a complex hierarchical structure[40]. An example of this hierarchical organization is the Hierarchical Taxonomy of Psychopathology (HiTOP) (**Box 3** and *Figure* 1). At the top of the hierarchy is the *p*-factor, a broad transdiagnostic liability to all forms of psychopathology[41]. Situated below the *p*-factor are narrower dimensions—internalizing, thought disorders, disinhibited externalizing, and antagonistic externalizing—specific to particular domains of psychopathology[42]. Each of these dimensions, in turn, subsumes still narrower symptom dimensions (e.g., fear, distress, substance use). Too often, simple summary scores ignore this structure, combining broad and narrow sources of variance[43], leading to attenuation of biology-psychopathology associations).

We show in example 1 of the supplementary material how failing to differentiate these multidimensional and hierarchical sources of variance from each other can confound relations with biological parameters. We provide an illustration of these concepts using Child Behaviour Checklist (CBCL) data from the ABCD study, which exhibits both multidimensionality and hierarchical structure. First, the CBCL is a multidimensional instrument that measures eight empirical syndromes using eight distinct subscales. Second, the CBCL has a hierarchical structure with variance attributable to three levels: 1) a *p*-factor; 2) internalizing and externalizing dimensions; and 3) the eight specific psychopathology syndromes. We used a bifactor model[44] within a structural equation modeling framework (see **Box 4**) to separate these dimensions into three orthogonal (i.e., uncorrelated) variance components and examined how much variance was unique to each level. The CBCL has three composite scales: 1) Total Problems, which summarizes the scores across the eight syndrome scales; 2) Internalizing Problems, which summarizes scores across the three internalizing scales; and 3) Externalizing Problems, which summarizes scores across the two externalizing scales. Less than 49% of the total variance is common across the eight scales, such that collapsing measurement of psychopathology into the unidimensional Total Problems score

misrepresents the data and would result in attenuation of biology-psychopathology associations unique to the *p*-factor by 30.2% (i.e., r_{xx} = .488), even assuming perfect reliability of the biological measure. This is despite the Total Problems score showing high reliability in terms of Cronbach's alpha (α = .949). However, it is possible for internal consistency reliability to be high in the presence of multidimensionality, such that reliability cannot be used as a **unidimensionality** statistic.

Results are worse for the other two composite scales, Internalizing Problems and Externalizing Problems, where variance uniquely attributable to these group dimensions is only 10.4% and 20.1%, resulting in a 67.8% and 55.2% attenuation of correlation coefficients with external variables, respectively, (i.e., r_{xx} = .104 and .201). We also demonstrate that high phenotypic complexity across the eight empirical syndrome scales due to the hierarchical organization leads to low internal consistency reliability for these individual scales (i.e., average ~42% variance is unique to each scale). This low reliability results in substantial attenuation bias, with correlations between symptoms and biological criterion variables being reduced from between 15% (r_{xx} = .721 for Somatic Complaints) to 48.2% (r_{xx} = .232 for the Anxious/Depressed scale).

4. Inadequate phenotypic resolution

The vast majority of biology-psychopathology association studies implicitly assume that measurement precision is uniform across the **latent trait** continuum, a concept referred to as **phenotypic resolution**[40]. Yet most measured psychiatric phenotypes lack sufficient coverage of the adaptive (low) end of the continuum, leading to differential phenotypic resolution across the range of the scale[45]. Consider anxiety. Low scores on a clinical scale are meant to represent the absence of pathological anxiety, but often there is little to no item content addressing the opposite end of the latent trait continuum. As a result, there will be

high error at the low end of the scale, making it difficult to conduct robust individual differences research. This problem is known as a 'multiplicative error-in-variable model,' in which the error is proportional to the distributional properties of the signal[33]. **Attenuation bias** will thus be present for participants who score at the lower end of the psychopathology continuum, which tends to be most individuals, particularly in studies of community-dwelling, non-clinical populations. The multiplicative error-in-variable model also results in marked heteroscedasticity (i.e., the distribution of the residuals or error terms in a regression analyses is unequal across different values of the measured values), which reduces statistical power[46].

Phenotypic resolution can be examined using item response theory (IRT) (**Box 4**). IRT provides total information functions, which plot the measurement precision of a phenotype as a function of the standardized latent trait distribution[47]. Typically, for unipolar psychiatric phenotypes, reliability is unacceptably low ($r_{xx} < .6$) below the mean[48]. Because reliability places an upper bound on the association with other variables[49], this decrease in measurement precision can markedly decrease **signal-to-noise ratio** in biology-psychopathology association studies.

In example 2 of the supplementary material, we provide an illustrative example of poor phenotypic resolution using CBCL data from the ABCD study, with results demonstrating that only a small portion of the sample have reliable scores for most of the CBCL scales. Specifically, we find unacceptably low reliability, even for basic research purposes ($r_{xx} < .6$), at or below one standard deviation below the mean for ten of the eleven scales (i.e., all scales except the Total Problems scale). The average proportion across CBCL scales of the ABCD sample that would not have interpretable scores due to low phenotypic resolution was 37.2%; more than half of the sample had uninterpretable scores for three of the

eleven CBCL scales. Thus, despite the promise of the ABCD study for providing a sample size sufficient to accurately assess biology-psychopathology associations, a large proportion of participants from the ABCD study have CBCL scores with unacceptably low reliability, which will have the unfortunate and counterproductive goal of attenuating biology-psychopathology associations.

5. Measurement Non-Invariance

Another challenge to the accurate assessment of biology-psychopathology associations is the assumption that a measure assesses a psychiatric construct similarly across groups and measurement occasions (i.e., **measurement invariance)**[50]. Yet there is ample evidence that measurement properties can vary (i.e., non-invariance) across demographic groups (e.g., sex) or unobserved/latent classes (i.e., homogeneous subpopulations/subgroups, clusters, or mixtures, embedded within the sample)[51]. Non-invariance can substantially bias results, because raw scores do not have the same substantive interpretation across groups. For example, a raw score of 10 on a particular scale may not correspond to the same level of psychopathology in males and females.

Invariance testing provides a rigorous means of evaluating the equivalence of model parameters across groups by imposing a series of increasingly restrictive equality constraints on the model parameter estimates within a factor analytic framework[50]. Typically, three levels of invariance are evaluated: 1) weak invariance; 2) strong invariance; and 3) strict invariance (**Table S3** of supplementary material for technical definitions)[50]. Unfortunately, only a small percentage of studies test for full measurement invariance[50]; thus, the combining of raw scores across discrete groups (e.g., sex, ethnicity, etc.) for biology-psychopathology associations remains problematic. In example 3 of the supplementary material, we provide a striking example of measurement non-invariance of the CBCL Total Problems scale (which is

the most reliable scale of the CBCL)[52] between male and female ABCD participants. Results demonstrate that CBCL raw scores are not comparable between male and female children at any point along the **latent trait** continuum. Thus, any study that pools the results on the CBCL Total Problems scale for male and female children and tests the association with biological variables will draw erroneous conclusions.

6. The heterogeneity problem

The **heterogeneity problem** is increasingly recognized as a key challenge for biological studies of psychiatric illness[34]. Heterogeneity can be described at person-centered and variable-centered levels[34]. Person-centered heterogeneity refers to the presence of clusters or subtypes within groups, such as a group of individuals diagnosed with major depression. To the extent that such clusters or subtypes are unrecognized and associated with distinct biological signatures, they will attenuate biology-psychopathology associations (i.e., mixing apples and oranges). This problem is exacerbated in case-control research because traditional DSM/ICD diagnoses likely encompass phenomenologically, etiologically, and biologically heterogeneous syndromes (**Box 2**). The result is the so-called 'jingle fallacy', in which divergent phenomena are arbitrarily equated, in this case because of the application of a common term[53]. Variable-centered heterogeneity describes admixtures of symptoms with divergent etiology, pathophysiology, course and/or treatment response[54] or failure to differentiate between narrower homogeneous and unidimensional symptom components.

Both person-centered and variable-centered heterogeneity have emerged as a critical issue in depression research. For example, an analysis of 3,703 participants in a clinical trial for the treatment of depression revealed a remarkable degree of person-centered disorder heterogeneity with 1,030 unique symptom profiles identified using the Quick Inventory of Depressive Symptoms (QIDS-16), 864 (83.9%) of which were endorsed by five or fewer

participants and 501 (48.6%) were endorsed by only one participant[55]. Thus, methodologies that explicitly accommodate potential clinical sample heterogeneity are a promising way forward in psychiatric research[56]. There is also evidence of variable-centered heterogeneity in depression, which has a clear multifactorial structure despite often being treated as a unitary construct based on sum scores on inventories, such as the Hamilton Rating Scale for Depression[57]. Indeed, Kendler et al. (2013) identified three distinct genetic factors that explained the co-occurrence of distinct subsets of DSM criteria/symptoms: cognitive and psychomotor symptoms, and mood and neurovegetative symptoms[58]. Heterogeneity has also been identified across depression symptoms in terms of etiology, risk factors, and impact on functioning[57]. These findings suggest that the analysis of narrower homogenous and unidimensional symptom components or even individual symptoms is likely to be a more informative and productive avenue for future biology-psychopathology association studies.

7. Method bias

Method bias is a common, yet often neglected, potential source of measurement error in biology-psychopathology association studies. Sources of method bias include response styles commonly encountered in self-report, such as social desirability (i.e., responses attributable to the desire to appear socially acceptable), acquiescence ('yea-saying'), disaquiescence ('nay-saying'), extreme (selecting extreme response categories in Likert-type ordinal scales), and midpoint (selecting middle categories in Likert-type ordinal scales) response styles[59]. Method bias can distort dimensional structure, obscure true relations between constructs, and compromise validity[60,]. Method bias is caused by method factors, which describe sources of systematic measurement error that contribute to an individual's observed score, thus attenuating subsequent analyses of association[60]. Indeed, method biases are one of the most important sources of measurement error[59]. Between one-fifth and one-third (18-32%) of the variance in self-report measures is attributable to method factors[60].

Method factors and the resulting method bias represent serious threats to study validity, because as systematic sources of error variance they attenuate and otherwise distort the empirical relationship between variables of interest[59].

Recommendations for precision psychiatric phenotyping

In this section, we outline some recommendations for enhancing the precision of psychiatric phenotyping and, ultimately, increasing the robustness and reproducibility of biology-psychopathology association studies (**Table 1** and *Figure 1*).

1. Dimensional Sampling and Measurement

To overcome the limitations of categorical nosological systems, some have advocated for studying dimensional phenotypes that cut across traditional diagnostic categories, a view that closely aligns with the National Institute of Mental Health (NIMH) Research Domain Criteria (RDoC)[2] initiative. Psychometrically, mental disorders show a dimensional rather than a taxonomic structure[61] and dimensional measures of psychopathology exhibit greater reliability and validity than categorical diagnoses[23]. Indeed, the highly polygenic architecture of many psychopathology phenotypes implies that they are dimensionally distributed quantitative traits[62]. Greater statistical power can be further achieved in biological studies through a dimensional enhancement strategy, involving the recruitment of participants with subthreshold and non-clinical levels of symptoms to leverage symptom variation across the full spectrum of severity[63]. The chances of sampling bias and clinical heterogeneity will be reduced, and effect size estimates will be less biased, with dimensional (vs. case-control study) designs[27]. Dimensional sampling strategies are potentially more economical than case-control sampling, as dimensional designs do not rely on thorough clinical pre-screening of participants prior to their inclusion in the study[64]. Dimensional sampling is also more likely to yield samples more representative of the population than case-control sampling, as

dimensional sampling does not exclude individuals based on arbitrary clinical cut-offs and hierarchical exclusion rules[43]. However, to ensure sampling of the full spectrum of symptom or syndrome severity, participants likely to have elevated levels of the target psychopathology dimensions can be over-sampled.

2. Deep Phenotyping and Use of Standardized Measures

Existing large-scale databases—such as the UK Biobank[65]—have a large number of participants who completed an array of measures. However, a limitation of these databases is minimal phenotyping of specific psychopathology phenotypes[32]. To address problems of minimal and inconsistent phenotyping, we recommend comprehensive assessment using a **deep phenotyping** approach with standardized psychopathology measures that can be widely adopted (e.g., **Box 3**), and which are better suited for data pooling via established psychiatric research consortia (e.g. ENIGMA, PGC)[32]. Broadband assessment of multiple dimensions of psychopathology should be undertaken due to the highly comorbid nature of mental health problems[64]. An advantage of deep phenotyping is that it enables the identification and accommodation of comorbidity, as well as person-centered and variable-centered heterogeneity. Deep phenotyping also facilitates greater comparability across studies and the potential harmonization of datasets. Examples of deep phenotyping can be found in existing cohorts[30,31].

3. Use of Homogenous Unidimensional Scales and Hierarchical Modeling

Construct homogeneity and **unidimensionality** are important qualities of scales used to assess psychopathology and enable researchers to isolate the specific sources of variance associated with biological measures[66]. Relatedly, due to the potential empirical overlap of symptom components/empirical syndromes at low levels of the psychopathology hierarchy, it is important that the measures chosen assess homogeneous components with

high discriminant validity to avoid redundancy[43]. We thus advocate for a 'splitting' approach in which psychopathological constructs are dissected into finer-grained, lower-order homogenous constructs to isolate specific variance while taking account of the hierarchical organization of these phenotypes[67]. The study conducted by Tiego et al. (2022)[68] provides an example of a splitting approach that identified significant associations between polygenic risk for schizophrenia and psychometric measures of schizotypy in a non-clinical sample that were otherwise obscured by the use of raw scores or a 'lumping approach'.

Unidimensionality of a construct can be evaluated using factor analysis within a structural equation modeling framework (**Box 4**).

Table 1

Relation of Sources of Imprecision in Psychopathology Phenotyping to Proposed Solutions

Problem	Solution
1. sampling bias	dimensional sampling and measurement
2. minimal and inconsistent phenotyping	deep phenotyping and use of standardized measures
3. phenotypic complexity	Use of homogenous unidimensional scales, test for multidimensionality and model hierarchical relations between dimensional constructs
4. poor phenotypic resolution	increasing phenotypic resolution by adding items assessing the adaptive end of the continuum
5. measurement non-invariance	testing for and accommodating measurement invariance
6. the heterogeneity problem:	
i) person-centered heterogeneity	i) mixture modeling
ii) variable-centered heterogeneity	ii) broadband assessment of psychopathology and hierarchical modeling
7. method bias	multi-method assessment

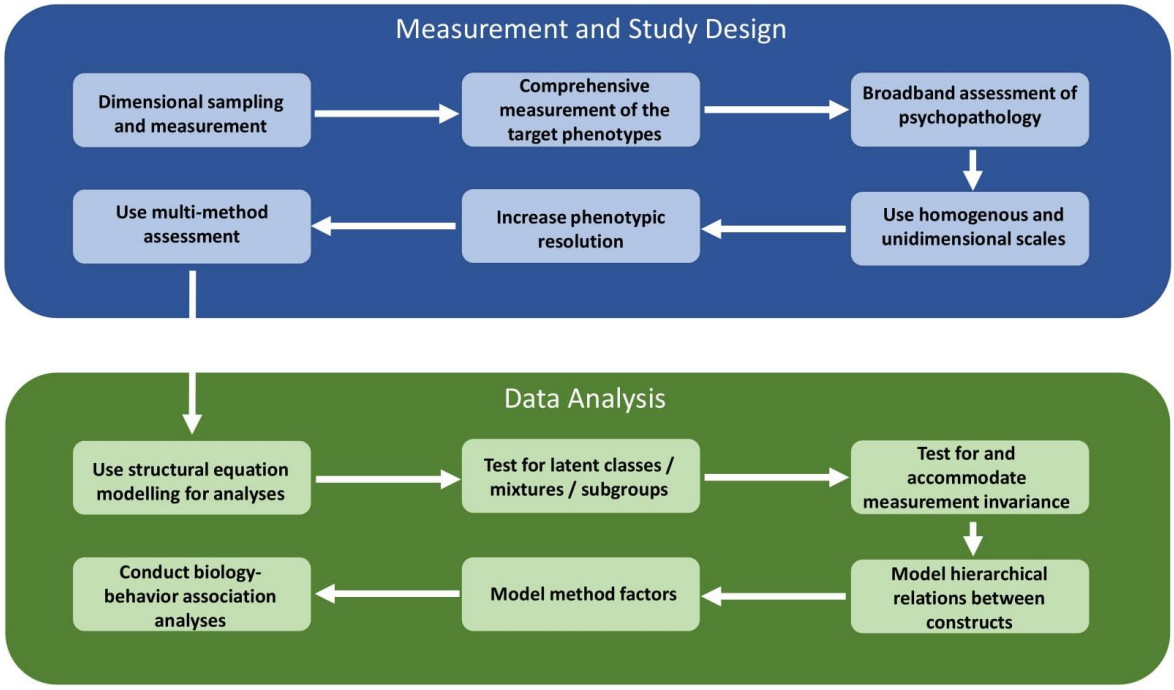

Figure 3. Example workflow for a precision psychiatric phenotyping approach in the context of a biology-psychopathology association study.

Psychiatric symptoms are intrinsically hierarchical. Even homogenous scales typically contain sources of variance spanning multiple levels of the hierarchy[43]. Failure to account for this structure leads to measurement contamination, and reduced reliability and validity for investigating biological associations (cf. Supplementary Example 1). Phenotypic complexity, multidimensionality, the **heterogeneity problem**, and the **comorbidity problem** can all be addressed via hierarchical modeling. There are two approaches to modeling the hierarchical structure of psychopathology: *bottom-up* and *top-down*. Bottom-up approaches leverage higher-order factor models and confirmatory factor analysis within a structural equation modeling framework (**Box 4**), with narrower psychiatric syndromes modeled at the first stage and broader spectra modeled at the second (for a tutorial, see Ref. [69]). Using a bifactor model, hierarchical sources of variance can be partitioned into a common factor (e.g., *p*-factor) and

orthogonal specific factors (e.g., internalizing, externalizing; see Supplementary Example 1 for a detailed illustration)[44]. An alternative bottom-up approach uses hierarchical clustering, where questionnaire items or subscales are organized into homogenous clusters based on shared features[70].

The top-down approach is the bass-ackwards method[71]. The bass-ackwards method is useful for explicating complex hierarchical structures top-down and involves extracting an increasing number of orthogonal principal components to represent the major dimensions in a multi-level hierarchy. The first unrotated principal component captures covariance amongst items or subscales from psychopathology questionnaires at the broadest level. In the second iteration of the method, two orthogonally rotated principal components are extracted; followed by three at the next iteration and so on. Component correlations are calculated between adjacent levels to evaluate continuity versus differentiation of psychopathology components. Proceeding further down the hierarchy, the covariance structure becomes differentiated into dimensions that are increasingly narrow conceptually and empirically, until distinct behavioral syndromes / symptom constellations are isolated. An example of the bass-ackwards method in the ABCD data is provided by Michelini et al. (2019)[72].

4. Increasing Phenotypic Resolution

To address the issue of low phenotypic resolution, items can be carefully selected within an item response theory (IRT) framework (**Box 5**) so that they assay psychopathological severity across the full length of the latent-trait continuum, offering psychometric precision at all levels of the measured construct[40]. Alternatively, it is possible to select measures that have already been optimized within an IRT framework to increase measurement precision across the entire latent-trait continuum (e.g., the Computerized adaptive assessment of personality disorder; CAT-PD[73]). For unipolar traits, it is possible to

bolster measurement precision with items from a related construct that represents the opposite (i.e., adaptive) end of the continuum[74]. We demonstrate the utility of this approach in example 4 of the supplementary material, where we bolster the lower end of the CBCL Attention Problems latent trait continuum by pooling the items from this scale with items taken from the Early Adolescent Temperament Questionnaire – Revised (EATQ-R)[17] Effortful Control subscale, which measures the adaptive end of the attentional control/attentional problems continuum.

5. Address Measurement Non-Invariance

Measurement invariance should be thoroughly evaluated across groups, including sex/gender, race/ethnicity, and developmental stage. There are multiple resources for invariance testing, including analytic flow charts and checklists[50]. Differential item function (DIF) testing within an IRT framework provides a powerful approach to invariance testing, but requires larger sample sizes and involves more restrictive assumptions[75]. Where full invariance does not hold, partial invariance can be considered by freely estimating one or more model parameters in the comparison group[76]. Alternatively, researchers can utilize Bayesian approximate invariance testing, which is useful when there are many small, trivial differences between group parameters of no substantive interest, but which in combination result in poor model fit[76]. Groups or subsamples with partial non-invariance of their model parameters can still be meaningfully compared in some circumstances[76]. Measurement non-invariance can be accommodated in several ways. Groups or subsamples with fully non-invariant measurement parameters for psychiatric phenotypes should be analyzed separately. It is also possible to circumvent issues of measurement non-equivalence within both factor analytic and IRT frameworks by removing items identified as having non-invariant factor loadings or intercepts, or slope and threshold parameters, to ensure the equivalence of the latent variable across groups. However, in these instances researchers should be aware of

changing the substantive interpretation of the construct by narrowing its scope and breadth (i.e., attenuation paradox).

6. Mixture modeling

In contrast to situations where subgroups are easily identified and differentiated based on manifest, discrete characteristics such as sex and ethnicity, there are situations where subgroups embedded within the data are not directly observed, resulting in person-centered heterogeneity. Thus, prior to conducting biology-behavior association studies, it is important to verify that the psychiatric phenotypes can be treated as continuous dimensions in the sample. Mixture modeling provides a useful approach for investigating person-centered heterogeneity[77]. Mixture modeling is a particularly promising approach because it can identify latent classes or clinical subtypes, which often characterize psychopathology phenotypes[77]. Entropy provides a summary measure of classification accuracy of participants based on the posterior probabilities of class membership within a mixture modeling analysis. It can range between 0.00 to 1.00, with higher entropy indicating better classification accuracy. When entropy is high (e.g., $\geq$.80) class membership can be used as a discrete categorical variable for subsequent analyses to compare results between classes. However, where entropy is low, classes must be compared using alternative analytic approaches that take into account the probabilistic nature of class membership. By identifying and analyzing subtypes, the confounding impact of sample heterogeneity on studies of the associations between biology and psychopathology can be reduced[34]. In example 5 of supplementary material, we apply mixture modeling to the Attention Problems CBCL scale, using data from the ABCD 2-year follow-up. Results reveal evidence for two latent classes with different empirical distributions and item response profiles on the CBCL. These observations suggest that failure to account for the latent categorical structure of the Attention Problems scale could lead to erroneous results in biology-psychopathology association studies.

7. Multimethod assessment

A fundamental tenet of psychometrics is that measurement of a psychological attribute represents a trait-method unit, combining a person's true score with systematic measurement error related to the assessment method[66]. Thus, at least two different assessment methods are required to differentiate the true score for a trait measure from method effects[78]. The recommended approach to circumventing issues of method bias is to use multimethod assessment and then implement statistical remedies to identify and exclude the method factors and decompose an observed score into true score, method variance (systematic error), and random measurement error[60,78]. The optimal statistical method for removing method variance is the trait method minus one [T(M-1)] model estimated within a structural equation modeling framework (**Box 4**)[79].

In example 6 in the supplementary material, we apply the T(M-1) method to the new composite scale we constructed in example 4, which combined CBCL Attention Problems scale items and the EATQ-R Effortful Control subscale items of the ABCD data. The purpose of applying the T(M-1) model was to control for method variance associated with subjective report by the primary caregivers and in doing so increase signal-to-noise ratio. To do so, we incorporated neurocognitive measures of the target attention problems construct, specifically stop signal reaction time from the stop signal task and d-prime as an estimate of working memory from the n-back task, both of which are well-established endophenotypes of ADHD[80,81]. We were then able to specify the neurocognitive measures as the reference method, such that loadings from the CBCL and EATQ-R caregiver report items on the target Attention Problems factor captured only that variance shared with the neurocognitive measures. A methods factor captured the residual variance in subjective report by the primary caregivers that was unique to these measures[79]. We found that the Attention Problems factor was associated with polygenic risk for ADHD. By contrast, the methods factor that captured

variance specific to caregiver-report measures of attention problems and attention control abilities was not related to polygenic risk for ADHD (*Figure S27*). Thus, the T(M-1) model yielded a genetic association that was otherwise obscured by standard analyses.

Conclusions

It has been suggested that large, consortia-sized samples are necessary to discover robust and reproducible biology-psychopathology associations. Yet larger sample sizes are not sufficient to resolve the issues introduced by imprecise or otherwise suboptimal psychiatric phenotypes. As a field, we must first improve our measurement techniques. We recommended broadband, transdiagnostic assessment of hierarchically-organized, unidimensional, and homogenous psychopathology dimensions across the full range of the severity spectrum. We encourage greater focus on deep phenotyping, measurement invariance, phenotypic resolution and person-centered and variable-centered heterogeneity. A voluminous psychometrics literature—and the worked examples featured in this review—make clear that this multi-faceted strategy will increase validity, reliability, effect sizes, statistical power, and ultimately replicability.

Box 1 – The relationship between measurement reliability and observed effect size

The relation between measurement reliability and the observed effect size[20] is pertinent to many fields of research. Here, we discuss the issue in relation to psychiatric phenotypes in the context of associations with neurobiology and/or genetics. Constraints on the precision with which psychological attributes can be measured are captured by true score theory (also known as classical test theory), according to which variance reflecting a psychological measurement includes a stable component that reflects a person's 'true score' and measurement error[82]:

$$\sigma^2_{observed} = \sigma^2_{true} + \sigma^2_{error} \quad (1)$$

Thus, according to true score theory, all psychological measurement incorporates measurement error (i.e., 'error-in-variables model'[49]), such that our measurements contain some unknown quantity of measurement error, some of which reflects: 1) systematic error attributable to other sources of variance that are not of substantive interest, for example **method bias**; and 2) random measurement error, for example attributable to poor psychometric properties of a measure of a psychopathology phenotype[83]. Measurement error attenuates associations between variables[49]. This bias is intuitively demonstrated with respect to the Pearson coefficient of product-moment correlation (*r*), which forms the basis of many analyses conducted in the literature on biology-psychopathology associations and can be used as an estimate of effect size. It has been demonstrated that the correlation coefficient, *r*, which is the sample realization of the population parameter rho (ρ), is always a biased estimate of the true association between two variables, x and y:[49]

$$r_{ox,oy} = r_{tx,ty}\sqrt{(r_{xx}r_{yy})} \quad (2)$$

where $r_{ox,oy}$ is the observed correlation, $r_{tx,ty}$ is the true correlation, and r_{yy} and r_{xx} are the reliability coefficients for variables *x* and *y*.

In most cases, the measurement error will be uncorrelated between the variables, resulting in greater dispersion in the data and an **attenuation bias** of the correlation coefficient and, by extension, smaller and less accurate **effect sizes**[38,49]. An additional consequence of greater dispersion in the data and less clustering around the true effect is increased sampling variability, particularly at smaller sample sizes. Relatedly, the standard errors (*SE*) for the correlation coefficient increase as a function of smaller samples, n, and smaller effect sizes, resulting in reduced **efficiency** and precision of estimation[84].

$$SE_r = \sqrt{\frac{1- r^2}{n-2}} \quad (3)$$

Since the probability value of the correlation coefficient is based on the distribution of Student's *t* with *n* – 2 degrees of freedom ($t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$), smaller effect sizes, as well as smaller samples lead to lower statistical power. These issues are especially pertinent to measuring psychopathology phenotypes in **biomarker** research and, critically, will not be resolved simply by increasing sample sizes[38]. Assuming sample homogeneity (i.e., the **homogeneity assumption**), increased sample sizes will only reduce sampling variability ($\sqrt{N}$) but not proportionally decrease measurement error. The estimates themselves will remain downwardly **biased** if measurement error is present. Finally, inasmuch as the resulting sample statistic fails to converge on the correct population parameter, it is less likely to be replicated in subsequent samples[21].

Box 2 – Limitations of traditional approaches to psychiatric nosology

Existing diagnostic systems, such as the *DSM-5* and the *ICD-11* have clinical utility, facilitating treatment and communication between mental health professionals and consumers of mental health services[85]. However, the psychopathological concepts invoked by modern nosology may have a tenuous relation with biological correlates, undermining our attempts to link measurement of phenotypes with **biomarkers**[3]. The limitations of such nosological schemes for informing our understanding of the biology of mental disorders have long been recognized. Initially developed to capture psychiatric signs and symptoms without detailed consideration of etiology or pathophysiology[3], diagnostic criteria have since been reified as reflecting, rather than merely indexing, the natural phenomenology of the proposed disease entities themselves, resulting in a conflation of diagnostic criteria with the proposed underlying disorder[86]. Philosophically, the field has fallen prey to the question-begging fallacy, in which diagnostic categories are investigated as if they are real entities without first asking whether the categories are valid in the first place.

The limitations of traditional nosologies introduce a substantial source of phenotypic imprecision due to questionable validity. Problematically, current diagnostic systems define mental disorders as **polythetic-categorical constructs**. Prototypical symptoms occurring in pre-specified number and combination are conceptualized as forming discrete taxa, underpinning binary diagnostic decisions. However, it is known that mental disorders have a dimensional rather than a taxonomic structure[61], with frequency and severity of symptoms extending as a continuum from the clinical to the subclinical and into the non-clinical range. A related issue is that individuals are generally diagnosed using hierarchical exclusion rules in diagnostic checklists, by which comorbid conditions may be ruled out based on meeting criteria for another disorder. These factors can lead to artificial ‘prototypical cases’ with elevated symptoms and no comorbidity, as well as distort the covariance structure of the data,

which can impact subsequent analyses[87]. Additionally, focusing on a particular diagnostic category assumes homogeneity of symptoms and mechanisms (i.e., the **homogeneity assumption**), but individuals with the same diagnosis may exhibit little to no overlap in symptoms (i.e., the **heterogeneity problem**)[34]. Co-morbidity between putatively distinct disorders (i.e., **the comorbidity problem**)[88], and issues of arbitrary clinical cut-offs and the ignoring of the clinical significance of subthreshold symptomatology are well-documented limitations of current psychiatric taxonomies[89]. These limitations obfuscate the search for the neurobiological correlates of psychiatric symptoms and constitute an impediment to future research in this domain[90].

Box 3 The Hierarchical Taxonomy of Psychopathology

The Hierarchical Taxonomy of Psychopathology (HiTOP) model is a potentially useful framework for precision psychiatric phenotyping. HiTOP is a data-driven approach to psychiatric nosology that organizes symptoms into homogenous, hierarchically-organized dimensions (*Figure 1*)[42]. The problem of arbitrary diagnostic thresholds, subthreshold/subclinical symptomatology, and low power is addressed by measuring psychopathology continuously with no artificial demarcation point designating health from disorder[42]. The comorbidity problem and heterogeneity problem are addressed by organizing co-occurring problems into homogenous dimensions[42]. For example, the high comorbidity of major depressive disorder and generalized anxiety disorder are seen to reflect the operation of common etiological mechanisms, which are captured by the Distress subfactor and more broadly situated under the Internalizing spectrum within the HiTOP model.

The development of an omnibus measure of the HiTOP model is nearing completion and will be open-source and freely available for use without charge in both computerized and paper-and-pencil formats[91]. In the meantime, several existing instruments can be used to reliably assess HiTOP dimensions in youth and adults (https://hitop.unt.edu/clinical-tools/hitop-friendly-measures). HiTOP-conformant measures enable broadband, transdiagnostic assessment of psychopathology at multiple levels of the hierarchy, from broad superspectra dysfunction and spectra to narrower subfactors and empirical syndromes. HiTOP-conformant measures focus on narrow homogenous and unidimensional constructs with high discriminant validity facilitating high reliability and valid inference[43,66] for association studies with biology. At the lowest levels of the hierarchy, HiTOP encompasses even narrower symptom components with (e.g., anhedonia, insomnia) and maladaptive traits[42]. The latter provides a measure of the lower range and adaptive end of the psychopathology continuum. Combining measures of traits and psychopathology thus

improves phenotypic resolution across the full spectrum. Notably, the higher-order spectra of the HiTOP model are invariant across sexes and different age groups[92]. HiTOP dimensions, including the broad superspectra and spectra, as well as narrower subfactors and symptom components, can serve as phenotypic targets for neuroscience-informed RDoC domains[93].

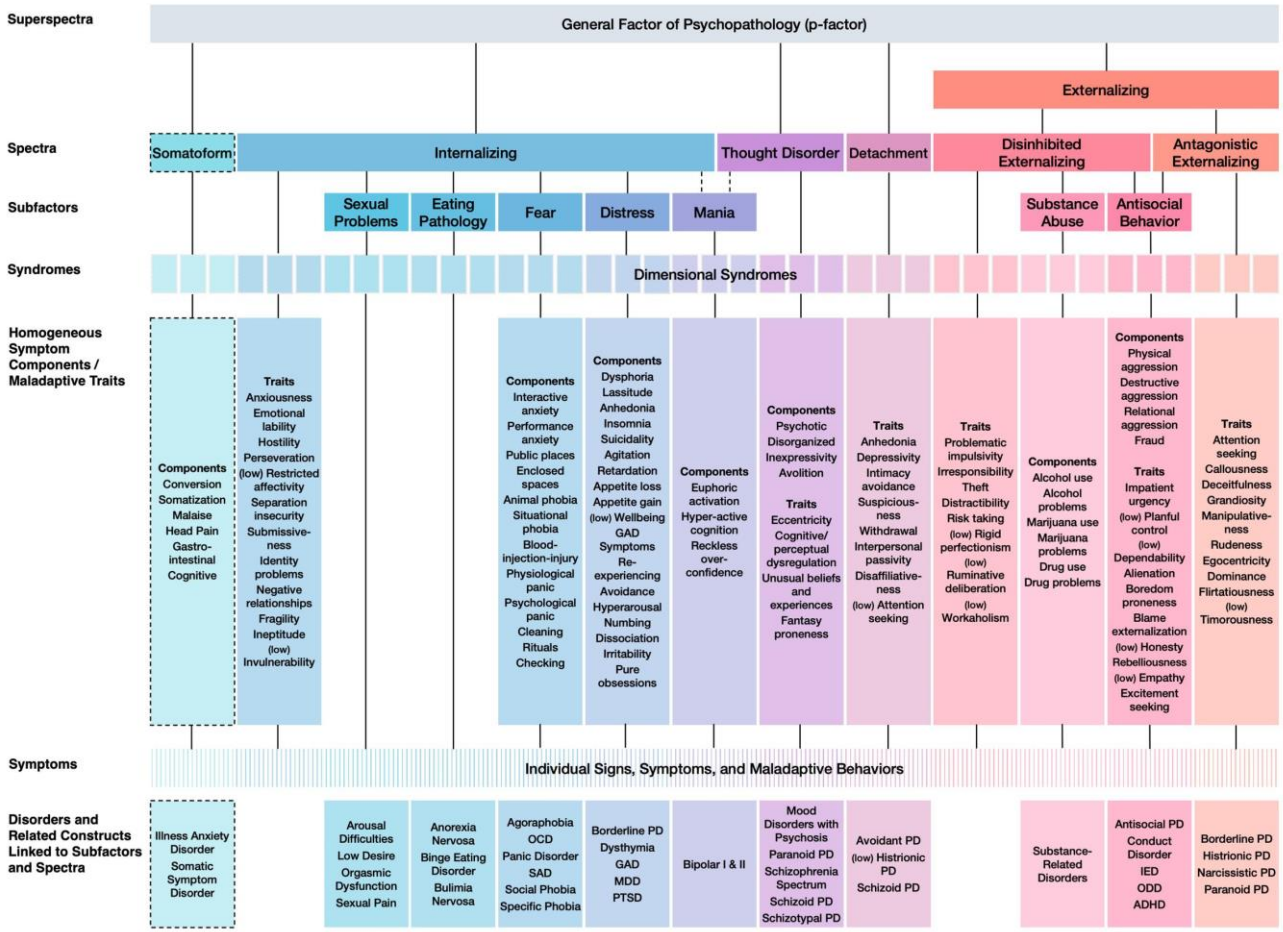

Figure 1. The Hierarchical Taxonomy of Psychopathology (HiTOP) model. https://osf.io/5y9te

The broadest dimensions, reflecting common liabilities to psychopathology, are situated at the top of the hierarchy with the narrowest traits and symptom components situated at the bottom, reflecting liabilities to specific problems.

Box 4. Structural equation modeling (SEM)

Hierarchical modeling, measurement invariance, mixture modeling, and the T(M-1) model can be done within a structural equation modeling framework (SEM). SEM is a statistical technique that combines factor analysis, canonical correlation, and multiple regression[94]. SEM can be used to extract the common variance from **factor indicators** of the construct of interest. The resulting factor, also known as a latent variable, is a purer measure of the construct of interest because only variance common to all variables that reflect the dimension of interest are included as shared variance[94]. In the common factor model estimated within the SEM framework, reflective latent variables (i.e., an underlying factor is conceptualized as causing the covariance in the indicators) are estimated by decomposing observed variables into variance shared with the other factor indicators and variance that is unique to the variable (i.e., variance attributable to a separate construct and measurement error). The formula is expressed as:

$$x_i = a_x + \lambda_x \xi_i + \theta \varepsilon_i \quad (4)$$

where x_i is a measured variable (i.e., observed/manifest variable), a_x is an intercept, λ_x is a factor loading determining the influence of a factor ξ_i on the measured variable, and $\theta\varepsilon_i$ is the unique variance/error of the measured variable that is not explained by the factor loading (*Figure 2*). This model formalizes the following: (1) the target psychopathology phenotype is unobserved and must be inferred by one or more measured variables (e.g., questionnaire items); (2) measured variables are imperfect indices of the target construct and incorporate measurement error; (3) factor indicators are not necessarily equally important measures of the target latent variable, as indicated by differences in the strength of the factor loadings (i.e., λx).

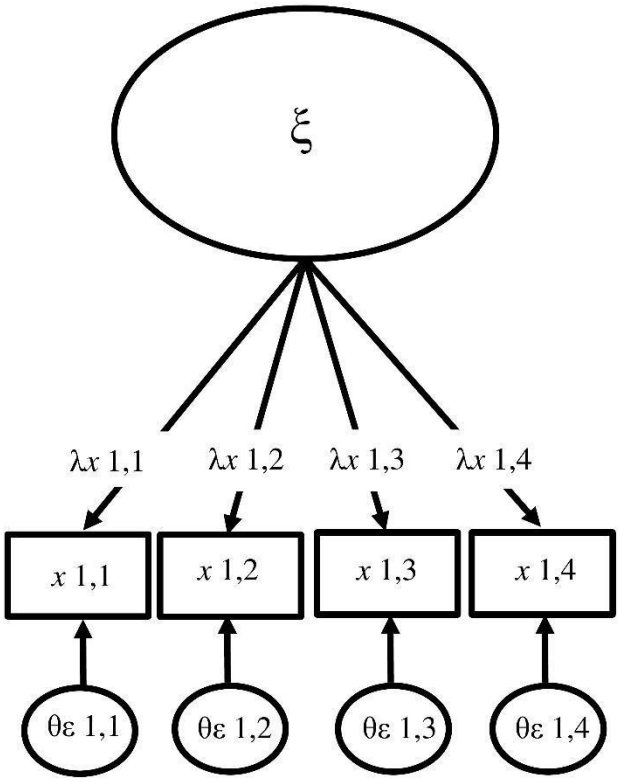

Figure 2. Reflective latent variable (i.e., common factor) model in which the unobserved psychobiological attribute (factor or latent construct; ξ), is conceptualized as explaining the variance/covariance in the measured variables ($x_{1,1} – x_{1,4}$) via their factor loadings ($\lambda x_{1,1}$ - $\lambda x_{1,1}$), which are linear regression coefficients. The indicator error variances (also residual variances or uniquenesses; $\theta\varepsilon_{1,1}$ - $\theta\varepsilon_{1,4}$) capture the variance in each measured variable not explained by the factor (i.e., variance not shared with the other indicator variables).

In a structural regression model, SEM enables estimation of regression path coefficients between factors within the model. Thus, SEM estimates the empirical relations between predictor variables and criterion variables with measurement error excluded from the final model[94]. An additional advantage of using SEM is that hypothesized multiple dependence relations can be examined concurrently, along with complex interactions[94]. By contrast, some researchers use a two-step factor score regression technique in which factor scores estimates are derived from the latent variables as manifest variables and then

incorporated into subsequent regression analyses. It is important to note that factor score estimates are not the same as latent variables due to factor score indeterminacy. In simple terms, factor score indeterminacy reflects the fact that an infinite set of factor scores can be estimated for the same analysis that will be equally consistent with the factor loadings. This is because the number of observed variables is less than the number of common and unique factors to be estimated[95]. The degree of factor score indeterminacy is related to the number of factor indicators and their communalities (i.e., how much variance is explained in the variables by the factor) and is represented by a validity coefficient, which will vary between studies[95]. Factor score estimates can, therefore, misrepresent the rank ordering of individuals along the factor[95]. The degree to which factor score estimates preserve the correlations amongst the factors in the analysis (i.e., correlational accuracy) and are not contaminated by variance from orthogonal factors (i.e., univocality) will also vary between studies[95]. Furthermore, the use of factor score estimates can potentially bias the parameter estimates of the regression models[96]. Thus, we recommend against this approach in favour of SEM.

Ideally, biological measurements should be incorporated directly into latent models to capitalize on the increased measurement precision and statistical power (e.g., Kim et al., 2007[97]). However, it should be noted that SEM is a large sample technique, in general requiring samples of greater than 200[98]. Thus, it may not be feasible for many research studies examining biological variables. Several SEM packages are commercially available, such as Mplus (http://www.statmodel.com/), and freely available as open-source software, such as lavaan in R (https://lavaan.ugent.be/). The HiTOP Consortium provided a primer for conducting SEM research in the context of dimensional hierarchical models of psychopathology[69] and there are several excellent entry-level texts for SEM, such as Kline et a. (2015)[98].

Box 5 – Item Response Theory

Item response theory (IRT) is a sophisticated approach to psychometric scale construction, evaluation, and refinement and has been increasingly recommended for, and applied, in psychopathology research[99]. IRT encapsulates a set of measurement models and statistical methods that can be used to empirically model item level data[99]. The two-parameter logistic (2PL) model for dichotomous item response data and its extension for polytomous item response data, the graded response (GR) model, are the most commonly used models[45,100]. Two main parameters of interest are generated through IRT analysis: 1) a slope (also 'discrimination') parameter (α); and 2) a threshold (also severity or location) parameter (β). Slope parameters are akin to factor loadings and indicate how well an item measures the latent trait. They are measured in a logistic metric, generally ranging between $\pm$ 2.8, with higher values indicating that an item is more discriminating between different levels of a latent trait[99]. Threshold parameters indicate the location on the latent trait continuum where an item is most sensitive to different levels of the latent trait. They are measured in a standardized metric (i.e., $M = 0$, $SD = 1$) generally ranging between $\pm$ 3, with more extreme values indicating that an item is sensitive to lower and higher levels of symptom severity[99]. These item-level parameters enable the amount of measurement precision, or '*information*', to be quantified. Item information is additive and can be combined to represent the total measurement precision of items across the latent trait continuum[47]. Information (*I*) can then be transformed into a standard metric of internal consistency reliability $\left[r_{xx} = 1 - \left(\frac{1}{I}\right)\right]$ [100]. Items can thus be carefully selected to optimize measurement precision across the whole latent trait continuum. Furthermore, items with high local dependence (i.e., correlated residual variances) can be identified as redundant and removed. Despite the appeal of IRT for optimizing phenotypic precision in psychopathology research it has not been utilized widely for identifying associations between psychometric constructs and biological measures.

Acknowledgements

J.T. was supported the Tuner Impact Fellowship from the Turner Institute for Brain and Mental Health, Monash University Australia. AF is supported by the Sylvia and Charles Viertel Foundation, the National Health and Medical Research Council (IDs: 1146292 & 1197431) and the Australian Research Council (ID: DP200103509). AJS was supported by the National Institute of Mental Health (IDs: MH131264 & MH121409) and University of Maryland. The following members of the HiTOP Consortium read and endorsed the current manuscript: Rany Abend, Natalie Goulter, Nicholas Eaton, Antonia N. Kaczkurkin, Robin Nusslock.

References

1 Perkins, E. R., Latzman, R. D. & Patrick, C. J. Interfacing neural constructs with the Hierarchical Taxonomy of Psychopathology: 'Why' and 'how'. *Pers Ment Health* **14**, 106-122, doi:https://doi.org/10.1002/pmh.1460 (2020).

2 Insel, T. *et al.* Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *Am J Psychiatry* **167**, 748-751, doi:https://doi.org/10.1176/appi.ajp.2010.09091379 (2010).

3 Hyman, S. E. Can neuroscience be integrated into the DSM-V? *Nat Rev Neurosci* **8**, 725-732, doi:https://doi.org/10.1038/nrn2218 (2007).

4 Singh, I. & Rose, N. Biomarkers in psychiatry. *Nature* **460**, 202-207, doi:https://doi.org/10.1038/460202a (2009).

5 First, M. B. *et al.* Clinical applications of neuroimaging in psychiatric disorders. *Am J Psychiatry* **175**, 915-916, doi:https://doi.org/10.1176/appi.ajp.2018.1750701 (2018).

6 Marek, S. *et al.* Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654-660, doi:https://doi.org/10.1038/s41586-022-04492-9 (2022).

7 Poldrack, R. A. *et al.* Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci* **18**, 115, doi:https://doi.org/10.1038/nrn.2016.167 (2017).

8 Saggar, M. & Uddin, L. Q. Pushing the boundaries of psychiatric neuroimaging to ground diagnosis in biology. *eneuro* **6**, ENEURO.0384-0319.2019, doi:https://doi.org/10.1523/eneuro.0384-19.2019 (2019).

9 Sha, Z., Wager, T. D., Mechelli, A. & He, Y. Common dysfunction of large-scale neurocognitive networks across psychiatric disorders. *Biol Psychiatry* **85**, 379-388, doi:https://doi.org/10.1016/j.biopsych.2018.11.011 (2019).

10 Smoller, J. W. *et al.* Psychiatric genetics and the structure of psychopathology. *Mol Psychiatry* **24**, 409-420, doi:https://doi.org/10.1038/s41380-017-0010-4 (2019).

11 Nour, M. M., Liu, Y. & Dolan, R. J. Functional neuroimaging in psychiatry and the case for failing better. *Neuron* **110**, 2524-2544, doi:https://doi.org/10.1016/j.neuron.2022.07.005 (2022).

12 Sullivan, P. F. *et al.* Psychiatric genomics: An update and an agenda. *Am J Psychiatry* **175**, 15-27, doi:https://doi.org/10.1176/appi.ajp.2017.17030283 (2018).

13 Kundu, P., Inati, S. J., Evans, J. W., Luh, W.-M. & Bandettini, P. A. Differentiating BOLD and non-BOLD signals in fMRI time series using multi-echo EPI. *NeuroImage* **60**, 1759-1770, doi:https://doi.org/10.1016/j.neuroimage.2011.12.028 (2012).

14 Parkes, L., Fulcher, B., Yücel, M. & Fornito, A. An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. *NeuroImage* **171**, 415-436, doi:https://doi.org/10.1016/j.neuroimage.2017.12.073 (2018).

15 Kong, R. *et al.* Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. *Cereb Cortex* **31**, 4477-4500, doi:https://doi.org/10.1093/cercor/bhab101 (2021).

16 Visscher, P. M. *et al.* 10 years of GWAS discovery: Biology, function, and ranslation. *Am J Hum Genet* **101**, 5-22, doi:https://doi.org/10.1016/j.ajhg.2017.06.005 (2017).

17 Volkow, N. D. *et al.* The conception of the ABCD study: From substance use to a broad NIH collaboration. *Dev Cogn Neurosci* **32**, 4-7, doi:https://doi.org/10.1016/j.dcn.2017.10.002 (2018).

18 Lilienfeld, S. O. The research domain criteria (RDoC): An analysis of methodological and conceptual challenges. *Behav Res Ther* **62**, 129-139, doi:https://doi.org/10.1016/j.brat.2014.07.019 (2014).

19 Xing, X.-X. & Zuo, X.-N. The anatomy of reliability: A must read for future human brain mapping. *Sci Bull* **63**, 1606-1607, doi:https://doi.org/10.1016/j.scib.2018.12.010 (2018).

20 Zuo, X. N., Xu, T. & Milham, M. P. Harnessing reliability for neuroscience research. *Nat Hum Behav* **3**, 768-771, doi:https://doi.org/10.1038/s41562-019-0655-x (2019).

21 Nikolaidis, A. *et al.* Suboptimal phenotypic reliability impedes reproducible human neuroscience. *bioRxiv*, 2022.2007.2022.501193, doi:https://doi.org/10.1101/2022.07.22.501193 (2022).

22 Falk, E. B. *et al.* What is a representative brain? Neuroscience meets population science. *PNAS* **110**, 17615-17622, doi:https://doi.org/10.1073/pnas.1310134110 (2013).

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

24 van der Sluis, S., Posthuma, D., Nivard, M. G., Verhage, M. & Dolan, C. V. Power in GWAS: Lifting the curse of the clinical cut-off. *Mol Psychiatry* **18**, 2-3, doi:https://doi.org/10.1038/mp.2012.65 (2013).

25 Fisher, J. E., Guha, A., Heller, W. & Miller, G. A. Extreme-groups designs in studies of dimensional phenomena: Advantages, caveats, and recommendations. *J Abnorm Psychol* **129**, 14-20, doi:https://doi.org/10.1037/abn0000480 (2020).

26 Angold, A., Costello, E. J., Farmer, E. M. Z., Burns, B. J. & Erkanli, A. Impaired but undiagnosed. *J Am Acad Child Adolesc Psychiatry* **38**, 129-137, doi:https://doi.org/10.1097/00004583-199902000-00011 (1999).

27 Preacher, K. J. in *Extreme groups designs in The encyclopedia of clinical psychology* Vol. 2 (eds R. L. Cautin & S.O. Lilienfeld) 1189-1192 (John Wiley & Sons, Inc., 2015).

28 Dong, H.-M. *et al.* Charting brain growth in tandem with brain templates at school age. *Sci Bull* **65**, 1924-1934, doi:https://doi.org/10.1016/j.scib.2020.07.027 (2020).

29 Peterson, R. E. *et al.* Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589-603, doi:https://doi.org/10.1016/j.cell.2019.08.051 (2019).

30 Liu, S. *et al.* Chinese color nest project : An accelerated longitudinal brain-mind cohort. *Dev Cogn Neurosci* **52**, 101020, doi:https://doi.org/10.1016/j.dcn.2021.101020 (2021).

31 Tobe, R. H. *et al.* A longitudinal resource for studying connectome development and its psychiatric associations during childhood. *Sci Data* **9**, 300, doi:https://doi.org/10.1038/s41597-022-01329-y (2022).

32 Sanchez-Roige, S. & Palmer, A. A. Emerging phenotyping strategies will advance our understanding of psychiatric genetics. *Nat Neurosci* **23**, 475-480, doi:https://doi.org/10.1038/s41593-020-0609-7 (2020).

33 Newson, J. J., Hunter, D. & Thiagarajan, T. C. The heterogeneity of mental health assessment. *Front Psychiatry* **11**, doi:https://doi.org/10.3389/fpsyt.2020.00076 (2020).

34 Feczko, E. *et al.* The heterogeneity problem: Approaches to identify psychiatric subtypes. *TiCS* **23**, 584 - 601, doi:https://doi.org/10.1016/j.tics.2019.03.009 (2019).

35 Schnack, H. G. & Kahn, R. S. Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Front Psychiatry* **7**, 50, doi:https://doi.org/10.3389/fpsyt.2016.00050 (2016).

36 Yang, Z. *et al.* Brain network informed subject community detection in early-onset schizophrenia. *Sci Rep* **4**, 5549, doi:https://doi.org/10.1038/srep05549 (2014).

37 Hodgson, K., McGuffin, P. & Lewis, C. M. Advancing psychiatric genetics through dissecting heterogeneity. *Hum Mol Genet.* **26**, R160-R165, doi:https://doi.org/10.1093/hmg/ddx241 (2017).

38 De Nadai, A. S., Hu, Y. & Thompson, W. K. Data pollution in neuropsychiatry—An under-recognized but critical barrier to research progress *JAMA Psychiatry* **79**, 97-98, doi:https://doi.org/10.1001/jamapsychiatry.2021.2812 (2022).

39 Reise, S. P., Bonifay, W. E. & Haviland, M. G. Scoring and modeling psychological measures in the presence of multidimensionality. *J Pers Assess* **95**, 129-140, doi:https://doi.org/10.1080/00223891.2012.725437 (2013).

40 van der Sluis, S., Verhage, M., Posthuma, D. & Dolan, C. V. Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. *PLOS ONE* **5**, e13929, doi:https://doi.org/10.1371/journal.pone.0013929 (2010).

41 Caspi, A. & Moffitt, T. E. All for one and one for all: Mental disorders in one dimension. *Am J Psychiatry* **AJP in Advance**, 1 - 14, doi:https://doi.org/10.1176/appi.ajp.2018.17121383 (2018).

42 Kotov, R. *et al.* The Hierarchical taxonomy of psychopathology (HiTOP): A quantitative nosology based on consensus of evidence. *Annu Rev Clin Psychol* **17**, 83-108, doi:https://doi.org/10.1146/annurev-clinpsy-081219-093304 (2021).

43 Clark, L. A. & Watson, D. Constructing validity: New developments in creating objective measuring instruments. *Psychol Assess* **31**, 1412-1427, doi:https://doi.org/10.1037/pas0000626 (2019).

44 Reise, S. P. The rediscovery of bifactor measurement models. *Multivariate Behav Res* **47**, 667 - 696, doi:https://doi.org/10.1080/00273171.2012.715555 (2012).

45 Reise, S. P. & Waller, N. G. Item response theory and clinical measurement. *Annu Rev Clin Psychol* **5**, 27-48, doi:https://doi.org/10.1146/annurev.clinpsy.032408.153553 (2009).

46 Rosopa, P. J., Schaffer, M. M. & Schroeder, A. N. Managing heteroscedasticity in general linear models. *Psychol Methods* **18**, 335-351, doi:https://doi.org/10.1037/a0032553 (2013).

47 Thomas, M. L. The value of item response theory in clinical assessment: A review. *Assessment* **18**, 291-307, doi:https://doi.org/10.1177/1073191110374797 (2011).

48 Streiner, D. L. Starting at the beginning: An introduction to coefficient alpha and internal consistency. *J Pers Assess* **80**, 99 103, doi:https://doi.org/10.1207/s15327752jpa8001_18 (2003).

49 Saccenti, E., Hendriks, M. H. W. B. & Smilde, A. K. Corruption of the Pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models. *Sci Rep* **10**, 438, doi:https://doi.org/10.1038/s41598-019-57247-4 (2020).

50 Vandenberg, R. J. & Lance, C. E. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organ Res Methods* **3**, 4-70, doi:https://doi.org/10.1177/109442810031002 (2000).

51 Miettunen, J., Nordstrom, T., Kaakinen, M. & Ahmed, A. O. Latent variable mixture modeling in psychiatric research: A review and application. *Psychol Med* **46**, 457-467, doi:https://doi.org/10.1017/S0033291715002305 (2016).

52 Achenbach, T. M. *The Achenbach System of Empirically Based Assessment (ASEBA): Development, findings, theory, and applications.*, (University of Vermont, Research Center for Children,Youth, & Families., 2009).

53 Kelly, E. L. *Interpretation of educational measurements.* . (World Book, 1927).

54 Fried, E. I. Moving forward: How depression heterogeneity hinders progress in treatment and research. *Expert Rev Neurother* **17**, 423-425, doi:https://doi.org/10.1080/14737175.2017.1307737 (2017).

55 Fried, E. I. & Nesse, R. M. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *J Affect Disord* **172**, 96-102, doi:https://doi.org/10.1016/j.jad.2014.10.010 (2015).

56 Wager, T. D. & Woo, C.-W. Imaging biomarkers and biotypes for depression. *Nature Medicine* **23**, 16-17, doi:10.1038/nm.4264 (2017).

57 Fried, E. I. & Nesse, R. M. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Med* **13**, 72, doi:https://doi.org/10.1186/s12916-015-0325-4 (2015).

58 Kendler, K. S., Aggen, S. H. & Neale, M. C. Evidence for multiple genetic factors underlying DSM-IV criteria for major depression. *JAMA Psychiatry* **70**, 599-607, doi:https://doi.org/10.1001/jamapsychiatry.2013.751 (2013).

59 Podsakoff, P. M., MacKenzie, S. B., Lee, J. & Podsakoff, N. P. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *J Appl Psychol* **88**, 879 - 903, doi:https://doi.org/10.1037/0021-9010.88.5.879 (2003).

60 Podsakoff, P. M., MacKenzie, S. B. & Podsakoff, N. P. Sources of method bias in social science research and recommendations on how to control it. *Annu Rev Psychol* **63**, 539-569, doi:https://doi.org/10.1146/annurev-psych-120710-100452 (2012).

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

62 Plomin, R., Haworth, C. M. & Davis, O. S. Common disorders are quantitative traits. *Nat Rev Genet* **10**, 872 - 878, doi:https://doi.org/10.1038/nrg2670 (2009).

63 Cuthbert, B. N. The RDoC framework: Facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry* **13**, 28-35, doi:https://doi.org/10.1002/wps.20087 (2014).

64 Stanton, K., McDonnell, C. G., Hayden, E. P. & Watson, D. Transdiagnostic approaches to psychopathology measurement: Recommendations for measure selection, data analysis, and participant recruitment. *J Abnorm Psychol* **129**, 21-28, doi:https://doi.org/10.1037/abn0000464 (2020).

65 Allen, N. *et al.* UK Biobank: Current status and what it means for epidemiology. *Health Policy Technol* **1**, 123-126, doi:https://doi.org/10.1016/j.hlpt.2012.07.003 (2012).

66 Strauss, M. E. & Smith, G. T. Construct validity: Advances in theory and methodology. *Annu Rev Clin Psychol* **5**, 1-25, doi:https://doi.org/10.1146/annurev.clinpsy.032408.153639 (2009).

67 Karcher, N. R., Michelini, G., Kotov, R. & Barch, D. M. Associations between resting-state functional connectivity and a hierarchical dimensional structure of psychopathology in middle childhood. *Biol Psychiatry: Cogn Neurosci Neuroimaging* **6**, 508-517, doi:https://doi.org/10.1016/j.bpsc.2020.09.008 (2021).

68 Tiego, J. *et al.* Dissecting schizotypy and its association with cognition and polygenic risk for schizophrenia in a non-clinical sample. *Schizophrenia Bulletin* (in press).

69 Conway, C. C., Forbes, M. K. & South, S. C. A Hierarchical Taxonomy of Psychopathology (HiTOP) Primer for Mental Health Researchers. *Clinical Psychological Science*, 21677026211017834 (2021).

70 Yim, O. & Ramdeen, K. T. Hierarchical cluster analysis: Comparison of three linkage measures and application to psychological data. *Quant Meth Psych* **11**, 8 - 21, doi:https://doi.org/10.20982/tqmp.11.1.p008 (2015).

71 Goldberg, L. R. Doing it all Bass-Ackwards: The development of hierarchical factor structures from the top down. *J Res Pers* **40**, 347-358, doi:https://doi.org/10.1016/j.jrp.2006.01.001 (2006).

72 Michelini, G. *et al.* Delineating and validating higher-order dimensions of psychopathology in the Adolescent Brain Cognitive Development (ABCD) study. *Transl Psychiatry* **9**, 261, doi:https://doi.org/10.1038/s41398-019-0593-4 (2019).

73 Simms, L. J. *et al.* Computerized adaptive assessment of personality disorder: Introducing the CAT–PD project. *Journal of personality assessment* **93**, 380-389, doi:https://doi.org/10.1080/00223891.2011.577475 (2011).

74 Greven, C. U., Buitelaar, J. K. & Salum, G. A. From positive psychology to psychopathology: The continuum of attention-deficit hyperactivity disorder. *Journal of child psychology and psychiatry* **59**, 203-212 (2018).

75 Stark, S., Chernyshenko, O. S. & Drasgow, F. Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *J Appl Psychol* **91**, 1292 - 1306, doi:https://doi.org/10.1037/0021-9010.91.6.1292 (2006).

76 van de Schoot, R. *et al.* Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front Psychol* **4**, 770, doi:https://doi.org/10.3389/fpsyg.2013.00770 (2013).

77 Clark, S. L. *et al.* Models and strategies for factor mixture analysis: An example concerning the structure underlying psychological disorders. *Struct Equ Modeling* **20**, 681-703, doi:https://doi.org/10.1080/10705511.2013.824786 (2013).

78 Eid, M., Lischetzke, T., Nussbeck, F. W. & Trierweiler, L. I. Separating trait effects from trait-specific method effects in multitrait-multimethod models: a multiple-indicator CT-C(M-1) model. *Psychol Methods* **8**, 38-60, doi:https://doi.org/10.1037/1082-989x.8.1.38 (2003).

79 Eid, M., Geiser, C. & Koch, T. Measuring method effects: From traditional to design-oriented approaches. *Curr Dir Psychol Sci* **25**, 275-280, doi:https://doi.org/10.1177/0963721416649624 (2016).

80 Aron, A. R. & Poldrack, R. A. The cognitive neuroscience of response inhibition: Relevance for genetic research in attention-deficit/hyperactivity disorder. *Biol Psychiatry* **57**, 1285-1292, doi:https://doi.org/10.1016/j.biopsych.2004.10.026 (2005).

81 Martinussen, R., Hayden, J., Hogg-Johnson, S. & Tannock, R. A Meta-Analysis of Working Memory Impairments in Children With Attention-Deficit/Hyperactivity Disorder. *Journal of the American Academy of Child & Adolescent Psychiatry* **44**, 377-384, doi:https://doi.org/10.1097/01.chi.0000153228.72591.73 (2005).

82 DeVellis, R. F. Classical test theory. *Med Care* **44**, S50-S59, doi:https://doi.org/10.1097/01.mlr.0000245426.10853.30 (2006).

83 Nunnally, J. C. & Bernstein, I. *Psychometric theory*. 3rd edn, (McGraw-Hill, 1994).

84 Antonakis, J., Bendahan, S., Jacquart, P. & Lalive, R. On making causal claims: A review and recommendations. *Leadersh Q* **21**, 1086-1120, doi:https://doi.org/10.1016/j.leaqua.2010.10.010 (2010).

85 Kendell, R. & Jablensky, R. Distinguishing between the validity and utility of psychiatric diagnoses. *Am J Psychiatry* **160**, 4-12, doi:https://doi.org/10.1176/appi.ajp.160.1.4 (2003).

86 Kendler, K. S. The phenomenology of major depression and the representativeness and nature of DSM criteria. *Am J Psychiatry* **173**, 771-780, doi:https://doi.org/10.1176/appi.ajp.2016.15121509 (2016).

87 Kotov, R., Ruggero, C. J., Krueger, R. F., Watson, D. & Zimmerman, M. The perils of hierarchical exclusion rules: A further word of caution. *Depress Anxiety* **35**, 903-904, doi:https://doi.org/10.1002/da.22826 (2018).

88 Caspi, A. *et al.* The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clin Psychol Sci* **2**, 119 - 137, doi:https://doi.org/10.1177/2167702613497473 (2014).

89 Allsopp, K., Read, J., Corcoran, R. & Kinderman, P. Heterogeneity in psychiatric diagnostic classification. *Psychiatry Research* **279**, 15-22, doi:https://doi.org/10.1016/j.psychres.2019.07.005 (2019).

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

91 Simms, L. J. *et al.* Development of measures for the hierarchical taxonomy of psychopathology (HiTOP): A collaborative scale development project. *Assessment*, 10731911211015309, doi:https://doi.org/10.1177/10731911211015309 (2021).

92 Lahey, B. B., Krueger, R. F., Rathouz, P. J., Waldman, I. D. & Zald, D. H. A hierarchical causal taxonomy of psychopathology across the life span. *Psychol Bull* **143**, 142 - 186, doi:http://dx.doi.org/10.1037/bul0000069 (2017).

93 Michelini, G., Palumbo, I. M., DeYoung, C. G., Latzman, R. D. & Kotov, R. Linking RDoC and HiTOP: A new interface for advancing psychiatric nosology and neuroscience. *Clin Psychol Rev* **86**, 102025, doi:https://doi.org/10.1016/j.cpr.2021.102025 (2021).

94 Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. E. *Multivariate data analysis*. Seventh edn, (Pearson Education Limited, 2014).

95 Grice, J. W. Computing and evaluating factor scores. *Psychol Methods* **6**, 430-450, doi:https://doi.org/10.1037/1082-989X.6.4.430 (2001).

96 Devlieger, I., Mayer, A. & Rosseel, Y. Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement* **76**, 741 - 770, doi:10.1177/0013164415607618 (2016).

97 Kim, J., Zhu, W., Chang, L., Bentler, P. M. & Ernst, T. Unified structural equation modeling approach for the analysis of multisubject, multivariate functional MRI data. *Hum Brain Mapp* **28**, 85-93, doi:https://doi.org/10.1002/hbm.20259 (2007).

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

99 Reise, S. P. & Rodriguez, A. Item response theory and the measurement of psychiatric constructs: Some empirical and conceptual issues and challenges. *Psychol Med* **46**, 2025-2039, doi:https://doi.org/10.1017/S0033291716000520 (2016).

100 de Ayala, R. J. *The theory and practice of item response theory* (The Guilford Press, 2009).