nature mental health

Review article

https://doi.org/10.1038/s44220-023-00057-5

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

Received: 20 October 2022

Accepted: 24 March 2023

Published online: 10 May 2023

Check for updates

Jeggan Tiego[1], Elizabeth A. Martin[2], Colin G. DeYoung[3], Kelsey Hagan[4], Samuel E. Cooper[5], Rita Pasion[6], Liam Satchell[7], Alexander J. Shackman[8], Mark A. Bellgrove[1], Alex Fornito[1] & the HiTOP Neurobiological Foundations Work Group*

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. The 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. Here, we discuss challenges, outline several ways forward and provide worked examples to demonstrate key problems and potential solutions. A precision phenotyping approach can enhance the discovery and replicability of associations between biology and psychopathology.

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 new 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, treatment and prognosis[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 prompted 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

[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, 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, Austin, TX, USA. [6]HEI-LAB, Lusófona University, Lisbon, Portugal. [7]Department of Psychology, University of Winchester, Winchester, UK. [8]Department of Psychology, University of Maryland, College Park, MD, USA. *A list of authors and their affiliations appears at the end of the paper. e-mail: jeggan.tiego@monash.edu

BOX 1

The relationship between measurement reliability and observed effect size

The relationship 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, a person's observed score on a psychological measurement reflects their 'true score' and 'random measurement error'[82]:

$$x = t + e \tag{1}$$

where x is the observed score, t is the true score, and e is random measurement error. Note that the error term, e, only represents random error, so the true score, t, can include systematic error unrelated to the construct of interest.

Thus, according to true score theory, all psychological measurement incorporates measurement error (that is, 'error-in-variables model'[49]). 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]:

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

where $r_{\mathrm{ox,oy}}$ is the observed correlation, $r_{\mathrm{tx,ty}}$ 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]. Relatedly, the standard error (s.e.) for the correlation coefficient increases as a function of smaller samples, n, and smaller effect sizes, r^2, resulting in reduced efficiency of estimation[83].

$$\mathrm{s.e.}_r = \sqrt{\frac{1-r^2}{n-2}} \tag{3}$$

Since the probability value of the correlation coefficient is based on the distribution of Student's t with $n-2$ degrees of freedom $\left(t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right)$, 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, 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].

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 statistically optimized psychiatric phenotypes (that is, observable characteristics or traits). 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: (1) the limitations of such measures are rarely discussed in comparison with the extensive literature devoted to improving biological measures; (2) prevalent practices to measure behavior are sub-optimal; and (3) 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, 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 effect 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,

BOX 2

Limitations of traditional approaches to psychiatric nosology

Existing diagnostic systems, such as DSM-5 and the ICD-11 have clinical utility, facilitating treatment and communication between mental health professionals and consumers of mental health services[84]. However, the psychopathological concepts invoked by modern nosology may have a tenuous relationship with biological correlates, undermining our attempts to link measurement of behavioral 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[85]. 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 (that is, diagnoses defined by an established minimum number of criteria, not all of which are required for diagnosis). Prototypical symptoms occurring in pre-specified numbers and combinations 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 the 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[86]. Additionally, focusing on a particular diagnostic category assumes homogeneity of symptoms and mechanisms (the homogeneity assumption—the assumption that different people with the same psychiatric diagnosis are phenotypically similar), but individuals with the same diagnosis may exhibit little to no overlap in symptoms (the heterogeneity problem—the grouping of cases with divergent symptom presentations into the same diagnostic category, or the grouping of symptoms with divergent etiology, pathophysiology, course and/or treatment response)[34]. Co-morbidity between putatively distinct disorders (that is, the comorbidity problem—psychiatric disorders co-occur in the same individuals more often than would be expected for independent entities, suggesting shared phenomenology and etiology)[87], and issues of arbitrary clinical cut-offs and ignoring of the clinical significance of subthreshold symptomatology are well-documented limitations of current psychiatric taxonomies[88]. These limitations obfuscate the search for the neurobiological correlates of psychiatric symptoms and constitute an impediment to future research in this domain[89].

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 (that is, observed biology–psychopathology associations are inversely proportional to measurement reliability), mandating larger and more expensive samples for adequate power and reproducibility[20] (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 psychometrically well developed and yet our understanding of the neurobiology and genetics of intelligence is incomplete. 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.

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 ref. 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 with a group of patients meeting a 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 Review, 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

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 homogeneous, hierarchically organized dimensions (Fig. 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 homogeneous 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, which is situated under the broader internalizing spectrum within the HiTOP model. Thus, 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.

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[90]. In the meantime, several existing instruments can be used to reliably assess HiTOP dimensions in youth and adults[91]. 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 homogeneous 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 (for example, 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 (that is, the reliability or precision of measurement of a phenotype along the full spectrum of the latent trait continuum). 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 Research Domain Criteria (RDoC) domains[93].

to representativeness, ethnic matching of biological measures and generalizability of predictions of behavior 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].

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 with averaging across two or more occasions, thereby attenuating 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 (for example, 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, it has been demonstrated[35] 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].

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 (that is, phenotypic complexity). Collapsing complex, multidimensional psychiatric phenotypes (for example, depression) into unitary scores has the potential to obscure biologically and clinically important sources of variance (for example, anhedonia versus 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 Fig. 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 (for example, fear, distress and substance abuse). Too often, simple summary scores ignore this structure, combining both broad and narrow sources of variance[43], leading to attenuation of biology–psychopathology associations.

We show in example 1 of the Supplementary Information 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

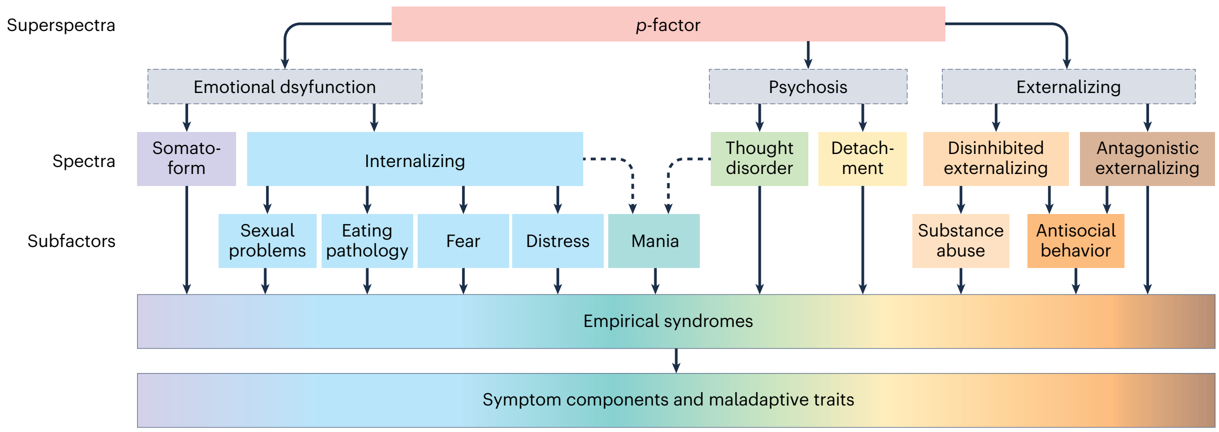


Fig. 1 | The HiTOP model. The broadest dimensions, reflecting common liabilities to psychopathology, are situated at the top of the hierarchy with the narrowest traits and symptom components situated at the bottom, reflecting liabilities to specific problems. Gray boxes with broken lines indicate hypothesized, but not yet confirmed, constructs. The broken single-headed arrows pointing to 'Mania' reflect preliminary relationships awaiting further confirmatory evidence.

BOX 4

Structural equation modeling

Hierarchical modeling, measurement invariance, mixture modeling and the T(M-1) model can be done within an SEM framework. 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 (that is, 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 (that is, variance attributable to a separate construct and measurement error). The formula is expressed as:

$$x_i = a_x + \lambda_x \xi_i + \theta\varepsilon_i \tag{4}$$

where x_i is a measured variable (that is, observed or 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 or error of the measured variable that is not explained by the factor loading (Fig. 2). This model formalizes the following: (1) the target psychopathology phenotype is unobserved and must be inferred by one or more measured variables (for example, 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 (that is, λ_x).

In a structural regression model, SEM enables estimation of regression path coefficients between factors within the model. Thus, SEM estimates the empirical relationships 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 relationships 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 (that is, 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 (that is, correlational accuracy) and are not contaminated by variance from orthogonal factors (that is, univocality) will also vary between studies[95]. The use of factor score estimates can also potentially bias the parameter estimates of the regression models[96]. Thus, we recommend against this approach in favor of SEM.

Ideally, biological measurements should be incorporated directly into latent models to capitalize on the increased measurement precision and statistical power that these models afford (for example, ref. 97). However, SEM generally requires sample sizes 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 ref. 98.

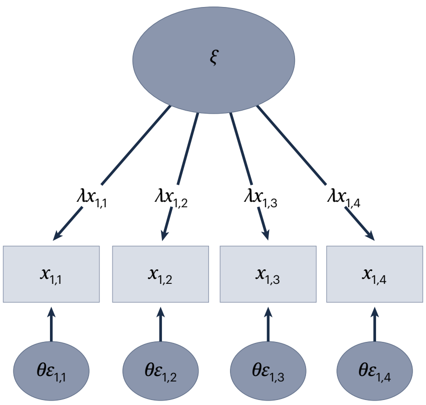


Fig. 2 | The reflective latent variable model. Reflective latent variable (common factor) model in which the unobserved psychobiological attribute (factor or latent construct; ξ), is conceptualized as explaining the variance/covariance in the measured variables ($x_{1,1}$–$x_{1,4}$) via their factor loadings ($\lambda x_{1,1}$–$\lambda x_{1,4}$), 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 (that is, variance not shared with the other indicator variables).

Behavior Checklist (CBCL) data from the ABCD study, which exhibits both multidimensionality and hierarchical structure. The CBCL is a multidimensional instrument that measures eight empirical syndromes using eight distinct subscales. The CBCL can be modeled as having a hierarchical structure with variance attributable to three levels, which approximates the scoring system typically applied with this instrument. We used a bifactor model[44] within a structural equation modeling (SEM) framework (Box 4 and Fig. 2) to separate these dimensions into three orthogonal (that is, 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. Common variance across the eight scales is quite reliable (r_{xx}= 0.847), such that collapsing measurement of psychopathology into the unidimensional total problems score would result in attenuation of biology–psychopathology associations unique to the *p*-factor by just 7.9%, assuming perfect reliability of the biological measure.

Results are worse for the other two composite scales, internalizing problems and externalizing problems, where reliable variance uniquely attributable to these group dimensions is only 3.1 and 2.3% (r_{xx} = 0.031 and 0.023), rendering these scales unreliable and unusable. We also demonstrate that high phenotypic complexity across the eight empirical syndrome scales leads to low residual variance for these individual scales (that is, an average of approximately 43.2% variance is unique to each scale).

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 (that is, 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 ($r_{xx} < 0.6$) below the mean[48]. Because reliability places an upper bound on associations 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 Information, 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 has reliable scores for most of the CBCL scales. Specifically, we find unacceptably low reliability, even for basic research purposes ($r_{xx} < 0.6$), at or below one standard deviation below the mean for ten of the eleven scales (that is, 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% and 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.

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 (that is, measurement invariance)[50]. Yet there is ample evidence that measurement properties can vary (that is, non-invariance) across demographic groups (for example, sex) or unobserved or latent classes (that is, homogeneous subpopulations or 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, four levels of invariance are evaluated: (1) configural invariance; (2) weak invariance; (3) strong invariance; and (4) strict invariance (Supplementary Table 3 contains technical definitions)[50]. Unfortunately, only a small proportion of studies test for full measurement invariance[50]; thus, combining raw scores across discrete groups (for example, sex and ethnicity) for biology–psychopathology associations remains problematic. In example 3 of the Supplementary Information, 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.

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 (that is, mixing apples and oranges). This problem is exacerbated in case–control research because traditional DSM and 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 a 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, three distinct genetic factors were identified that explained the co-occurrence of distinct subsets of DSM criteria and 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 homogeneous and unidimensional symptom components or even individual symptoms is likely to be a more informative and productive avenue for future biology–psychopathology association studies.

Method bias

Method bias (sources of systematic measurement error stemming from the measurement process, such as method effects, for constructs) 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 (that is, 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 relationships 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].

Table 1 | Sources of imprecision in psychopathology phenotyping and proposed solutions

Problem	Solution
Sampling bias	Dimensional sampling and measurement
Minimal and inconsistent phenotyping	Deep phenotyping and use of standardized measures
Phenotypic complexity	Use of homogeneous unidimensional scales, test for multidimensionality and model hierarchical relations between dimensional constructs
Poor phenotypic resolution	Increase phenotypic resolution by adding items assessing the adaptive end of the continuum
Measurement non-invariance	Test for and accommodate measurement non-invariance
The heterogeneity problem	
Person-centered heterogeneity	Mixture modeling
Variable-centered heterogeneity	Broadband assessment of psychopathology and hierarchical modeling
Method bias	Multi-method assessment

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 Fig. 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) 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 (versus 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 oversampled (Fig. 3).

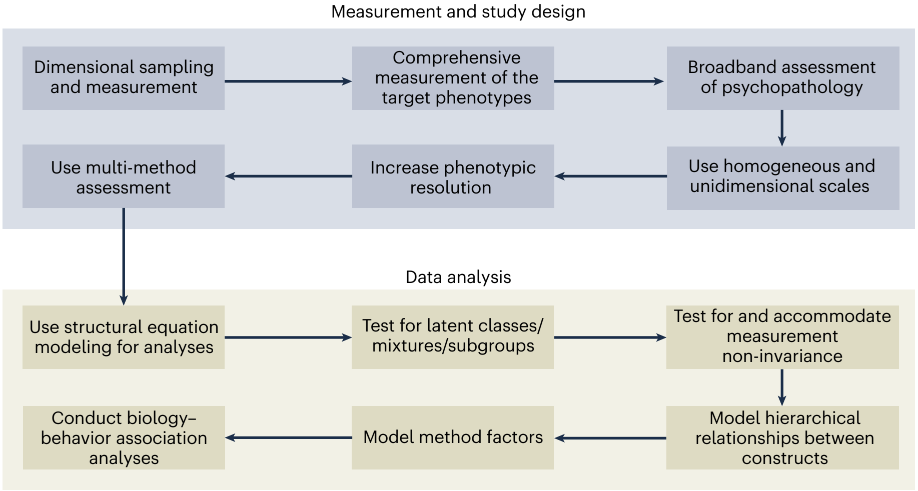

Measurement and study design
Dimensional sampling and measurement
Comprehensive measurement of the target phenotypes
Broadband assessment of psychopathology
Use homogeneous and unidimensional scales
Increase phenotypic resolution
Use multi-method assessment
Data analysis
Use structural equation modeling for analyses
Test for latent classes/ mixtures/subgroups
Test for and accommodate measurement non-invariance
Model hierarchical relationships between constructs
Model method factors
Conduct biology–behavior association analyses

Fig. 3 | Precision psychiatric phenotyping. Example workflow for a precision psychiatric phenotyping approach in the context of a biology–psychopathology association study.

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 (comprehensive assessment of one or more phenotypes) with standardized psychopathology measures that can be widely adopted (for example, Box 3), and which are better suited for data pooling via established psychiatric research consortia (for example, ENIGMA and 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].

Use of homogeneous unidimensional scales and hierarchical modeling

Construct homogeneity (that is, the assumption or evidence that a construct reflects variance in a single phenotype) and unidimensionality (that is, 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) are important qualities of scales used to assess psychopathology that enable researchers to isolate the specific sources of variance associated with biological measures[66]. Relatedly, owing to the potential empirical overlap of symptom components or 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 homogeneous constructs to isolate specific variance while taking account of the hierarchical organization of these phenotypes[67]. A previous study[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).

Psychiatric symptoms are intrinsically hierarchical. Even homogeneous 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 (compare with example 1 of the Supplementary Information). 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 an SEM 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 (for example, *p*-factor) and orthogonal specific factors (for example, internalizing, externalizing; see example 1 of the Supplementary Information for a detailed illustration)[44]. An alternative bottom-up approach uses hierarchical clustering, where questionnaire items or subscales are organized into homogeneous 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 of 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 or symptom constellations are isolated. An example of the bass-ackwards method in the ABCD data is provided in ref. 72.

Increasing phenotypic resolution

To address the issue of low phenotypic resolution, items can be carefully selected within an iIRT 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 (for example, the computerized adaptive assessment of personality disorder;

BOX 5

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 ±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 (that is, M=0, s.d.=1) generally ranging between ±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 $\left[r_{xx} = 1 - \left(\frac{1}{I}\right)\right]$ (ref. 100). Items can thus be carefully selected to optimize measurement precision across the whole latent-trait continuum. Furthermore, items with high local dependence (that is, 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.

CAT-PD[73]). For unipolar traits, it is possible to bolster measurement precision with items from a related construct that represents the opposite (that is, adaptive) end of the continuum[74]. We demonstrate the utility of this approach in example 4 of the Supplementary Information, 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.

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 cautious of changing the substantive interpretation of the construct by narrowing its scope and breadth (that is, the attenuation paradox).

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 the classification accuracy of participants based on the posterior probabilities of class membership within a mixture modeling analysis. It can range between 0 and 1.00, with higher entropy indicating better classification accuracy. When entropy is high (for example, ≥0.80) 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 the Supplementary Information, 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.

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 an SEM framework (Box 4)[79].

In example 6 of the Supplementary Information, 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 significantly related to polygenic risk for ADHD (Supplementary Fig. 27). 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. 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 homogeneous 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.

References

1. Perkins, E. R., Latzman, R. D. & Patrick, C. J. Interfacing neural constructs with the Hierarchical Taxonomy of Psychopathology: 'why' and 'how'. *Personal. Ment. Health* **14**, 106–122 (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 (2010).
3. Hyman, S. E. Can neuroscience be integrated into the DSM-V? *Nat. Rev. Neurosci.* **8**, 725–732 (2007).
4. Singh, I. & Rose, N. Biomarkers in psychiatry. *Nature* **460**, 202–207 (2009).
5. First, M. B. et al. Clinical applications of neuroimaging in psychiatric disorders. *Am. J. Psychiatry* **175**, 915–916 (2018).
6. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
7. Poldrack, R. A. et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* **18**, 115 (2017).
8. Saggar, M. & Uddin, L. Q. Pushing the boundaries of psychiatric neuroimaging to ground diagnosis in biology. *eNeuro* 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 (2019).
10. Smoller, J. W. et al. Psychiatric genetics and the structure of psychopathology. *Mol. Psychiatry* **24**, 409–420 (2019).
11. Nour, M. M., Liu, Y. & Dolan, R. J. Functional neuroimaging in psychiatry and the case for failing better. *Neuron* **110**, 2524–2544 (2022).
12. Sullivan, P. F. et al. Psychiatric genomics: an update and an agenda. *Am. J. Psychiatry* **175**, 15–27 (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 (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 (2018).
15. Kong, R. et al. Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. *Cereb. Cortex* **31**, 4477–4500 (2021).
16. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and ranslation. *Am. J. Hum. Genet.* **101**, 5–22 (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 (2018).
18. Lilienfeld, S. O. The Research Domain Criteria (RDoC): an analysis of methodological and conceptual challenges. *Behav. Res. Ther.* **62**, 129–139 (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 (2018).
20. Zuo, X. N., Xu, T. & Milham, M. P. Harnessing reliability for neuroscience research. *Nat. Hum. Behav.* **3**, 768–771 (2019).
21. Nikolaidis, A. et al. Suboptimal phenotypic reliability impedes reproducible human neuroscience. Preprint at *bioRxiv* 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. *Proc. Natl. Acad. Sci. USA* **110**, 17615–17622 (2013).
23. Markon, K. E., Chmielewski, M. & Miller, C. J. The reliability and validity of discrete and continuous measures of psychopathology: a quantitative review. *Psychol. Bull.* **137**, 856–879 (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 (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 (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 (1999).
27. Preacher, K. J. in *Extreme Groups Designs in the Encyclopedia of Clinical Psychology* Vol. 2 (eds. Cautin, R. L. & Lilienfeld, S. O.) 1189–1192 (John Wiley and Sons, 2015).
28. Dong, H.-M. et al. Charting brain growth in tandem with brain templates at school age. *Sci. Bull.* **65**, 1924–1934 (2020).
29. Peterson, R. E. et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).
30. Liu, S. et al. Chinese color nest project: an accelerated longitudinal brain-mind cohort. *Dev. Cogn. Neurosci.* **52**, 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 (2022).
32. Sanchez-Roige, S. & Palmer, A. A. Emerging phenotyping strategies will advance our understanding of psychiatric genetics. *Nat. Neurosci.* **23**, 475–480 (2020).
33. Newson, J. J., Hunter, D. & Thiagarajan, T. C. The heterogeneity of mental health assessment. *Front. Psychiatry* https://doi.org/10.3389/fpsyt.2020.00076 (2020).
34. Feczko, E. et al. The heterogeneity problem: approaches to identify psychiatric subtypes. *Trends Cogn. Sci.* **23**, 584–601 (2019).
35. Schnack, H. G. & Kahn, R. S. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. *Front. Psychiatry* **7**, 50 (2016).

36. Yang, Z. et al. Brain network informed subject community detection in early-onset schizophrenia. *Sci. Rep.* **4**, 5549 (2014).
37. Hodgson, K., McGuffin, P. & Lewis, C. M. Advancing psychiatric genetics through dissecting heterogeneity. *Hum. Mol. Genet.* **26**, R160–R165 (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 (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 (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 (2010).
41. Caspi, A. & Moffitt, T. E. All for one and one for all: mental disorders in one dimension. *Am. J. Psychiatry* **175**, 831–844 (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 (2021).
43. Clark, L. A. & Watson, D. Constructing validity: new developments in creating objective measuring instruments. *Psychol. Assess.* **31**, 1412–1427 (2019).
44. Reise, S. P. The rediscovery of bifactor measurement models. *Multivariate Behav. Res.* **47**, 667–696 (2012).
45. Reise, S. P. & Waller, N. G. Item response theory and clinical measurement. *Annu. Rev. Clin. Psychol.* **5**, 27–48 (2009).
46. Rosopa, P. J., Schaffer, M. M. & Schroeder, A. N. Managing heteroscedasticity in general linear models. *Psychol. Methods* **18**, 335–351 (2013).
47. Thomas, M. L. The value of item response theory in clinical assessment: a review. *Assessment* **18**, 291–307 (2011).
48. Streiner, D. L. Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J. Pers. Assess.* **80**, 99–103 (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 (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 (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 (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 and 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 (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 (2015).
56. Wager, T. D. & Woo, C.-W. Imaging biomarkers and biotypes for depression. *Nat. Med.* **23**, 16–17 (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 (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 (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 (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 (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 (2012).
62. Plomin, R., Haworth, C. M. & Davis, O. S. Common disorders are quantitative traits. *Nat. Rev. Genet.* **10**, 872–878 (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 (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 (2020).
65. Allen, N. et al. UK Biobank: current status and what it means for epidemiology. *Health Policy Technol.* **1**, 123–126 (2012).
66. Strauss, M. E. & Smith, G. T. Construct validity: advances in theory and methodology. *Annu. Rev. Clin. Psychol.* **5**, 1–25 (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 (2021).
68. Tiego, J. et al. Dissecting schizotypy and its association with cognition and polygenic risk for schizophrenia in a nonclinical sample. *Schizophr Bull.* https://doi.org/10.1093/schbul/sbac016 (2023).
69. Conway, C. C., Forbes, M. K. & South, S. C. A Hierarchical Taxonomy of Psychopathology (HiTOP) primer for mental health researchers. *Clin. Psychol. Sci.* https://doi.org/10.1177/21677026211017834 (2021).
70. Yim, O. & Ramdeen, K. T. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *Quant. Methods Psychol.* **11**, 8–21 (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 (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 (2019).
73. Simms, L. J. et al. Computerized adaptive assessment of personality disorder: introducing the CAT–PD project. *J. Pers. Assess.* **93**, 380–389 (2011).
74. Greven, C. U., Buitelaar, J. K. & Salum, G. A. From positive psychology to psychopathology: the continuum of attention-deficit hyperactivity disorder. *J. Child Psychol. 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 (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 (2013).
77. Clark, S. L. et al. Models and strategies for factor mixture analysis: an example concerning the structure underlying psychological disorders. *Struct. Equation Modell.* **20**, 681–703 (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 (2003).
79. Eid, M., Geiser, C. & Koch, T. Measuring method effects: from traditional to design-oriented approaches. *Curr. Dir. Psychol. Sci.* **25**, 275–280 (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 (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. *J. Am. Acad. Child Adolesc. Psychiatry* **44**, 377–384 (2005).
82. DeVellis, R. F. Classical test theory. *Med. Care* **44**, S50–S59 (2006).
83. Antonakis, J., Bendahan, S., Jacquart, P. & Lalive, R. On making causal claims: a review and recommendations. *Leadersh. Q.* **21**, 1086–1120 (2010).
84. Kendell, R. & Jablensky, R. Distinguishing between the validity and utility of psychiatric diagnoses. *Am. J. Psychiatry* **160**, 4–12 (2003).
85. Kendler, K. S. The phenomenology of major depression and the representativeness and nature of DSM criteria. *Am. J. Psychiatry* **173**, 771–780 (2016).
86. 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 (2018).
87. Caspi, A. et al. The p factor: one general psychopathology factor in the structure of psychiatric disorders? *Clin. Psychol. Sci.* **2**, 119–137 (2014).
88. Allsopp, K., Read, J., Corcoran, R. & Kinderman, P. Heterogeneity in psychiatric diagnostic classification. *Psychiatry Res.* **279**, 15–22 (2019).
89. Cuthbert, B. N. & Insel, T. R. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med.* **11**, 126 (2013).
90. Simms, L. J. et al. Development of measures for the hierarchical taxonomy of psychopathology (HiTOP): a collaborative scale development project. *Assessment* **29**, 3–16 (2021).
91. HiTOP Friendly Measures. *HiTOP Clinical Network* https://hitop.unt.edu/clinical-tools/hitop-friendly-measures (accessed 1 October 2022).
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 (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 (2021).
94. Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. E. *Multivariate Data Analysis* 7th edn (Pearson Education, 2014).
95. Grice, J. W. Computing and evaluating factor scores. *Psychol. Methods* **6**, 430–450 (2001).
96. Devlieger, I., Mayer, A. & Rosseel, Y. Hypothesis testing using factor score regression: a comparison of four methods. *Educ. Psychol. Meas.* **76**, 741–770 (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 (2007).
98. Kline, R. B. *Principles and Practice of Structural Equation Modeling* 4th edn (Guilford, 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 (2016).
100. de Ayala, R. J. *The Theory and Practice of Item Response Theory* (Guilford, 2009).

Acknowledgements

J.T. was supported by a Turner impact fellowship from the Turner Institute for Brain and Mental Health, Monash University, Australia. A.F. was supported by the Sylvia and Charles Viertel Foundation, the National Health and Medical Research Council (grant numbers: 1146292 and 1197431) and the Australian Research Council (grant number: DP200103509). A.J.S. was supported by the National Institute of Mental Health (grant numbers: AA030042, MH131264 and MH121409) and the University of Maryland.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44220-023-00057-5.

Correspondence and requests for materials should be addressed to Jeggan Tiego.

Peer review information *Nature Mental Health* thanks Xi-Nian Zuo, Terrie Moffitt and Gregory Miller for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

the HiTOP Neurobiological Foundations Work Group

Rany Abend[9], Natalie Goulter[10], Nicholas R. Eaton[11], Antonia N. Kaczkurkin[12] & Robin Nusslock[13]

[9]School of Psychology, Reichman University, Herzliya, Israel. [10]School of Psychology, Newcastle University, Newcastle upon Tyne, UK. [11]Departments of Psychology and Psychiatry and Behavioral Health, Stony Brook University, Stony Brook, NY, USA. [12]Department of Psychology, Vanderbilt University, Nashville, TN, USA. [13]Department of Psychology, Northwestern University, Evanston, IL, USA.

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

Correction to: *Nature Mental Health* (2023) **1**:304-315 https://doi.org/10.1038/s44220-023-00057-5, published online 10 May 2024.

https://doi.org/10.1038/s44220-024-00366-3

[redacted]

Check for updates

Jeggan Tiego, Elizabeth A. Martin, Colin G. DeYoung, Kelsey Hagan, Samuel E. Cooper, Rita Pasion, Liam Satchell, Alexander J. Shackman, Mark A. Bellgrove, Alex Fornito & the HiTOP Neurobiological Foundations Work Group*

In the version of the article initially published, in Box 1, equation (1) was for true score sample variance as a function of observed scored variance, and error variance, rather than the formula for an individual's true score as a function of their observed score and measurement error. We also incorrectly wrote that the error term includes systematic error. However, the error term captures only random error, such that the true score can include systematic error unrelated to the construct of interest. In the "Phenotypic complexity" section, we amended the text to clarify that a three-tiered hierarchical structure is not an intrinsic property of the CBCL, but one that is imposed by specification of a bifactor model, which we applied because it roughly reflects the scoring structure of the CBCL (i.e., Total Problems, Internalizing, and Externalizing composite scales, and eight empirical syndrome scales), and enables this measure to be partitioned into unique sources of variance for further analysis. We incorrectly specified the bifactor model (Supplementary Fig. 1) with a negative correlation between the Internalizing and Externalizing group factors and with too many error covariances (i.e., correlations between variances in the observed variables not explained by the factors). We re-estimated the model without specifying these additional parameters and obtained an adequate fit to the data. Based on the re-estimated parameters of this model, we recalculated the proportion of reliable variance unique to each of the three composite scales, and the residual variance unique to each of the eight CBCL syndrome scales. In the "Measurement non-invariance" section, we did not explicitly mention configural invariance, which is tested prior to weak invariance. We similarly omitted configural invariance from Supplementary Table 3, and the incorrect labels were assigned to each level of invariance. This, along with definitions, have been corrected. Under the solution column in Table 1 and the third box of Data Analysis in Fig. 3, we mistakenly wrote "measurement invariance" instead of "measurement non-invariance." These corrections have been made to the HTML and PDF versions of the article.

*A list of authors and their affiliations appears online.

nature mental health

Supplementary information https://doi.org/10.1038/s44220-023-00057-5

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

In the format provided by the
authors and unedited

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

Supplementary Information

62 pages

9 Tables

27 Figures

Words: 4,151

Table of contents

List of Tables

List of Figures

Example 1 – Phenotypic complexity

To demonstrate the problem of phenotypic complexity, we modeled the CBCL data from the two-year follow-up wave of the ABCD study cohort using a bifactor model, which enables variance to be partitioned into common and scale-specific components[1]. To evaluate model-data consistency we report the chi-square (χ^2) test statistic with associated model degrees of freedom and probability value (*p*); $p > .05$ indicates that the null hypothesis of exact fit of the model to the data cannot be rejected, but this statistic is overly sensitive to minor model misspecification in large samples, such as the current one[2]. Thus, we also report the root mean square error of approximation (RMSEA), standardized root mean squared residual (SRMR), and comparative fit index (CFI), where lower values of the RMSEA and SRMR, and higher values of the CFI, indicate a better-fitting model[2]. As a potential substitute for the χ^2 statistic, the matrix of correlation residuals informs on local model fit, where residuals below .10 indicate that the observed bivariate relationships between the variables are being closely reproduced by the model[2]. The model is displayed in Supplementary Figure 1. The model failed the exact fit test, ($\chi^2(16) = 587.893$, $p < .001$, RMSEA = .078, [95%CI = .073, .084], CFI = .972, SRMR = .027). However, it passed the local fit test in that all the correlation residuals were below .10, indicating that model misspecification error was minor and ignorable. It was therefore concluded that the model provided an acceptable fit to the data.

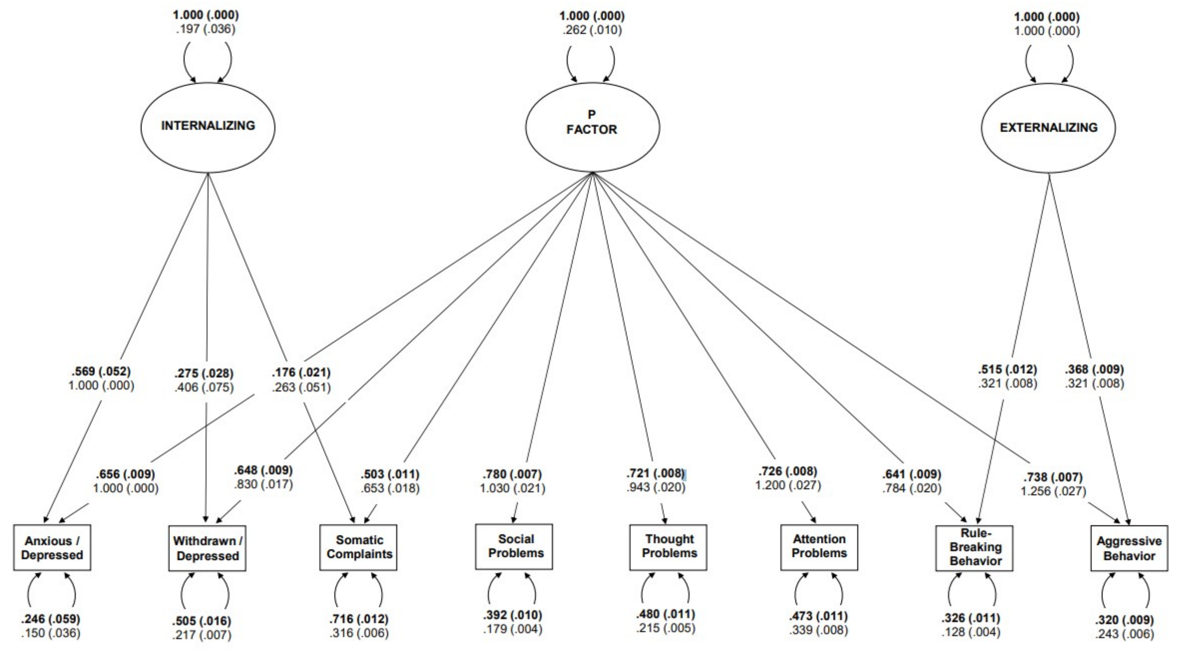
A shown in Supplementary Figure 2, residual variance (including measurement error) unique to each of the scales (but some possibly shared with one or more of the other subscales in the form of error covariances), after removal of the general and group factor variances, ranged from as low as 24.6% for the Anxious/Depressed subscale to 71.6% for the Somatic Complaints subscale, but averaged just 43.3% across the eight subscales.

Reliable variance unique to the Internalizing and Externalizing composite scales after removal of variance attributable to the general factor and subscales was just 3.1 and 2.3% respectively, rendering them unusable as standalone measures. Less extreme reductions in reliable variance would still attenuate relationships between these measures and criterion variables (e.g., genetic markers and imaging-derived phenotypes), thereby obscuring psychopathology-biological associations (assuming these measures are valid indices of psychopathology and identifiable and meaningful underlying biological substrates, respectively).

A recent landmark study by Marek et al. (2022)[4] reported a median effect size of *r* = 0.06 across all possible brain-wide associations between various MRI-derived measures of brain structure and function, and different metrics of cognitive ability as measured by the National Institute of Health (NIH) Toolbox[5], and personality and psychopathology[12], as measured by the CBCL[6]; short form[7,8] of the Urgency, (Lack of) Premeditation, (Lack of) Perseverance, Sensation Seeking, Positive Urgency (sUPPS-P) Behavioral Impulsivity scale[9-11]; the child version[12] of the Behavioral Inhibition / Behavioral Activation (BIS/BAS) scales[13]; and the Pediatric Psychosis Questionnaire − Brief Version[14,15]. However, using equation 1 from the main text, we can see that unreliability of measurement due to phenotypic complexity of one or more of these instruments may have resulted in attenuation bias in these observed brain-behavior associations. Conversely, we can correct for attenuation of the correlation coefficient using the formula,

$$r_{tx,ty} = \frac{r_{ox,oy}}{\sqrt{r_{yy} r_{xx}}}, \tag{1}$$

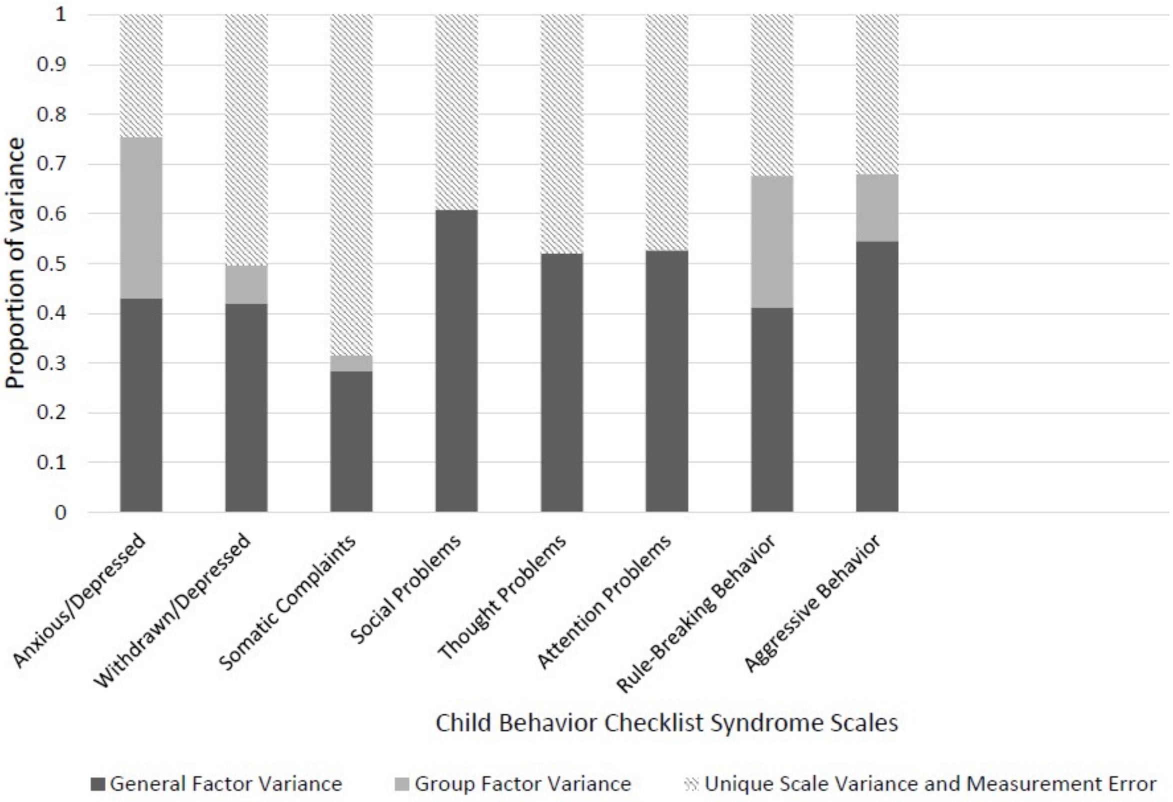
which indicates that, even if we assume zero error in the imaging-derived phenotypes, the true correlations could be considerably higher than those observed and reported. Considering that r_{es} = .10, .20, and .30 correspond with small, medium, and large effects sizes respectively[16], the true effect sizes could be meaningfully higher than those observed and reported when phenotypic complexity has not been taken into account. These disattenuated correlations also have major implications for statistical power and sample size planning[17]. We further note that while Marek et al. (2022)[4] address the notion of attenuation bias and disattenuation correction by arguing that the reliability of the behavioral phenotypes, including the CBCL scales, is at - or near -ceiling, these calculations rely on taking the alpha reliability estimates of the CBCL scales on face value (acceptable to high). Furthermore, as we demonstrate below in example 2, the reliability of a given psychopathology measure varies along the latent trait continuum and usually drops below acceptable levels below the mean. Attenuation of biology-behavior associations can be substantial when high phenotypic complexity (and low phenotypic resolution) is not considered.



Supplementary Figure 1. Bifactor model of the CBCL data obtained from the two-year follow-up data collection wave of the ABCD study cohort.

Note. **Model fit statistics** $\chi2$ (16) = 587.893, p < .001, RMSEA = .078, [95%CI **= .073, .084]. All correlation residuals were below .10.**

Model figure is displayed using symbols from the McArdle-McDonald reticular action model[18]. Observed (also measured or manifest) variables are represented as rectangles. Factors (latent variables or constructs) are represented as large ellipses. Error variances for observed variables, are symbolized with small double-headed arrows pointing to the rectangles. Double-headed, curved arrows pointing to factors are the latent variable variances. Straight, single-headed arrows from large ellipses to observed variables reflect factor loadings.



Supplementary Figure 2. Proportion of variance in the CBCL Scales in 5,820 participants from the two-year follow-up wave of the ABCD study cohort that is unique to the eight syndrome scales versus what is general factor variance (i.e., overarching *p*-factor), and what is specific to each of the two group factors (internalizing or externalizing).

Image taken from Tiego and Fornito (2022)[19]. Reprinted with permission.

Example 2 - Low phenotypic resolution

To illustrate the problem of low phenotypic resolution in psychiatric phenotypes, we first calculated the internal consistency reliability using Cronbach's alpha (α) for each of the eight syndrome scales and three composite scales of the CBCL. We then plotted the total information functions (TIFs) within an item response theory (IRT) framework for each of the eight CBCL empirical syndrome scales and the three CBCL composite scales (i.e., Internalizing, Externalizing, and Total Problems). A TIF represents the additive measurement precision (i.e., information) contributed by items on a questionnaire scale/subscale or other performance measure[20]. IRT is distinct from classical test theory in that it does not assume reliability is uniform across the latent-trait continuum. Rather than standard measures of reliability from classical test theory (e.g., Cronbach's α), a TIF plots the total information (i.e., measurement precision) contributed by the retained questionnaire items, which varies across different points of the latent trait continuum. We can then calculate the corresponding reliability in the population (where zero is the population mean and one the population standard deviation) [21,22] associated with each point of the latent trait continuum for each phenotype using the formula: $r_{xx} = 1 - \left({}^{1}/_{I}\right)$[23].

Although the classic metric of internal consistency reliability indexed using Cronbach's α demonstrated acceptable levels of reliability for all eleven scales (α = .68 - .95), IRT analysis revealed unacceptably low reliability even for basic research purposes ($r_{xx} < .6$)[24] at or below one standard deviation below the mean for all scales accept the Total Problems scale (Table S1). This low reliability is non-trivial when considering that scores on the CBCL are strongly positively skewed[25,26] with most children scoring at the lower end of the scale (Supplementary Figures 3 – 13). We therefore calculated the proportion of the sample with unreliable scores ($r_{xx} < 0.60$) for each of the CBCL scales (Supplementary Table 2). On average, 37.2% of the sample would have unreliable scores. More than half of the sample had unreliable scores for 3 of the 11 scales. In short, a substantial proportion of ABCD participants have scores with unacceptably low reliability, which will necessarily attenuate observed biology-psychopathology associations. This analysis demonstrates the problems posed by taking scale reliability estimates at face value.

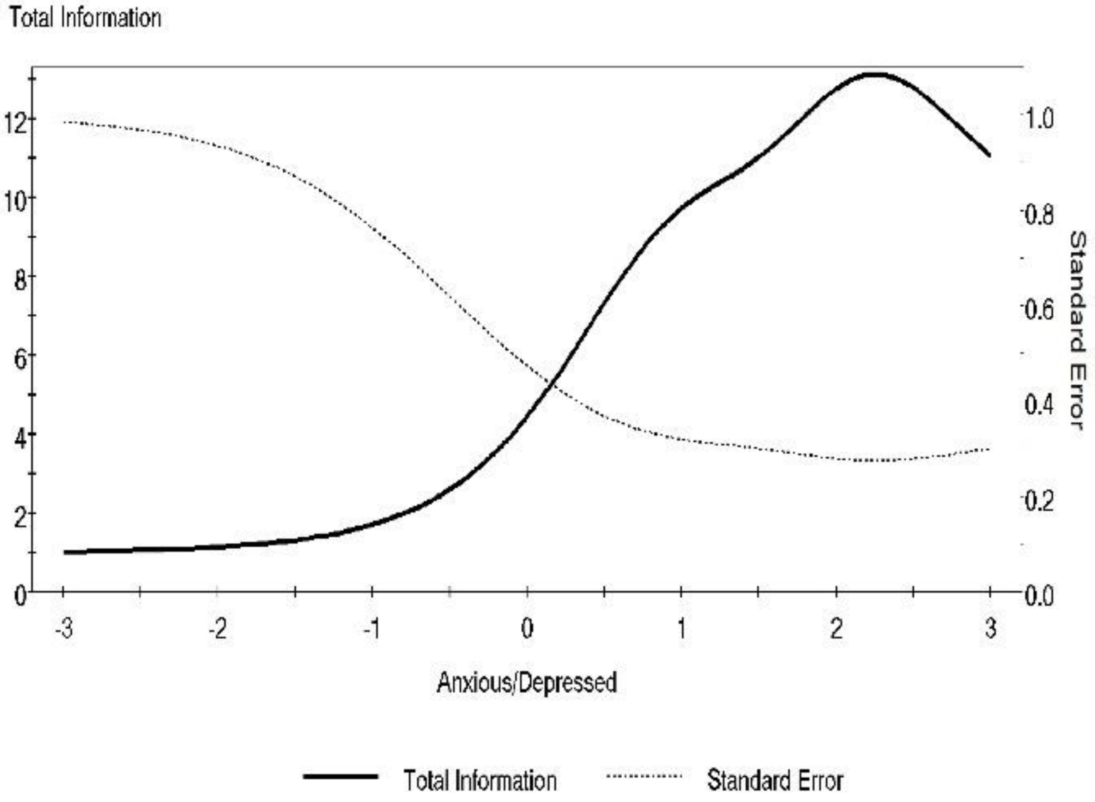
Supplementary Table 1

Reliability of the Child Behavior Checklist Scales Across the Latent Trait Continuum Estimated Using Unidimensional Item Response Theory Analysis

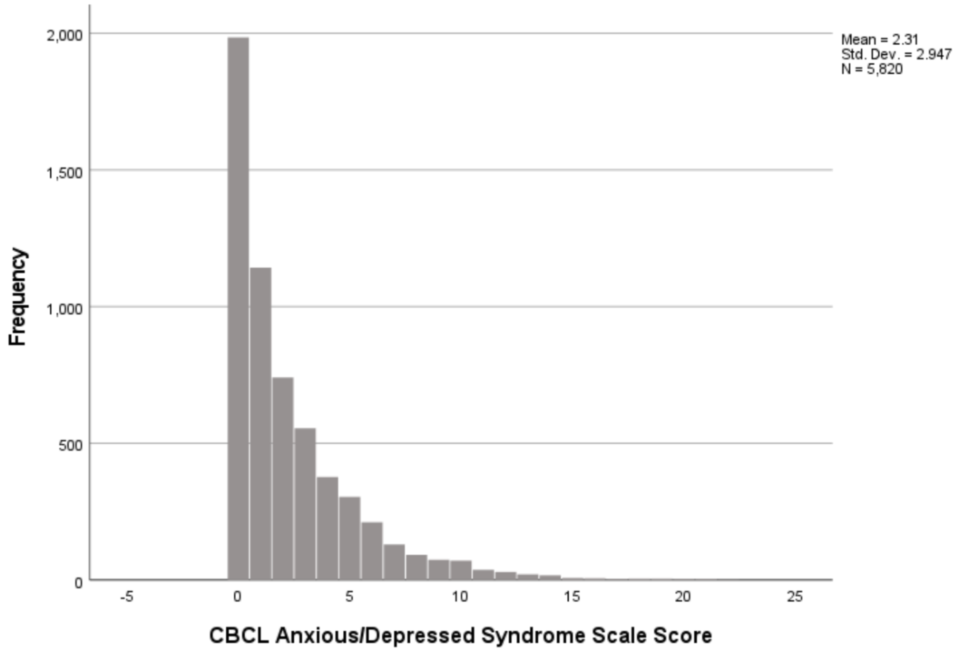
			Reliability r_{xx} (*I*) Across Latent Trait Continuum (θ)												
CBCL Scale	Number of Items	α	-3.0 *SD*	-2.5 *SD*	-2.0 *SD*	-1.5 *SD*	-1.0 *SD*	-0.5 *SD*	*M*	+0.5 *SD*	+1.0 *SD*	+1.5 *SD*	+2.0 *SD*	+2.5 *SD*	+3.0 *SD*
Anxious/Depressed	13	.813	.030	.061	.125	.241	.417	.616	.775	.863	.900	.911	.922	.922	.909
			(1.0304)	(1.0654)	(1.1431)	(1.3178)	(1.7141)	(2.6040)	(4.4470)	(7.3177)	(10.0273)	(11.2038)	(12.7643)	(12.7768)	(11.0446)
Withdrawn/Depressed	8	.765	.010	.021	.048	.104	.214	.389	.592	.755	.853	.895	.889	.887	.897
			(1.0097)	(1.0218)	(1.0500)	(1.1162)	(1.2721)	(1.6359)	(2.4489)	(4.0826)	(6.7991)	(9.5497)	(8.9773)	(8.8811)	(9.7193)
Somatic Complaints	11	.677	.031	.053	.091	.153	.251	.394	.575	.749	.853	.872	.863	.890	.884
			(1.0321)	(1.0561)	(1.0997)	(1.1805)	(1.3353)	(1.6497)	(2.3517)	(3.9830)	(6.8158)	(7.8103)	(7.3099)	(9.1264)	(8.6198)
Social Problems	11	.746	.020	.036	.066	.775	.211	.354	.541	.732	.862	.909	.901	.906	.911
			(1.0199)	(1.0371)	(1.0703)	(1.1356)	(1.2675)	(1.5470)	(2.1805)	(3.7244)	(7.2407)	(10.9852)	(10.0993)	(10.6863)	11.1783
Thought Problems	15	.677	.027	.045	.079	.140	.243	.391	.558	.700	.798	.867	.904	.909	.916
			(1.0275)	(1.0475)	(1.0862)	(1.1634)	(1.3207)	(1.6412)	(2.2647)	(3.3360)	(4.9490)	(7.4909)	(10.4228)	11.0323	(11.9506)
Attention Problems	10	.852	.018	.040	.091	.201	.405	.683	.897	.938	.913	.947	.917	.875	.839
			(1.0182)	(1.0419)	(1.1004)	(1.2522)	(1.6793)	(3.1581)	(9.7237)	(16.1762)	(11.4510)	(18.8380)	(12.0549)	(8.0111)	(6.2087)
Rule-Breaking Behavior	17	.715	.010	.018	.032	.064	.141	.311	.579	.793	.868	.878	.913	.940	.944
			(1.0103)	(1.0179)	(1.0333)	(1.0688)	(1.1635)	(1.4516)	(2.3757)	(4.8371)	(7.5997)	(8.1925)	(11.5183)	(16.5493)	(17.9945)
Aggressive Behavior	18	.876	.012	.243	.084	.214	.451	.298	.848	.903	.926	.944	.955	.954	.947
			(1.0117)	(1.321)	(1.0920)	(1.2727)	(1.8199)	(3.3513)	(6.5684)	(10.3334)	(13.4458)	(17.7986)	(22.3005)	(21.7362)	(18.8987)
Internalizing Problems	32	.874	.096	.162	.268	.416	.586	.737	.841	.902	.933	.946	.951	.952	.951
			(1.1062)	(1.1938)	(1.3657)	(1.7123)	(2.4164)	(3.8028)	(6.3015)	(10.2012)	(14.9264)	(18.4754)	(20.2725)	(20.9856)	(20.3838)
Externalizing Problems	35	.897	.025	.055	.126	.274	.506	.735	.871	.925	.945	.958	.968	.970	.970
			(1.0254)	(1.0586)	(1.1443)	(1.3770)	(2.0256)	(3.7776)	(7.7467)	(13.388)5	(18.3015)	(23.9771)	(30.9540)	(33.8423)	(33.4850)
Total Problems[1]	103	.949	.192	.314	.478	.652	.800	.888	.938	.962	.975	.981	.984	.985	.985
			(1.2382)	(1.4585)	(1.9144)	(2.8772)	(4.9036)	(8.9608)	(16.1290)	(26.6085)	(39.3269)	(52.3180)	(62.3948)	(66.4372)	(67.7770)

N = 5,820. CBCL = child behavior checklist. α = Cronbach's alpha internal consistency reliability. r_{xx} = internal consistency reliability. *I* = Information (r_{xx} = 1 – 1/*I*). Red color font type indicates unacceptably low reliability for basic research (r_{xx} < .60). [1]*n* = 5,81

A)

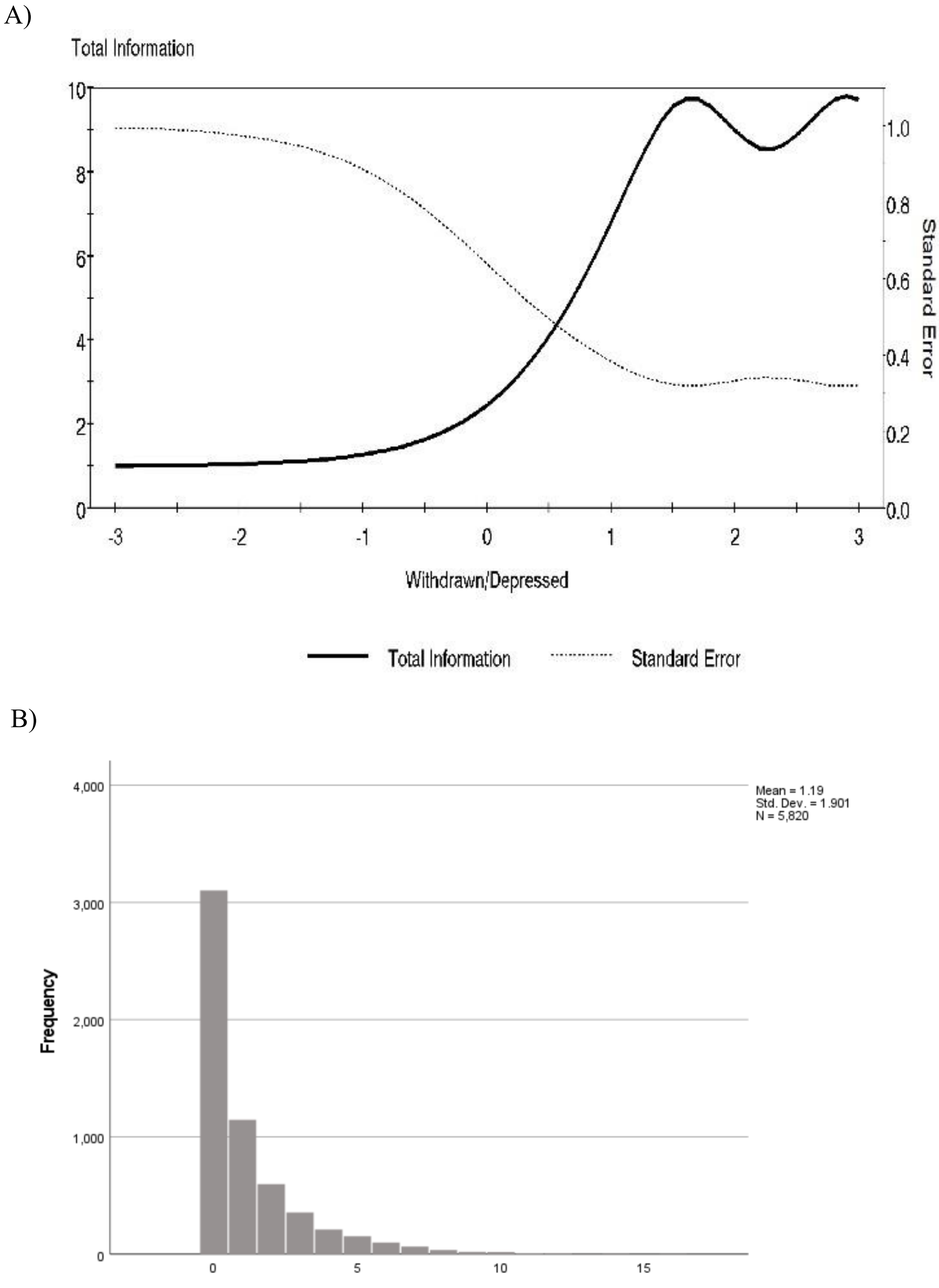


B)



Supplementary Figure 3. A) Total information function / curve (TIF/TIC) for the child behavior checklist Anxious/Depressed syndrome scale. B) Histogram of sum scale scores on the Anxious/Depressed syndrome scale.

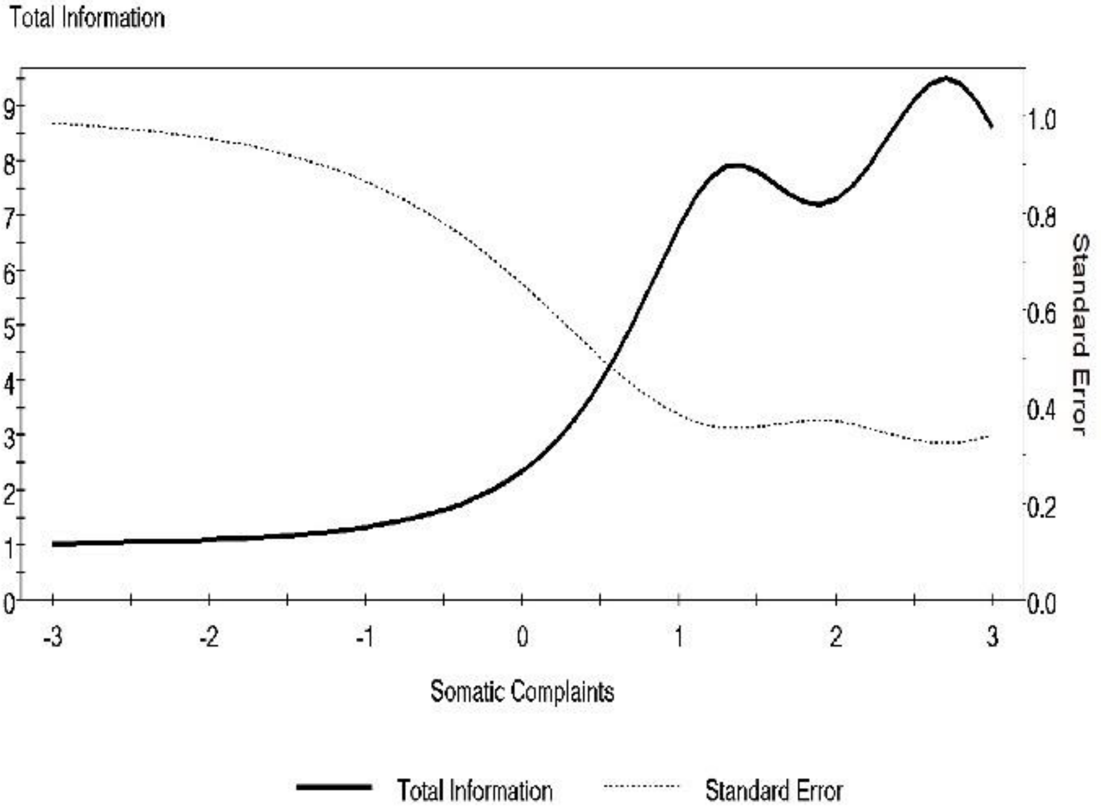
Note. N = 5,820. $r_{xx} = 1 - \left({}^{1}/_{I}\right)$. Standard error of the estimate (*SEE*) = $1/\sqrt{I}$.



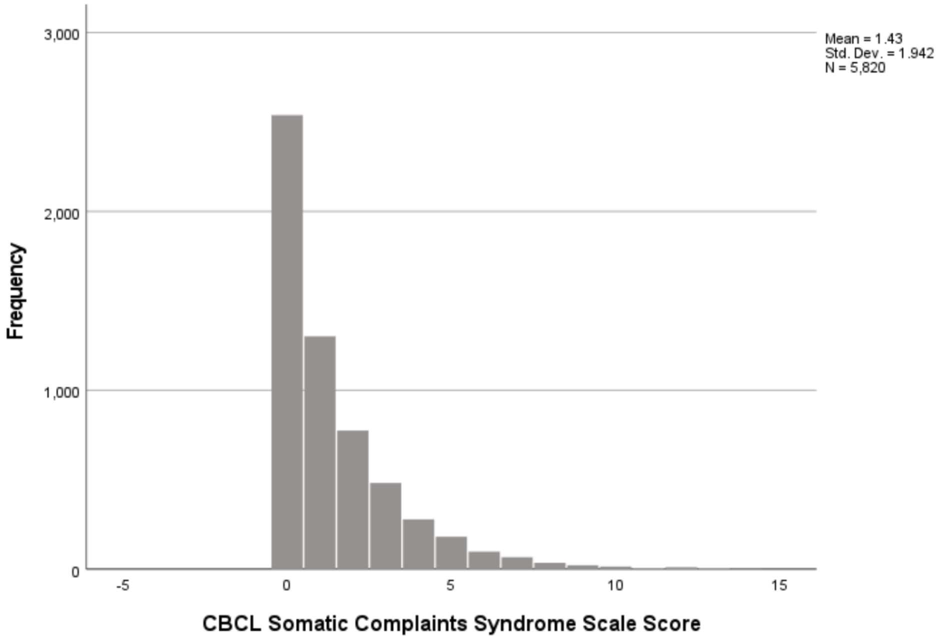
Supplementary Figure 4. A) Total information function / curve (TIF/TIC) for the child behavior checklist Withdrawn/Depressed syndrome scale. Taken from Tiego and Fornito (2022)[19]. Reprinted with permission. B) Histogram of sum scale scores on the Withdrawn/Depressed syndrome scale.

Note. N = 5,820. $r_{xx} = 1 - (1/I)$. Standard error of the estimate (*SEE*) = $1/\sqrt{I}$.

A)



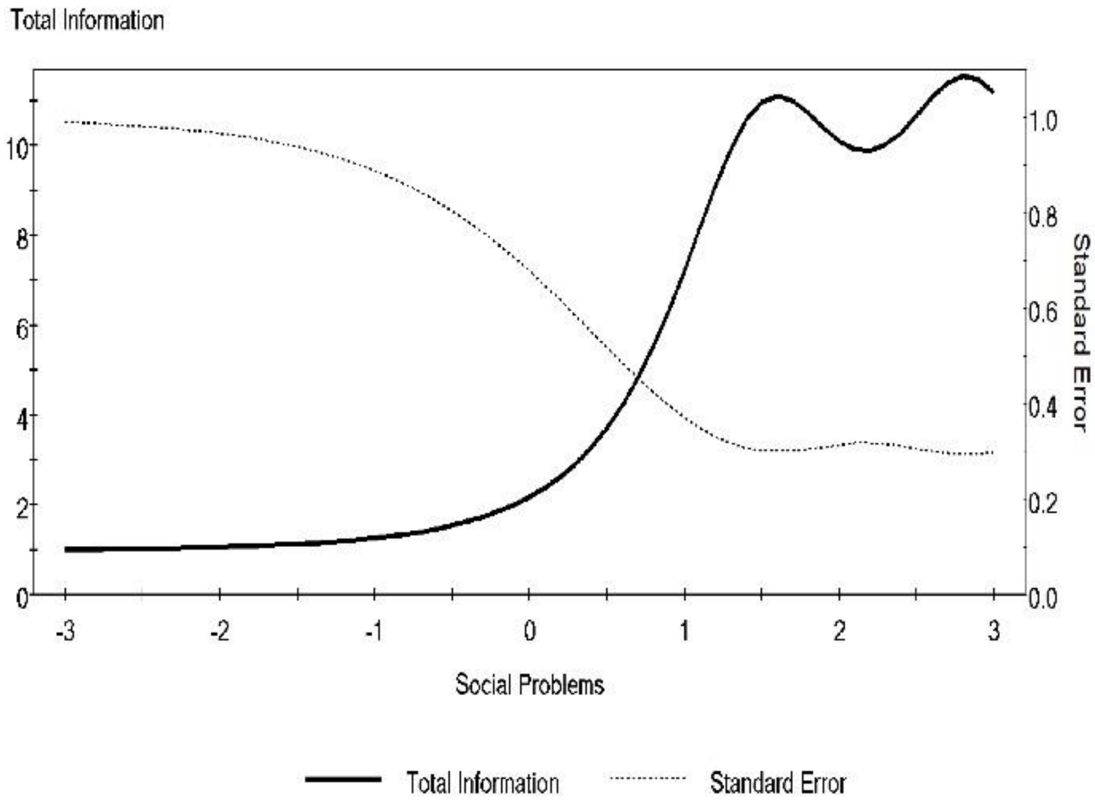
B)



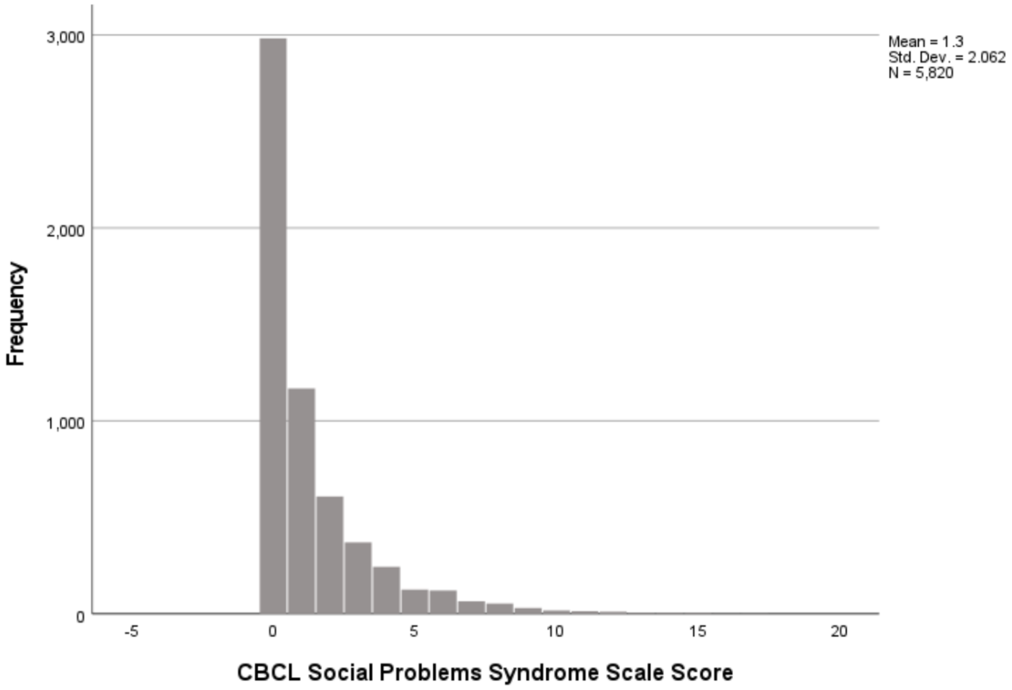
Supplementary Figure 5. A) Total information function / curve (TIF/TIC) for the child behavior checklist Somatic Complaints syndrome scale. B) Histogram of sum scale scores on the Somatic Complaints syndrome scale.

Note. N = 5,820. $r_{xx} = 1 - \left({}^{1}/_{I}\right)$. Standard error of the estimate (*SEE*) = $1/\sqrt{I}$.

A)



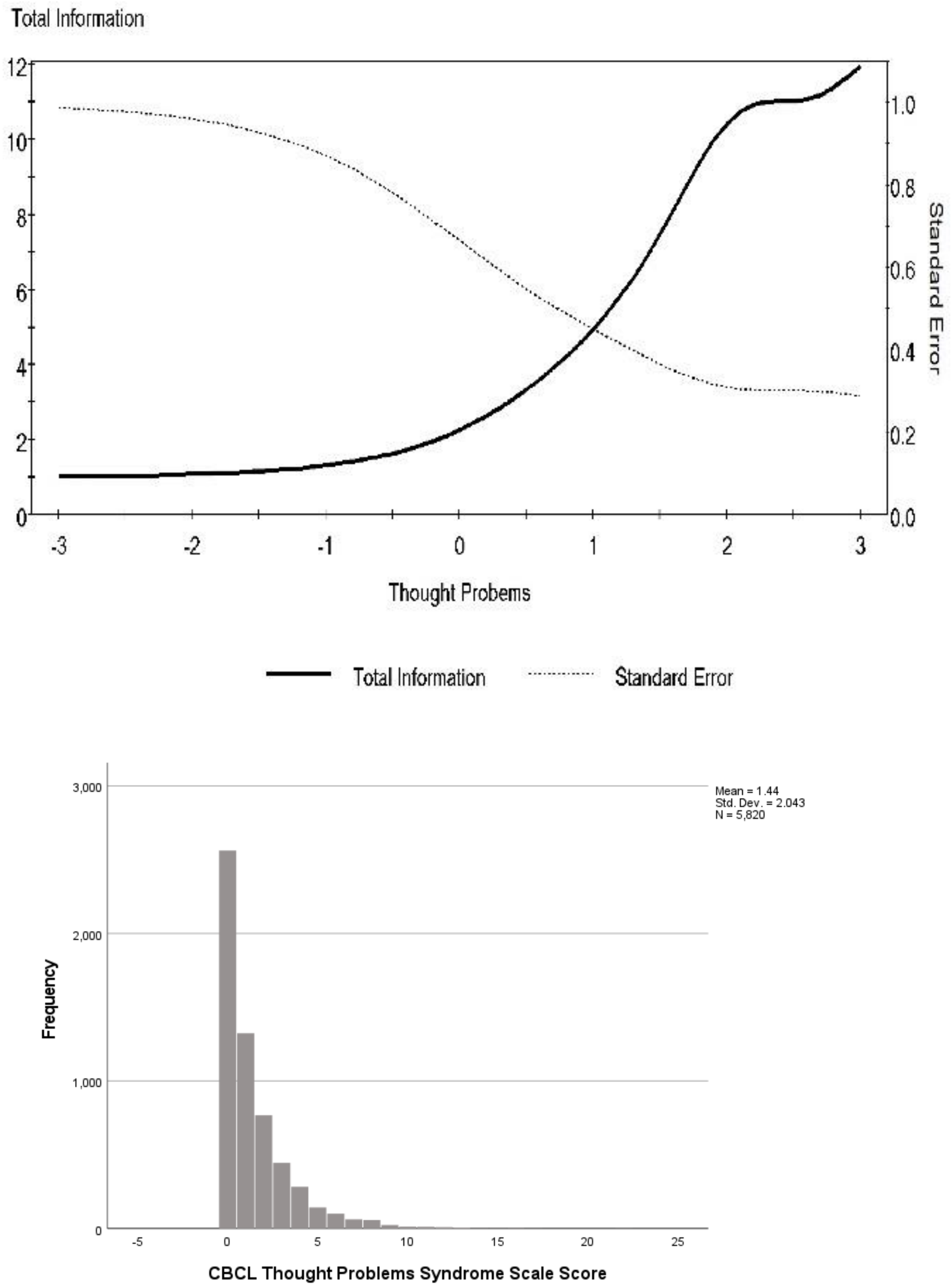
B)



Supplementary Figure 6. A) Total information function / curve (TIF/TIC) for the child behavior checklist Social Problems syndrome scale. B) Histogram of sum scale scores on the Social Problems syndrome scale.

Note. N = 5,820. $r_{xx} = 1 - \left({}^{1}/_{I}\right)$. Standard error of the estimate (*SEE*) = $1/\sqrt{I}$.

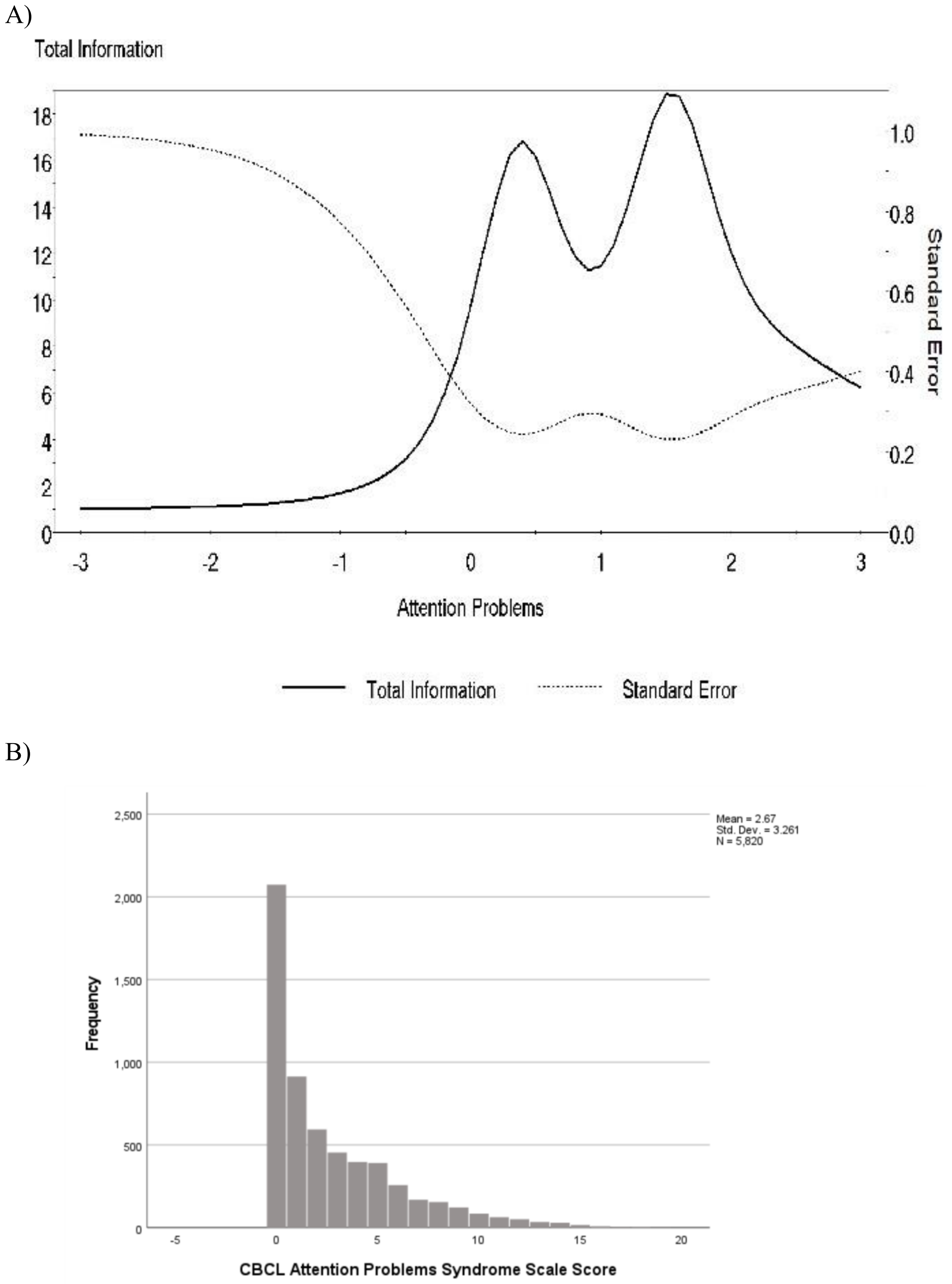
A)



B)

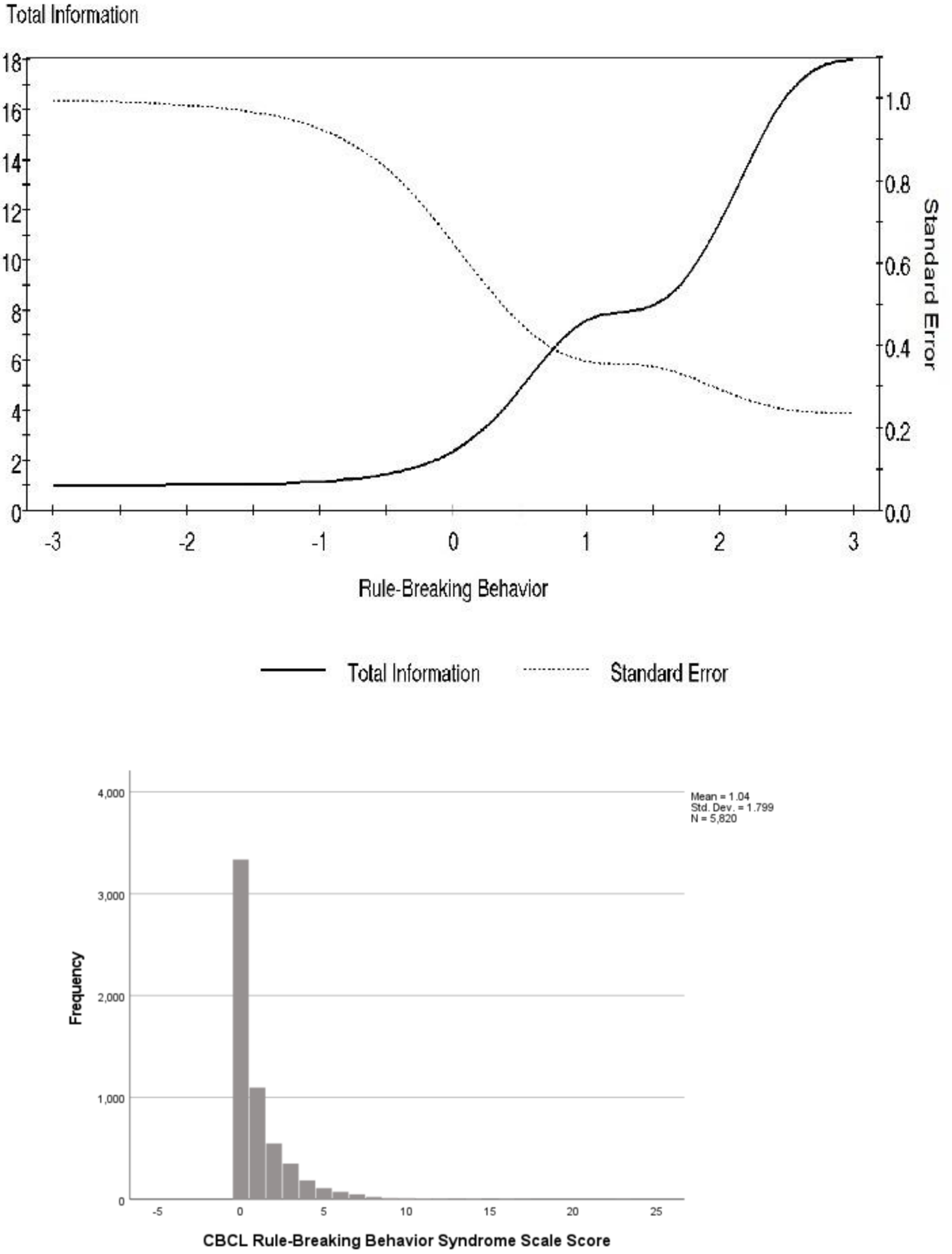
Supplementary Figure 7. A) Total information function / curve (TIF/TIC) for the child behavior checklist Thought Problems syndrome scale. B) Histogram of sum scale scores on the Thought Problems syndrome scale.

Note. N = 5,820. $r_{xx} = 1 - \left({}^{1}/_{I}\right)$. Standard error of the estimate (*SEE*) = $1/\sqrt{I}$.



Supplementary Figure 8. A) Total information function / curve (TIF/TIC) for the child behavior checklist Attention Problems syndrome scale. B) Histogram of sum scale scores on the Attention Problems syndrome scale.

Note. N = 5,820. $r_{xx} = 1 - \left(1/_{I}\right)$. Standard error of the estimate (*SEE*) = $1/\sqrt{I}$.

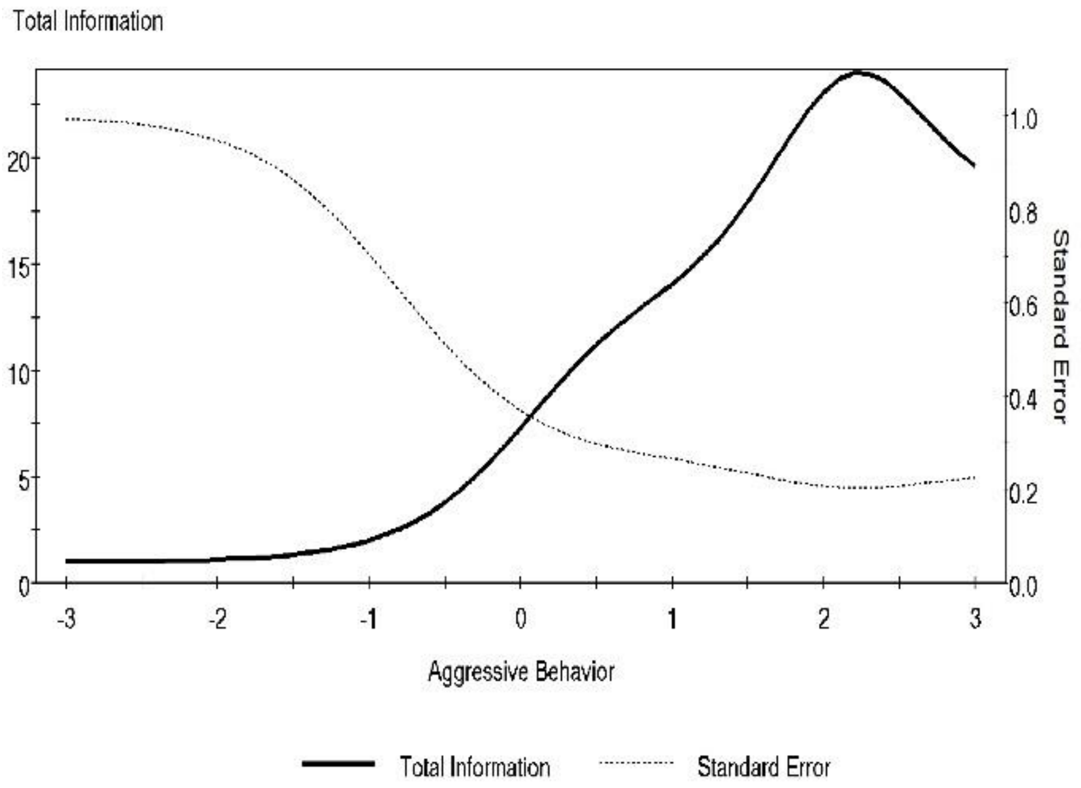

A)

B)

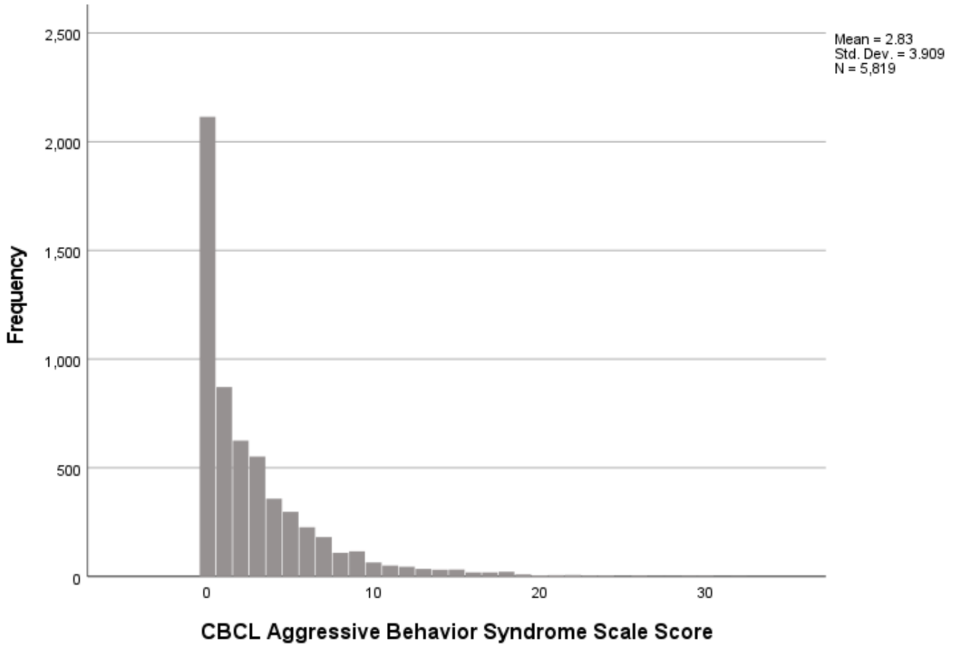
Supplementary Figure 9. A) Total information function / curve (TIF/TIC) for the child behavior checklist Rule-Breaking Behavior syndrome scale. B) Histogram of sum scale scores on the Rule-Breaking Behavior syndrome scale.

Note. N = 5,820. $r_{xx} = 1 - \left({}^{1}/_{I}\right)$. Standard error of the estimate (*SEE*) = $1/\sqrt{I}$.

A)

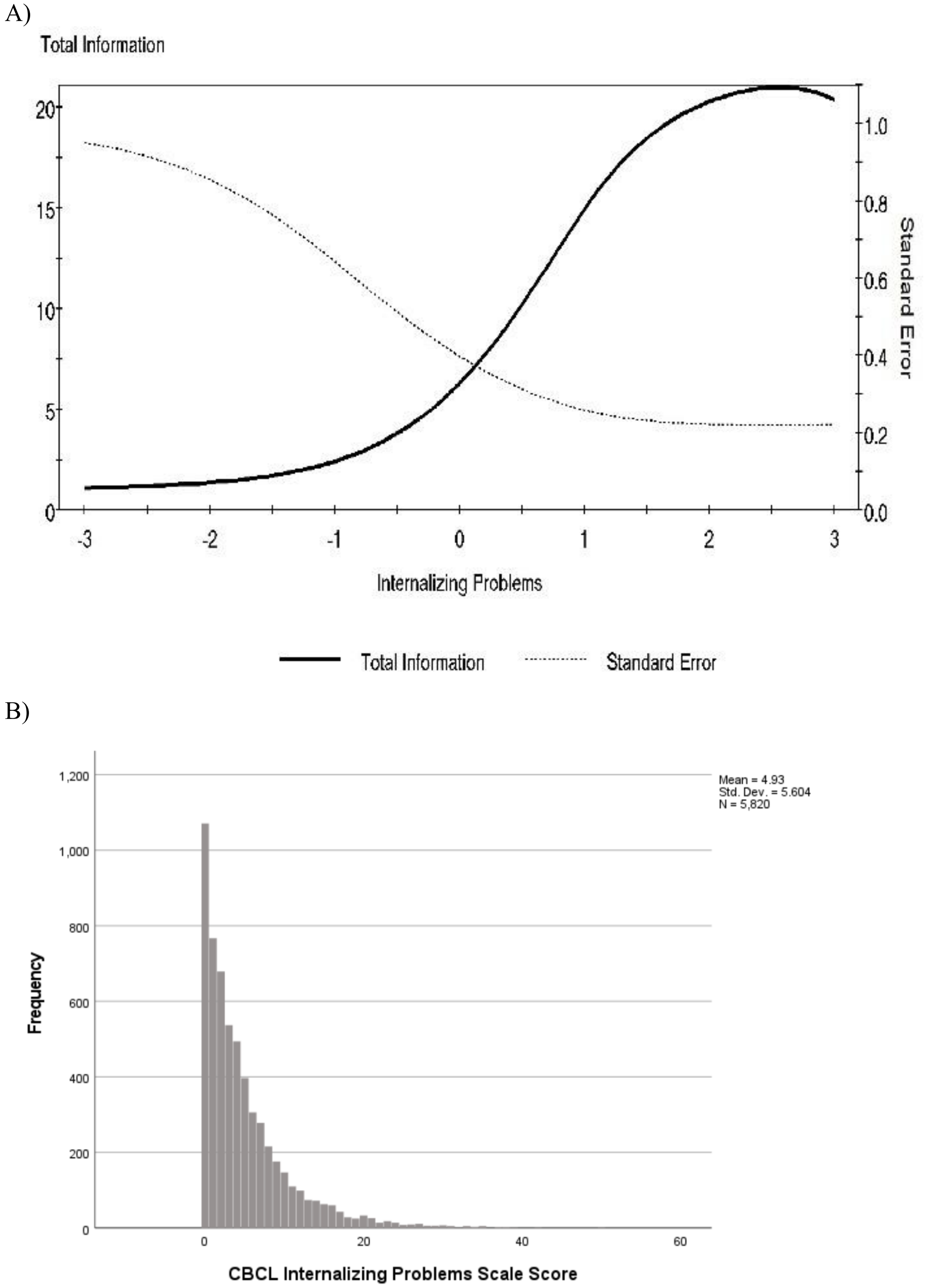


B)



Supplementary Figure 10. A) Total information function / curve (TIF/TIC) for the child behavior checklist Aggressive Behavior syndrome scale. B) Histogram of sum scale scores on the Aggressive Behavior syndrome scale.

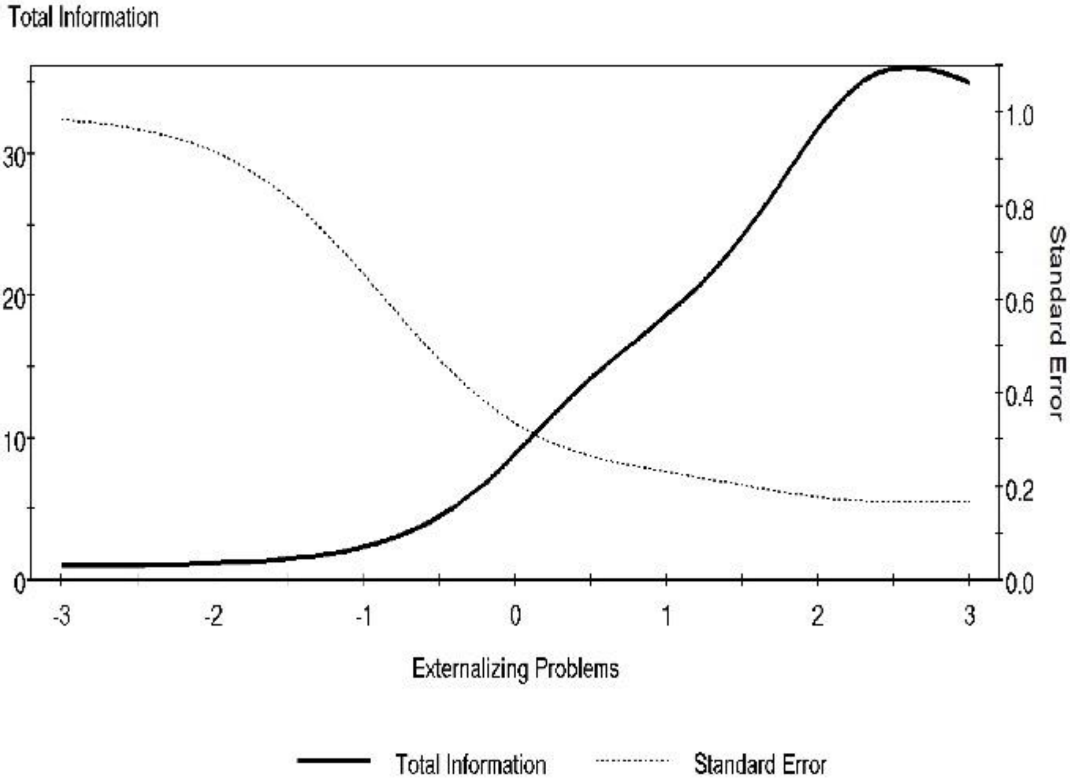
Note. N = 5,819. $r_{xx} = 1 - \left({}^{1}/_{I}\right)$. Standard error of the estimate (*SEE*) = 1/$\sqrt{I}$.



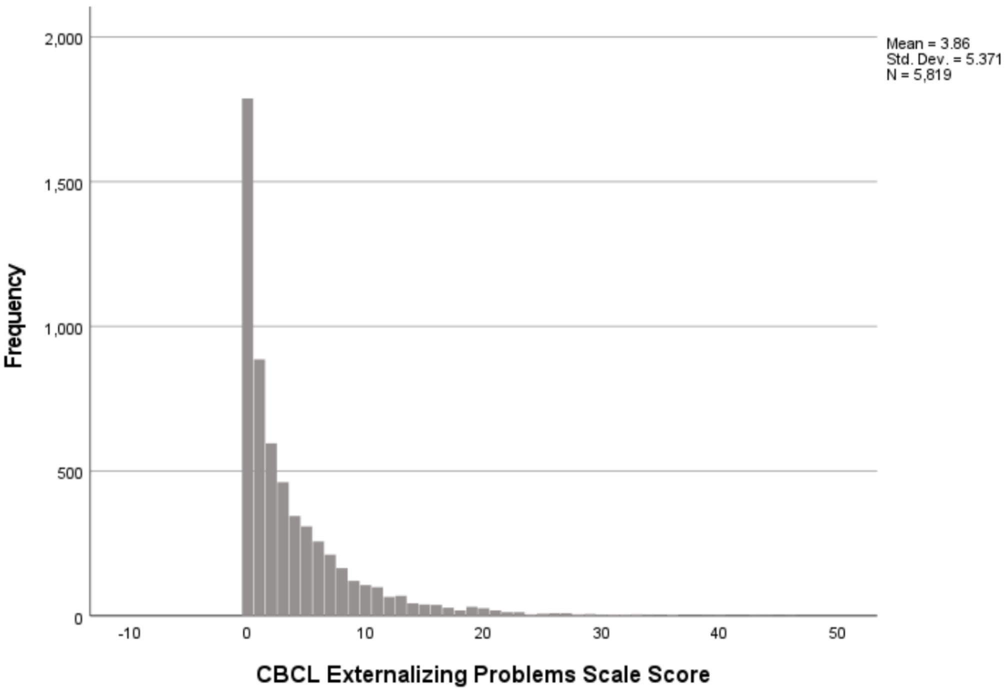
Supplementary Figure 11. A) Total information function / curve (TIF/TIC) for the child behavior checklist Internalizing Problems scale. B) Histogram of sum scale scores on the Internalizing Problems scale.

Note. N = 5,820. $r_{xx} = 1 - ({}^{1}/_{I})$. Standard error of the estimate (*SEE*) = $1/\sqrt{I}$.

A)

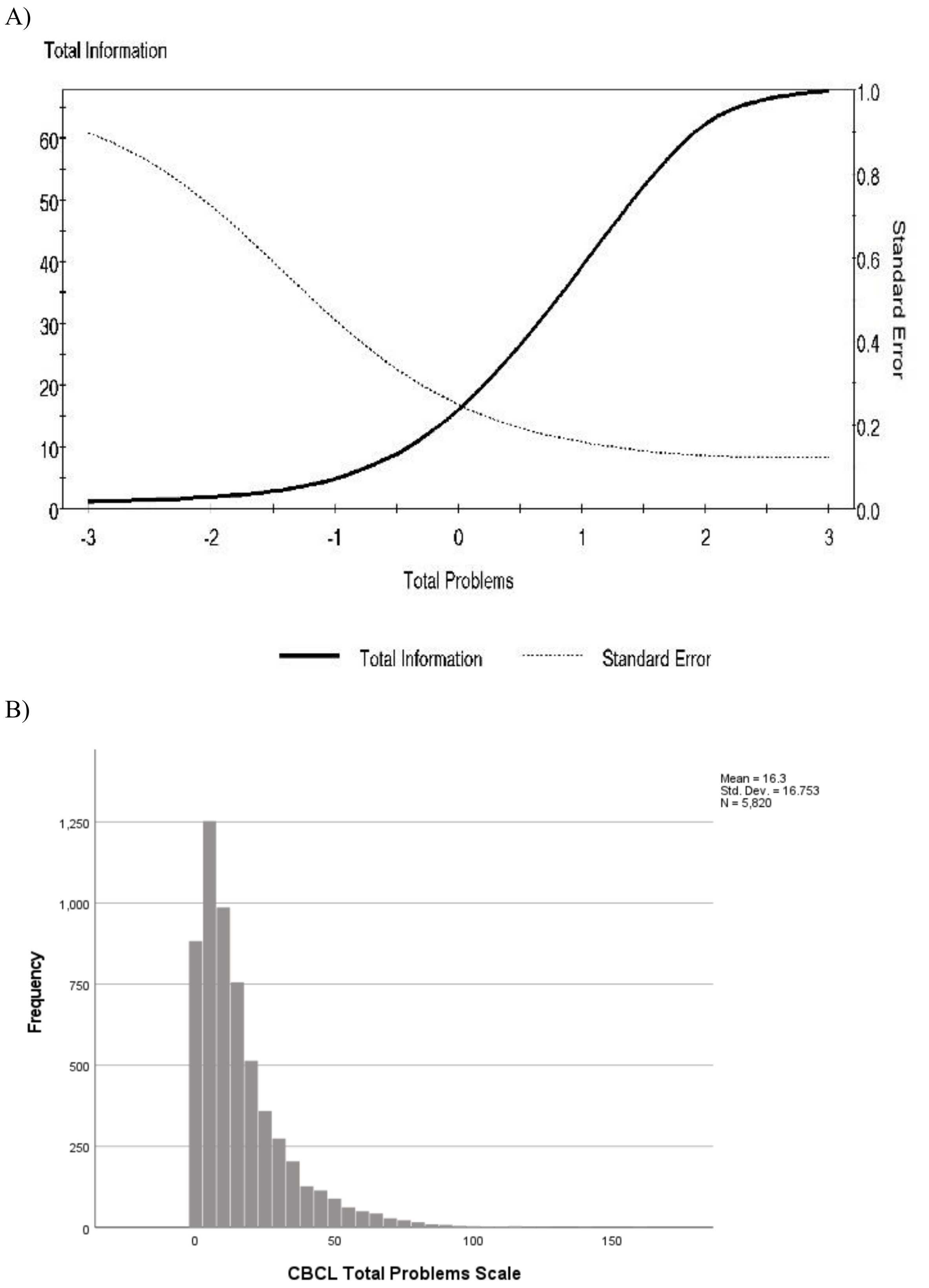


B)



Supplementary Figure 12. A) Total information function / curve (TIF/TIC) for the child behavior checklist Externalizing Problems scale. B) Histogram of sum scale scores on the Externalizing Problems scale.

Note. $N = 5{,}819$. $r_{xx} = 1 - (1/_I)$. Standard error of the estimate (*SEE*) = $1/\sqrt{I}$.



Supplementary Figure 13. A) Total information function / curve (TIF/TIC) for the child behavior checklist Total Problems scale. B) Histogram of sum scale scores on the Total Problems scale.

Note. N = 5,820. $r_{xx} = 1 - \left({}^{1}/{}_{I}\right)$. Standard error of the estimate (*SEE*) = $1/\sqrt{I}$.

Supplementary Table 2

Proportion of the Sample from the Two-Year Follow-Up Wave of Data Collection from the ABCD Study Cohort that Did Not Meet Minimal Acceptable Standards of Measurement Reliability on Each of the Eleven Child Behavior Checklist Scales

CBCL Scale	θ $I < 2.5$	Raw Score at $I < 2.5$	n $r_{xx} < .60$	%N $r_{xx} < .60$
Anxious/Depressed	-0.600	0.5329	1,985	34.1
Withdrawn/Depressed	0.000	0.5207	3,103	53.3
Somatic Complaints	0.000	0.8081	2,539	43.6
Social Problems	0.100	0.7009	2,983	51.3
Thought Problems	0.100	0.9287	2,563	44.0
Attention Problems	-0.700	0.3498	2,074	35.6
Rule-Breaking Behavior	0.000	0.3885	3,336	57.3
Aggressive Behavior	-0.800	0.2757	2,115	36.3
Internalizing Problems	-1.000	0.9160	1,071	18.4
Externalizing Problems	-0.900	0.3649	1,788	30.7
Total Problems	-1.700	0.9401	453	7.78

Note. N = 5,820. CBCL = Child behavior checklist. θ = latent trait continuum in standardized metric (i.e., M = 0, SD = 1). I = Information. n = size of subsample. r_{xx} = internal consistency reliability.

Example 3 - Measurement non-invariance.

Measurement invariance for questionnaires can also be evaluated within an IRT framework (Box 5 main text), where it is called differential item functioning (DIF)[27,28]. DIF refers to the property of a measurement instrument in which the item parameters estimated within an IRT framework differ as a function of group membership, such that there is bias in interpreting and comparing the raw scores between groups. When DIF is of sufficient magnitude across many items it can result in differential test functioning (DTF), by which scores cannot be meaningfully compared between groups because they correspond to different levels of the latent trait being measured[27-29]. This has serious implications for biology-psychopathology association studies, because psychometric and substantive group differences in observed scores may obscure meaningful associations with psychiatric biomarkers. It is worth mentioning that DIF can also be associated with latent classes or mixtures (see example 5), which represent unobserved groups that vary in their slope and threshold parameters (Box 5 main text). These differences can be detected using IRT mixture modeling[30-32].

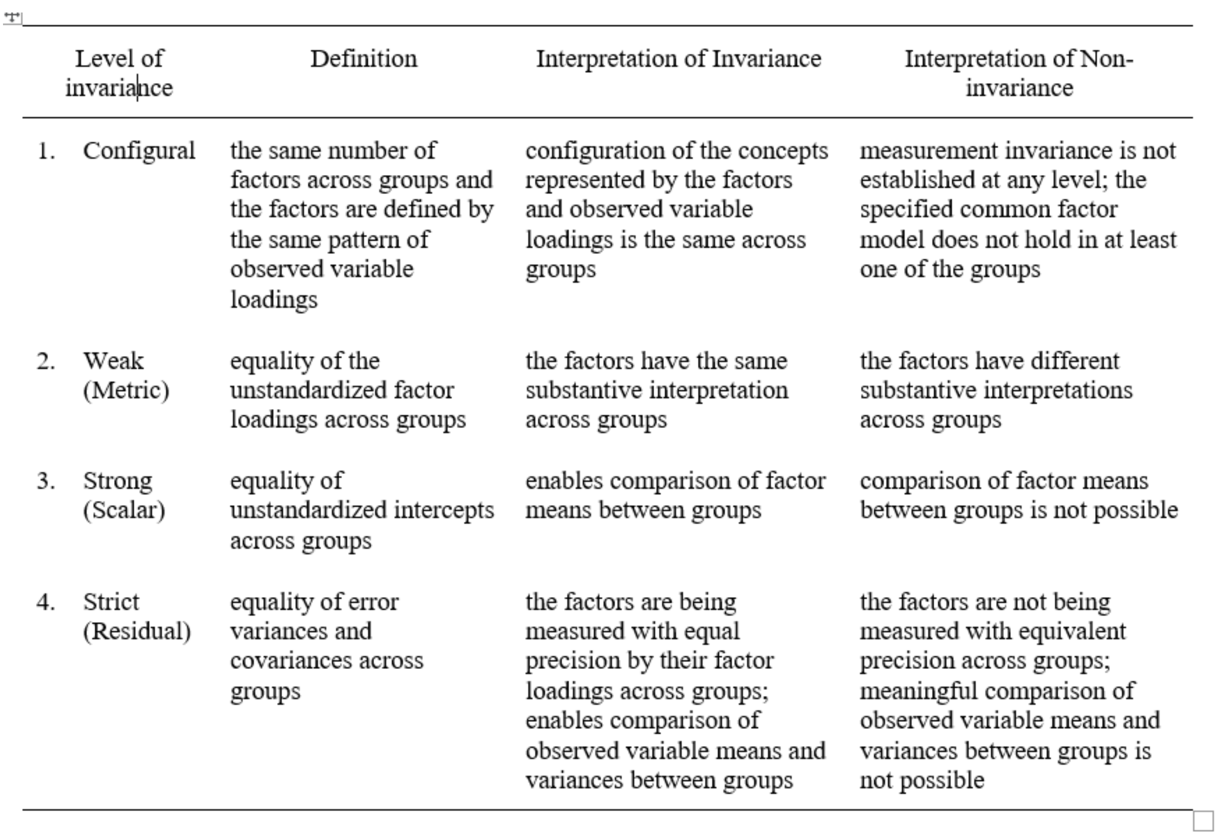
DIF assessment is an essential, but often overlooked, part of the validation process for psychiatric phenotypes[33]. DIF is a more powerful approach for detecting non-invariance than traditional factor analysis approaches, but requires larger sample sizes and more restrictive assumptions[34]. There are multiple approaches to DIF testing, but the preferred method when equivalence between any items has not yet been established is to use an iterative two-step procedure[35]. Here, all items are anchored to a common metric (i.e., all items scaled to the same latent trait distribution) and their slope and threshold parameters freely estimated one at a time. The difference in model fit is tested for statistical significance using the Wald χ^2 test[35]. Each item is tested for statistically significant group differences in slope and threshold parameters, as well as overall DIF (slope and threshold parameters) using the χ^2 test statistic with corresponding degrees of freedom (*df*). Differences in the threshold (severity/location) parameters indicate that item response categories are differentially sensitive to different levels of the latent trait between groups[29]. Statistically significant differences in slope parameters indicate that questionnaire items provide different degrees of information and precision of measurement across groups[29].

By way of example, we tested for DIF in the Total Problems scale of the CBCL for male and female ABCD participants using the two-year follow-up data. We focused on the Total Problems scale because it has the highest reliability of all the CBCL scales as indexed by Cronbach's α and information values across the latent trait continuum (Supplementary Table 1). We evaluated item-level performance prior to overall model fit[23,36]. The monotonicity assumption was assessed by inspecting the option response functions and ensuring that the probability of endorsement of each successive response category on CBCL items increased monotonically as a function of increasing severity on the CBCL total problems latent trait continuum[23]. We removed three items (72, 105, 106) with substantially elevated standard errors for their threshold parameters in males and females, suggesting poor fit of the model. The fit of the graded response (GR) model to each item was assessed with a generalization of the S-χ^2 item-fit statistic[37] at a lower significance threshold to account for the very large sample [$p < .001$]. No items demonstrated poor fit to the GR model based on this probability threshold. Many items demonstrated local dependence (LD) based on exceeding the recommended threshold for the standardized LD χ^2 statistics [i.e., > 10][38]. However, there was good reason to believe that these inflated LD statistics and apparent local dependencies between items were attributable to the large number of zero-frequency cells in the bivariate contingency tables[39] for the CBCL data, which is common for clinical scales with low endorsement rates resulting in sparseness of the observed data[23]. For this reason, we retained all remaining items regardless of whether they had elevated LD ($\chi^2 > 10$).

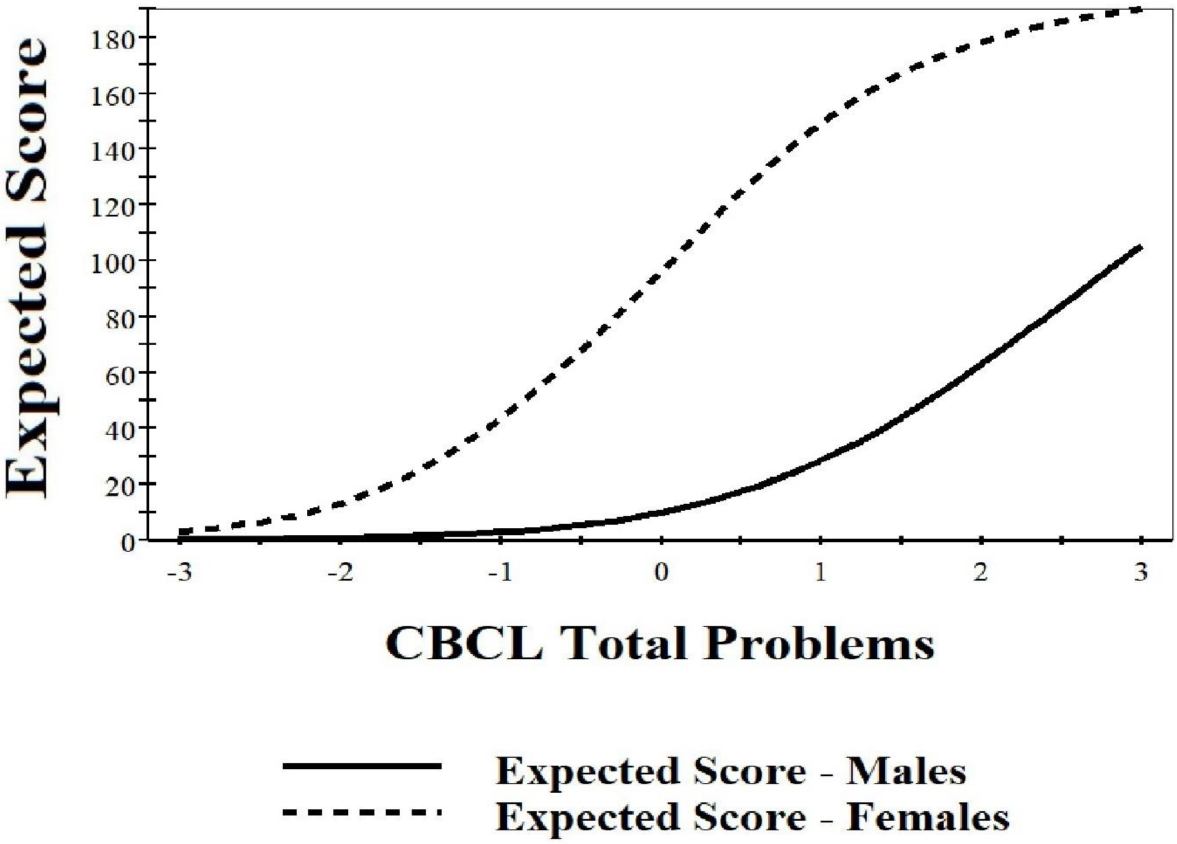
We determined substantial DIF between the sexes, such that there was evidence of DTF as can be seen in the test characteristic curves displayed in Supplementary Figure 14. Test characteristic curves plot the expected raw score for a group (*y* axis) as a function of their values on the underlying latent trait continuum (*x* axis)[22,29,40]. As can be seen in Supplementary Figure 14, the test characteristic curves were not coincident at any point along the latent trait continuum, indicating DTF. In other words, raw scores on the CBCL Total Problems scale cannot be directly compared between male and female children, because they correspond to different levels of the underlying Total Problems latent trait. For example, a raw score of 10 in males (equivalent to the mean of the latent trait) does not index the same level of severity in the underlying latent trait construct as it does in females (roughly equivalent to two standard deviations below the mean of the latent trait). These differences will confound any analysis that pools scores for males and females. The differences observed here are substantial and would confound any attempts to correlate this measure with biological variables that are pooled for male and female children.

Supplementary Table 3

Levels of Measurement Invariance Typically Evaluated within a Factor Analytic Framework for Continuous Indicators



Level of invariance	Definition	Interpretation of Invariance	Interpretation of Non-invariance
1. Configural	the same number of factors across groups and the factors are defined by the same pattern of observed variable loadings	configuration of the concepts represented by the factors and observed variable loadings is the same across groups	measurement invariance is not established at any level; the specified common factor model does not hold in at least one of the groups
2. Weak (Metric)	equality of the unstandardized factor loadings across groups	the factors have the same substantive interpretation across groups	the factors have different substantive interpretations across groups
3. Strong (Scalar)	equality of unstandardized intercepts across groups	enables comparison of factor means between groups	comparison of factor means between groups is not possible
4. Strict (Residual)	equality of error variances and covariances across groups	the factors are being measured with equal precision by their factor loadings across groups; enables comparison of observed variable means and variances between groups	the factors are not being measured with equivalent precision across groups; meaningful comparison of observed variable means and variances between groups is not possible



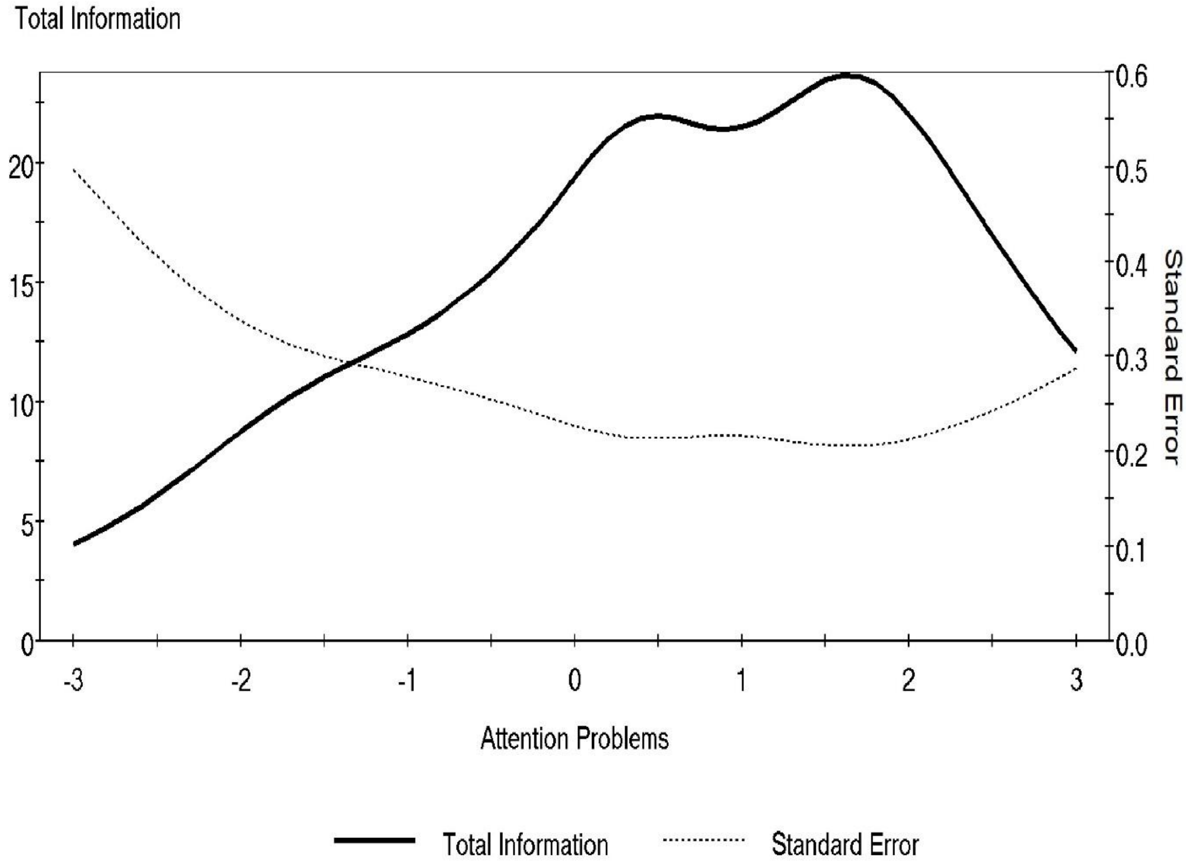
Supplementary Figure 14. Test characteristic curves showing the relationship of expected raw score (*y* axis) as a function of a participants' standing on the CBCL Total Problems latent trait continuum (*x* axis) for males (*n* = 3,025) and females (*n* = 2,795).

Image taken from Tiego and Fornito (2022)[19]. Reprinted with permission.

Example 4 – Increasing phenotypic resolution

Although attention deficit hyperactivity (ADHD)-related problems are dimensionally distributed in the developmental population[41], the Attention Problems scale, along with many other CBCL scales, are strongly positive skewed[6,25]. This is due to the fact that the CBCL was developed for maximal criterion-validity in differentiating referred from non-referred youth (i.e., using empirical criterion-keying)[25]. Thus, subscale items index symptoms that are only relevant for a small proportion of children with clinically-significant attention problems. As a result, there will be high precision of measurement at the upper end of the Attention Problems latent trait continuum where there is adequate item coverage, but very poor precision at the adaptive end of the continuum where attentional functioning is normal or even better than normal (Supplementary Table 1 & Supplementary Figure 8)[42].

Along with the CBCL, parents/guardians of child study participants in the ABCD study also completed the Early Adolescent Temperament Questionnaire – Revised (EATQ-R).[43] The EATQ-R measures the three higher-order dimensions of temperament: negative affectivity, positive affectivity, and effortful control (i.e., constraint). Effortful control is the self-regulatory domain of temperament (i.e., the developmental precursor of conscientiousness) and constitutes a protective factor against developmental psychopathology, especially disinhibited externalizing problems such as ADHD [44-47]. Thus, it stands to reason that high effortful control (i.e., high attentional control) represents the adaptive end of the attention problems continuum. We reran the latent trait model with IRT on the CBCL Attention Problems syndrome scale items incorporating the Effortful Control subscale items of the EATQ-R. The total information function is displayed in Supplementary Figure 15 and shows that measurement precision was markedly increased, with marginal reliability at r_{xx} = .94 and reliability not dropping below r_{xx} = .75 even at three standard deviations below the mean. However, inclusion of additional items must meet the assumptions of unidimensional IRT, including unidimensionality and fit of item data to the (two parameter logistic or graded response) IRT model.[23]



Supplementary Figure 15. Total information curve for the Attention Problems syndrome scale incorporating Effortful Control items from Early Temperament Questionnaire – Revised in 5,823 participants from the ABCD study. Marginal reliability estimate is r_{xx} = 0.94 and reliability does not decrease below r_{xx} = 0.75 even at -3*SD*.

Example 5 – Investigating sample heterogeneity with mixture modeling

One area of psychiatric research in which biological and etiological heterogeneity has been increasingly recognized and accommodated is in the study of attention deficit hyperactivity disorder (ADHD)[48-51]. Attempting to explicitly account for heterogenous subtypes has led to the discovery of unique neuroimaging biomarkers[52,53]. In line with these findings and by way of example, we conducted a factor mixture modeling (FMM) analysis of the attention problems syndrome scale of the CBCL in the two-year follow-up wave of data of the ABCD study cohort. FMM is a type of latent variable analysis that combines latent class analysis (LCA) with the common factor modeling (CFM) approach[54-56], and can be used for identifying discrete, or even probabilistic, classes (also "mixtures" or clinical subtypes/subgroups) that are latent (i.e., not directly observed) and embedded within multivariate dimensional data. FMM is particularly useful for analyzing zero-inflated data, which is characteristic of clinical phenomena measured in non-clinical samples[57]. Zero-inflated distributions can compromise correlational studies by violating distributional assumptions and attenuating linear relationships[57,58]. In these cases, FMM identifies individuals with little-to-no symptoms (i.e., a zero-inflated class) and distinguishes them from the rest of the distribution, resulting in differentiation into distinct sub-groups.

We first confirmed that the attention problems construct was unidimensional (i.e., absence of variable-centred heterogeneity) and identified the best-fitting model in the ABCD sample using Bayesian structural equation modelling (SEM). We conducted a thorough sensitivity analysis by varying the priors for the factor loadings and residual covariances (Supplementary Figure 16 & Supplementary Table 4)[59,60]. We then conduced LCA to determine the upper bound on the number of potential classes that could be embedded within the data[54]. We determined that five classes based on item response patterns could be discerned as the best fitting categorical latent class model (see Supplementary Table 6) and the upper bound for the number of FMM subtypes that would best account for the data (i.e., because FMM takes into account the factor structure and dimensionality of the data, as well as the categorical structure of person-centred subtypes, the number of classes best accounting for the data is less than that determined by LCA).

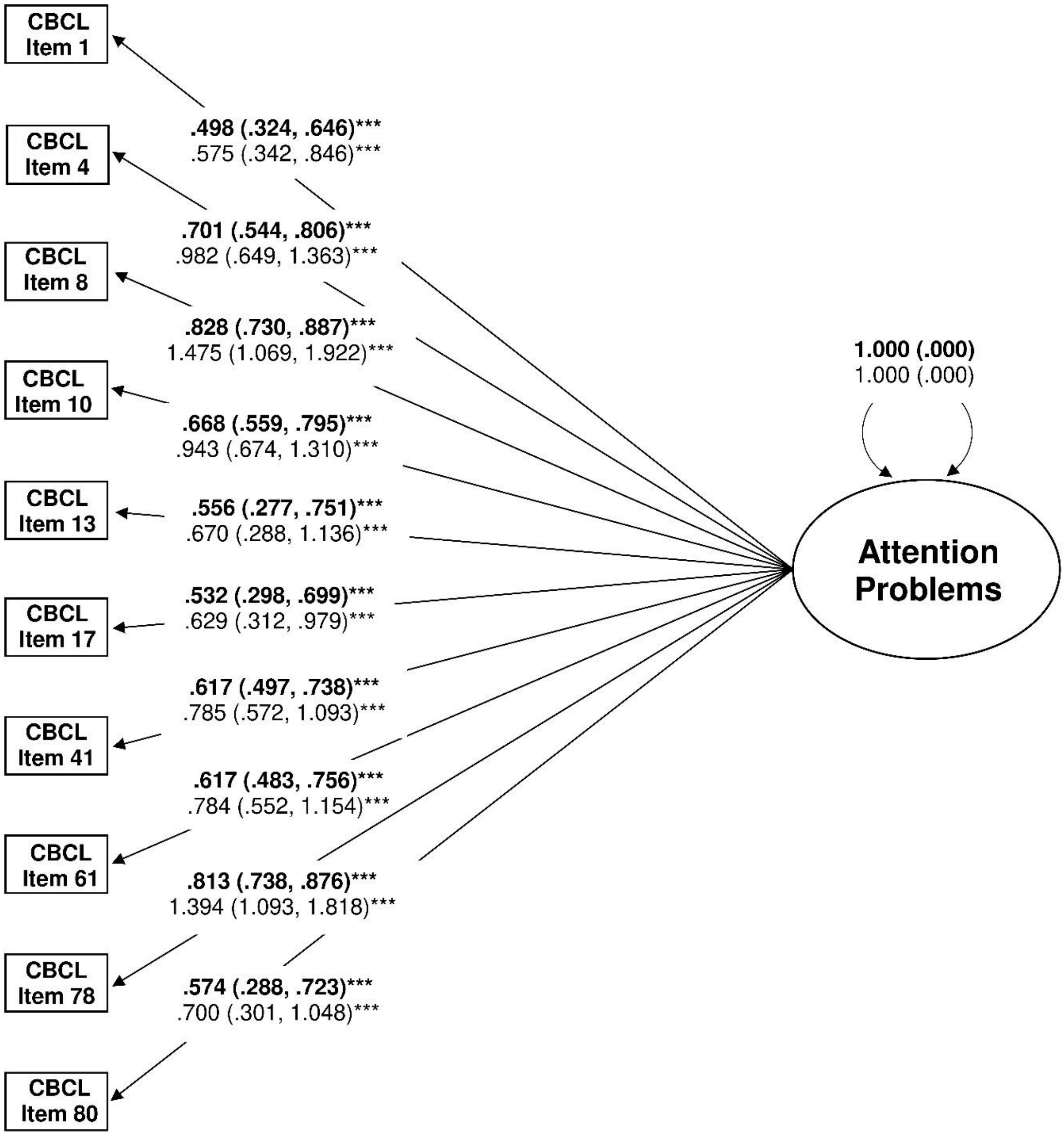
We then began testing FMMs, beginning with the simplest, a one-factor one-class model[54], before moving to one-factor two-class models using the most restrictive and parsimonious FMM (i.e., FMM-1, different latent means only) before progressively relaxing equality constraints on the factor variance-covariance matrix (i.e., FMM-2); the item thresholds (i.e., FMM-3), and the factor loadings (i.e., FMM-4), as well as specifying zero-inflated FMM models for the ≥ two-class models, to determine the best fitting model as indicated by the log likelihoods (lower is better), entropy (ranges between 0.000 – 1.000, with higher values indicating better class separation), and the Bayesian information criterion (BIC; lower values denoting the preferred model)[54]. We found that a two-class, one-factor model FMM-3 provided the best fit to the data as revealed by the BIC and better class separation than the three-class one-factor zero-inflated FMM-3, which was little better than chance class assignment (see Supplementary Table 7). Although class separation was poor for the two-class, one-factor FMM-3 model as shown by the low entropy, these two classes demonstrated distinct item response profiles (Supplementary Figures 17 – 26) with the smaller class 2 (*n* = 853, 14.66%) endorsing more severe symptoms on seven of the ten items (1 “acts young”; 4 “fails to finish”; 8 “concentrate”; 10 “sit still”; 41 “impulsive”; 61 “poor school”; 78 “inattentive”) than the bigger class 1 (*n* = 4,967, 85.34%). Thus, whilst the latent variable variables have a similar interpretation across classes due to the same pattern of factor loadings, they have different variances, and neither latent means nor raw scores can be directly and meaningfully compared due to class varying thresholds (i.e., systematic differences in item response category endorsement unrelated to the latent variable)[54]. Failure to check for and identify these mixtures may confound subsequent biology-psychopathology associations studies. As class separation was poor based on the entropy (*E* = .614), covariates (e.g. biological variables) would need to be compared across classes by including them as auxiliary variables and using the DCAT or BCH procedures as implemented in Mplus[61] for categorial and continuous variables, respectively[62,63]. This method avoids biased estimates in class comparisons, whilst preserving uncertainty in class membership without causing shifts in latent classes[64].

Supplementary Table 4

Summary of Fit Statistics for Competing Bayesian Confirmatory Factor Analysis Models for the ASRS-5 in the Adult ADHD Cohort

	Model*	95%*CI* $\Delta\chi^2$		PPP	Prior PPP
		LL	*UL*		
1	One-factor model factor loading priors N(0.90,.100), residual covariances priors IW(5,10)	-30.183	32.444	.483	.990
2	One-factor model factor loading priors N(0.90,.050), residual covariances priors IW(5,10)	-30.104	32.330	.478	.989
3	One-factor model factor loading priors N(0.80,.100), residual covariances priors IW(5,10)	-29.989	32.313	.477	.989
4	One-factor model factor loading priors N(0.80,.050), residual covariances priors IW(5,10)	-29.893	32.549	.484	.986
5	One-factor model factor loading priors N(0.70,.100), residual covariances priors IW(5,10)	-30.070	32.948	.482	.990
6	One-factor model factor loading priors N(0.70,.050), residual covariances priors IW(5,10)	-29.712	32.955	.474	.988
7	One-factor model factor loading priors N(0.60,.100), residual covariances priors IW(5,10)	-29.774	32.790	.477	.994
8	One-factor model factor loading priors N(0.60,.050), residual covariances priors IW(5,10)	-29.912	32.102	.482	.989
9	One-factor model factor loading priors N(0.50,.100), residual covariances priors IW(5,10)	-28.719	32.727	.473	.994
10	One-factor model factor loading priors N(0.50,.050), residual covariances priors IW(5,10)	-29.422	32.366	.482	.991
11	One-factor model factor loading priors N(0.90,.100), residual covariances priors IW(3,10)	-30.927	31.909	.483	.991
12	One-factor model factor loading priors N(0.90,.050), residual covariances priors IW(3,10)	-30.085	32.495	.482	.988
13	One-factor model factor loading priors N(0.80,.100), residual covariances priors IW(3,10)	-29.545	32.141	.487	.988
14	One-factor model factor loading priors N(0.80,.050), residual covariances priors IW(3,10)	-30.203	31.916	.484	.986
15	**One-factor model factor loading priors N(0.70,.100), residual covariances priors IW(3,10)**	**-30.080**	**33.170**	**.489**	**.990**
16	One-factor model factor loading priors N(0.70,.050), residual covariances priors IW(3,10)	-30.008	32.398	.479	.989
17	One-factor model factor loading priors N(0.60,.100), residual covariances priors IW(3,10)	-30.238	33.001	.474	.994
18	One-factor model factor loading priors N(0.60,.050), residual covariances priors IW(3,10)	-29.078	32.726	.472	.989
19	One-factor model factor loading priors N(0.90,.100), residual covariances priors IW(1,10)	-30.516	32.576	.483	.990
20	One-factor model factor loading priors N(0.90,.050), residual covariances priors IW(1,10)	-30.583	32.058	.481	.988
21	One-factor model factor loading priors N(0.80,.100), residual covariances priors IW(1,10)	-30.639	32.554	.484	.988
22	One-factor model factor loading priors N(0.80,.050), residual covariances priors IW(1,10)	- 30.344	32.701	.479	.986
23	One-factor model factor loading priors N(0.70,.100), residual covariances priors IW(1,10)	-30.133	32.877	.482	.991
24	One-factor model factor loading priors N(0.70,.050), residual covariances priors IW(1,10)	-29.524	32.921	.472	.987
25	One-factor model factor loading priors N(0.60,.100), residual covariances priors IW(1,10)	-29.819	32.227	.479	.994
26	One-factor model factor loading priors N(0.60,.050), residual covariances priors IW(1,10)	-29.154	33.052	.471	.989

Note. number of free parameters = 75; $\Delta\chi^2$ = 95% confidence interval for the difference between the observed and replicated chi-square values. PPP = posterior predictive probability value. Prior PPP = prior posterior predictive probability value. *All models used default normal priors for the item thresholds ~N(0.00,5.00). Base model with no priors for the factor loadings or error covariances failed to converge. Bold typeface denotes best fitting model. (*N* = 5,820).



Supplementary Figure 16. One-factor model of CBCL attention problems empirical syndrome scale in the two-year follow-up wave of data collection of the ABCD study (N = 5,820).

Note. Model fit statistics were q = 75; 95%CI $\Delta\chi^2$ = -30.080, 33.170; PPP = 0.489; Prior PPP = 0.990. Freely estimated residual covariances omitted for clarity (see Table S5).

Supplementary Table 5

Standardized Residual Covariances Between CBCL Attention Problems Items in the Best-Fitting Bayesian One-Factor Model

Variables	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. CBCL 1									
2. CBCL 4	0.209** (0.04, 0.375)								
3. CBCL 8	0.223 (-0.022, 0.495)	0.388*** (0.156, 0.596)							
4. CBCL 10	0.200* (0.012, 0.367)	0.142 (-0.100, 0.331)	0.470*** (0.244, 0.635)						
5. CBCL 13	0.219 (-0.007, 0.374)	0.217 (-0.077, 0.474)	0.286 (-0.231, 0.631)	0.101 (-0.172, 0.357)					
6. CBCL 17	0.158 (-0.002, 0.301)	0.276** (0.052, 0.489)	0.261 (-0.099, 0.575)	0.088 (-0.168, 0.312)	0.458*** (0.253, 0.600)				
7. CBCL 41	0.263** (0.094, 0.396)	0.277** (0.095, 0.419)	0.310*** (0.126, 0.491)	0.416*** (0.243, 0.534)	0.138 (-0.063, 0.346)	0.165 (-0.044, 0.346)			
8. CBCL 61	0.170* (0.006, 0.310)	0.388*** (0.186, 0.521)	0.361* (0.022, 0.539)	0.095 (-0.103, 0.248)	0.211 (-0.074, 0.415)	0.114 (-0.063, 0.318)	0.225** (0.070, 0.352)		
9. CBCL 78	0.196 (-0.031, 0.442)	0.370*** (0.171, 0.575)	0.648*** (0.504, 0.741)	0.369** (0.107, 0.529)	0.263 (-0.148, 0.596)	0.347** (0.029, 0.627)	0.421*** (0.239, 0.574)	0.356** (0.102, 0.526)	
10. CBCL 80	0.177 (-0.002, 0.332)	0.228 (-0.021, 0.482)	0.226 (-0.171, 0.599)	0.098 (-0.158, 0.363)	0.543*** (0.351, 0.681)	0.493*** (0.305, 0.628)	0.187 (-0.004, 0.394)	0.190 (-0.035, 0.398)	.320 (0.039, .651)

Note. 95% credibility intervals in brackets. *** one-tailed *p* < .001; ** one-tailed *p* < .01; *one-tailed *p* < .025.

Supplementary Table 6

Results of Exploratory Latent Class Analysis of the CBCL Attention Problems Empirical Syndrome Scale in the Two-Year Follow-Up Wave of Data from the ABCD Study

			Likelihood Ratio Δ^2				Lo-Mendell-Rubin Likelihood Ratio Test [3]		Bootstrapped Likelihood Ratio Test [3,4]		
C	*q*	*LL*	LR Δ^2 *df*	LR Δ^2	LR Δ^2 *p*	*E*	LMR	LMR *p*	2 *Δ*LL*	BLRT *p*	BIC
1 [1]	20	-34,859.934	58,621	10919.886	1.000						69893.249
2 [1]	41	-27,822.391	58,848	5169.178	1.000	.893	12028.061	<.001	12094.131	<.001	57981.168
3 [1]	62	-26,456.642	58,910	4311.943	1.000	.885	2534.377	<.001	2548.298	<.001	55614.920
4 [1]	83	-23,888.045	58,889	3989.107	1.000	.814	338.514	.011	340.373	<.001	55456.597
5 [1]	**104**	**-23,756.006**	**58,888**	**3965.017**	**1.000**	**.864**	**248.922**	**.046**	**250.289**	**<.001**	**55388.358**
6 [1]	125	-24,418.128	58,869	3794.840	1.000	.763	213.869	.007	15.044	<.001	55355.365
7 [2]	146	-24,614.995	58,851	3698.435	1.000	.816	122.991	.035	123.666	<.001	55413.748
8 [2]	167	-24,453.058	58,830	3590.495	1.000	.762	128.755	.038	129.462	<.001	55484.101
9 [2]	188	-23,000.892	58,812	3539.246	1.000	.761	-999	-999	-999	-999	55571.474
10 [2]	209	-23,954.556	58,786	3421.109	1.000	.768	175.272	.736	-999	-999	55671.257

Note. C = number of classes; *q* = number of free parameters; *LL* = log likelihood; LR Δ^2 *df* = degrees of freedom for the likelihood ratio chi-square test. LR Δ^2 = Likelihood ratio chi-square test of the difference between the observed versus expected frequency tables for the categorical latent class indicators. LR Δ^2 *p* = probability value for the likelihood ratio chi-square test; *E* = entropy; LMR = Lo-Mendell-Rubin adjusted Likelihood Ratio Test when comparing the *k* to *k* – 1 class model; LMR *p* = probability value for the Lo-Mendell-Rubin adjusted Likelihood Ratio Test. 2*Δ*LL* = Two times the log likelihood difference between *k* and *k* – 1 models for the bootstrapped likelihood ratio test. BLRT *p* = probability value for the bootstrapped likelihood ratio test. BIC = Bayesian Information Criterion; *N* = 646.

[1] Best loglikelihood values initially obtained using 80 and 16, then replicated using 160 and 32, random starting value perturbations and final stage optimizations. [2] Best loglikelihood values initially obtained using 320 and 64, then replicated using 640 and 128 random starting value perturbations and final stage optimizations.

[3] Number of initial stage random starts for the k-1 class analysis model = 20; Number of final stage optimizations for the k-1 class analysis model = 4

[4] Difference in the number of estimated parameters for *k* versus *k* – 1 models for the BLRT was 21.

Bold typeface indicates preferred model based on converging evidence across fit statistics.

Supplementary Table 7

Results of Exploratory Factor Mixture Modeling of CBCL Attention Problems in the Two-Year Follow-Up Wave of Data from the ABCD Study

Classes	Model	*LL*	LR Δ^2 *df*	LR Δ^2	LR Δ^2 *p*	Entropy	BIC
1		-27,271.773	58,932	4,007.773	1.0000		55,245.420
2	FMM-1 [1]	-27,729.425	58,853	5,191.192	1.0000	.895	58,051.317
	FMM-2 [2]	-28,039.115	58,932	4,031.630	1.0000	.564	55,253.961
	FMM-3 [2]	**-24,616.476**	**58,920**	**3,660.710**	**1.0000**	**.614**	**54,902.574**
3	FMM-1[1]	-26,465.635	58,923	4,342.573	1.0000	.882	55,673.358
	FMM-2 [4]	-27,728.688	58,928	4001.844	1.0000	.472	55,270.346
	ZI FMM-1[1]	-26,613.997	58,919	4,363.670	1.0000	.881	55,730.333
	ZI FMM-3 [3]	-23,511.314	58,907	3627.279	1.0000	.516	54,892.252
4	FMM-1 [1]	-25,554.856	58,926	4,169.173	1.0000	.850	55,435.462
	FMM-2 [1]	-29,625.937	58,925	4,000.532	1.0000	.348	55,294.206
	ZI FMM-1[1]	-25,896.665	58,929	4,191.474	1.0000	.851	55,428.748
	ZI FMM-2 [4]	-28,842.051	58,925	3,990.111	1.0000	.409	55,285.069

Note. LL = log likelihood; LR Δ^2 *df* = degrees of freedom for the likelihood ratio chi-square test. LR Δ^2 = Likelihood ratio chi-square test of the difference between the observed versus expected frequency tables for the categorical latent class indicators. LR Δ^2 *p* = probability value for the likelihood ratio chi-square test. BIC = Bayesian Information Criterion; FMM = factor mixture modeling; ZI = zero-inflated model; *N* = 5,820.

[1] Estimated using the robust maximum likelihood estimator (MLR) divided by the scaling correction factor for non-normality of ordinal data. Best loglikelihood values initially obtained using 80 and 16, then replicated using 160 and 32 random starting value perturbations and final stage optimizations.

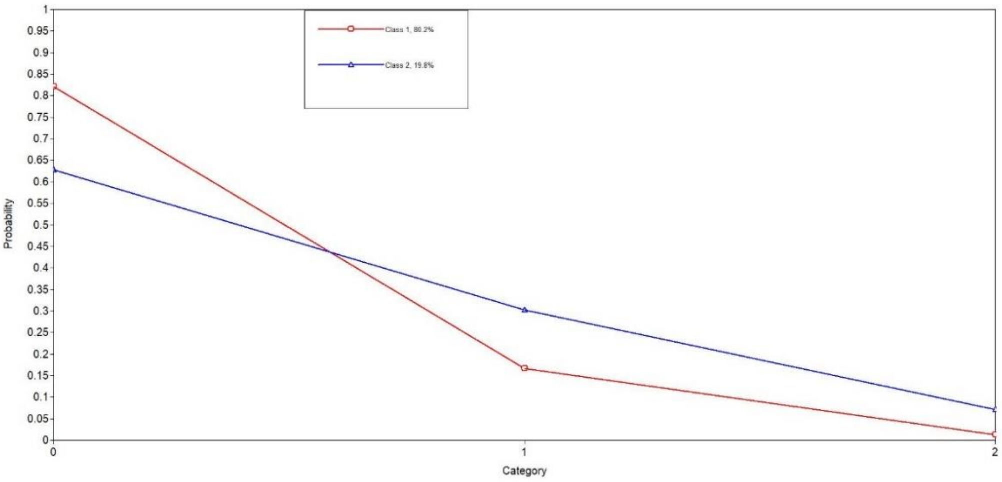
[2] Best loglikelihood values initially obtained using 160 and 32, then replicated using 320 and 64 random starting value perturbations and final stage optimizations.

[3] Best loglikelihood values initially obtained using 320 and 64, then replicated using 640 and 128 random starting value perturbations and final stage optimizations.

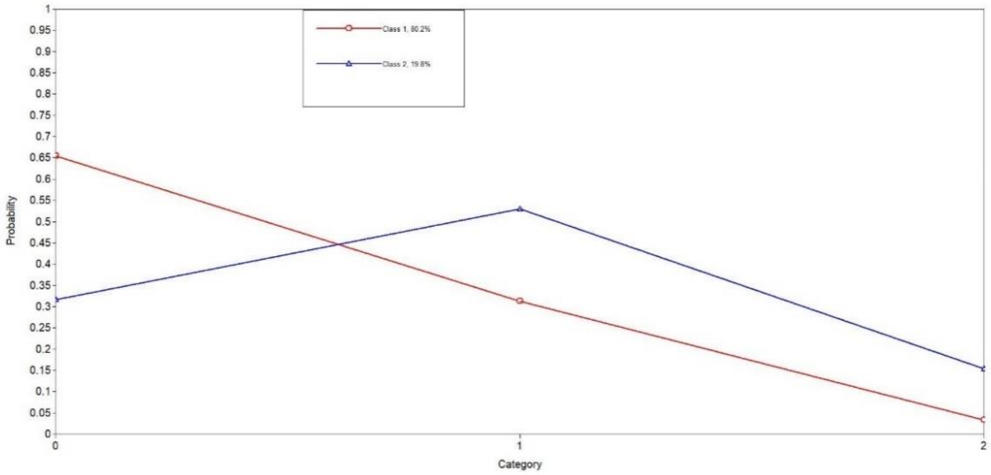
[4] The best log likelihood was not replicated across runs.

Bold typeface indicates preferred model based on fit statistics.

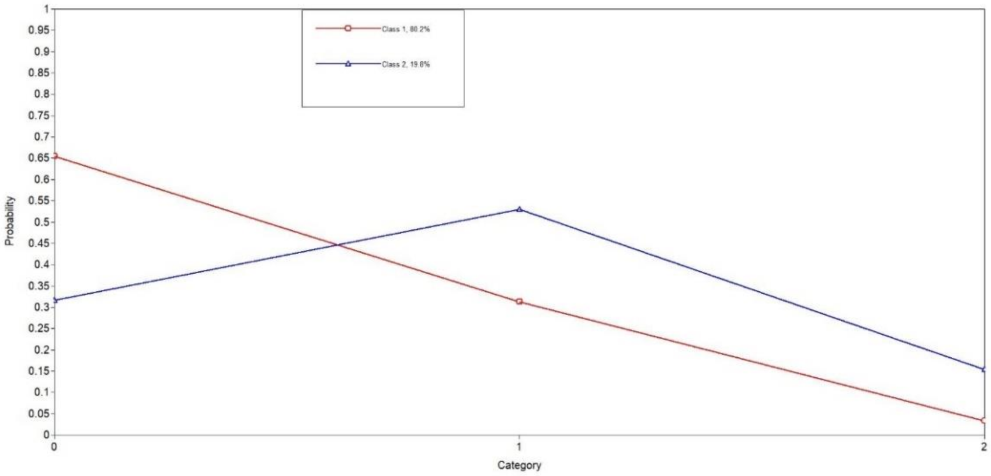
The following models were misspecified and did not converge on trustworthy estimates and therefore the results were not reported for these models: 2C FMM-4; 2C ZI (converged, but had zero cases in the zero-inflated class); 3C FMM-3; 3C FMM-4; 3C ZI FMM-2; 3C ZI FMM-4; 4C FMM-3; 4C FMM-4; 4C ZI FMM-3; 4C ZI FMM-4.



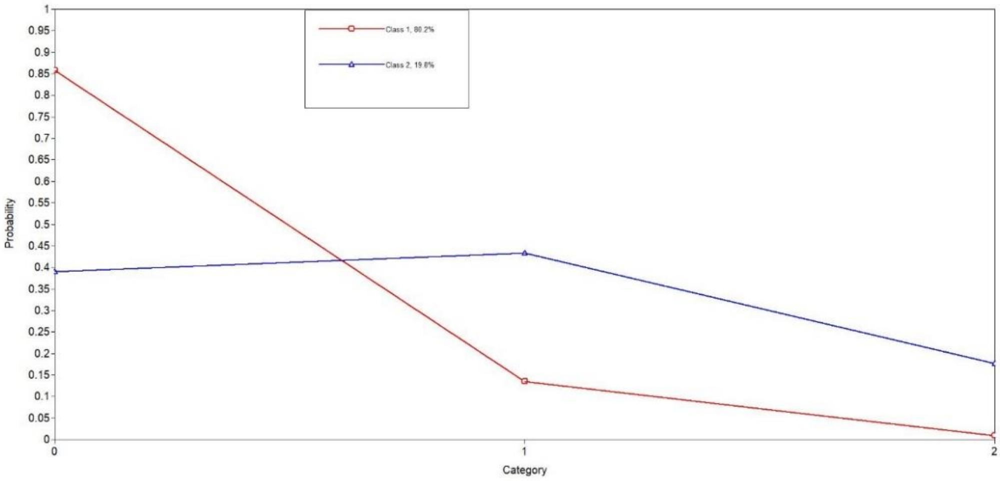
Supplementary Figure 17. Item Probability Plot for CBCL Item 1 “Acts Young” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True.*



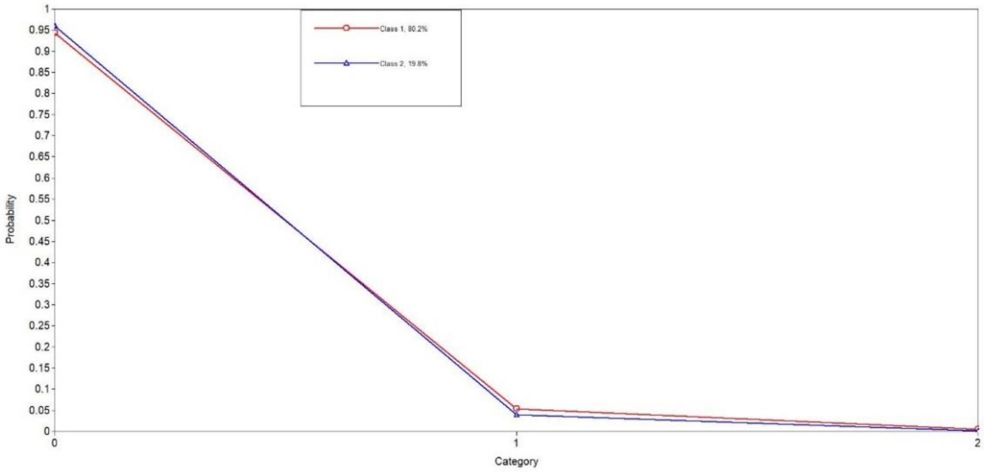
Supplementary Figure 18. Item Probability Plot for CBCL Item 4 “Fails to Finish” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True.*



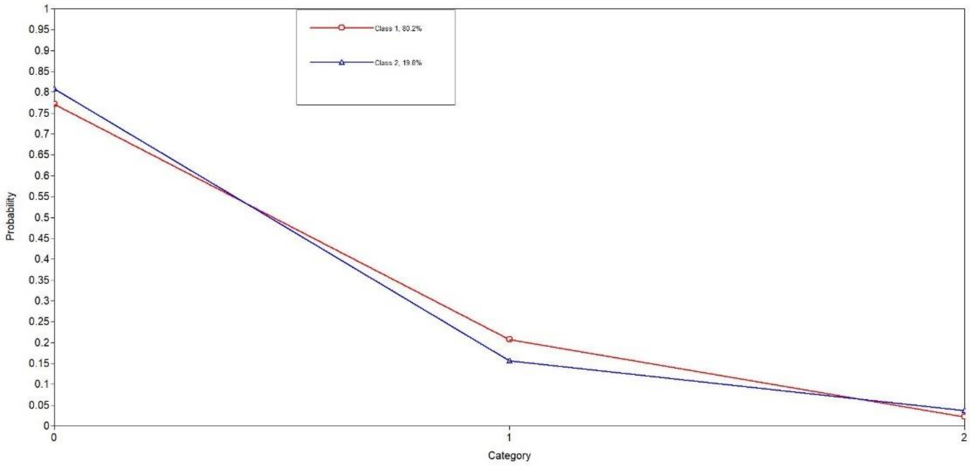
Supplementary Figure 19. Item Probability Plot for CBCL Item 8 “Concentrate” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True.*



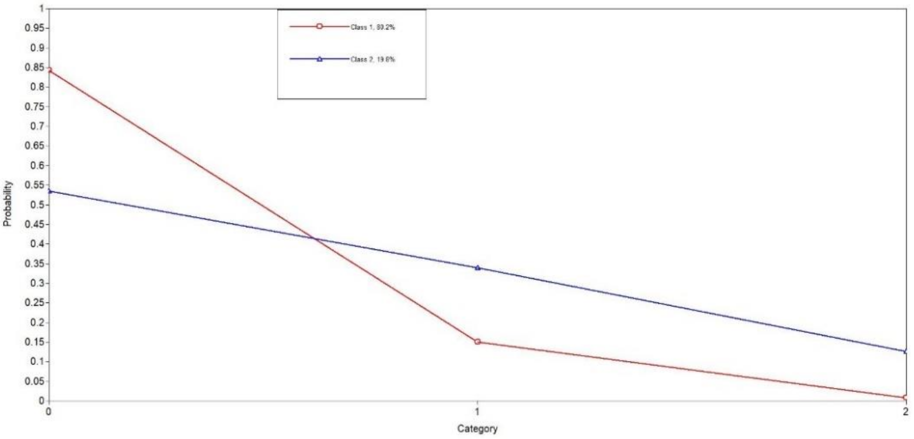
Supplementary Figure 20. Item Probability Plot for CBCL Item 10 “Sit Still” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True*.



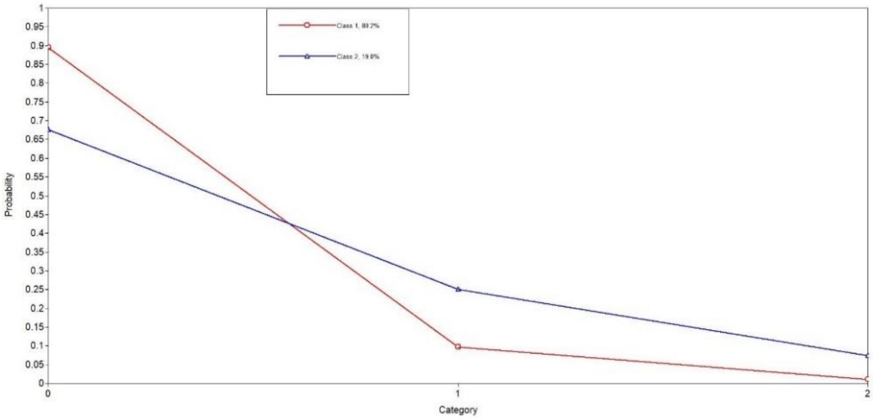
Supplementary Figure 21. Item Probability Plot for CBCL Item 13 “Confused” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True*.



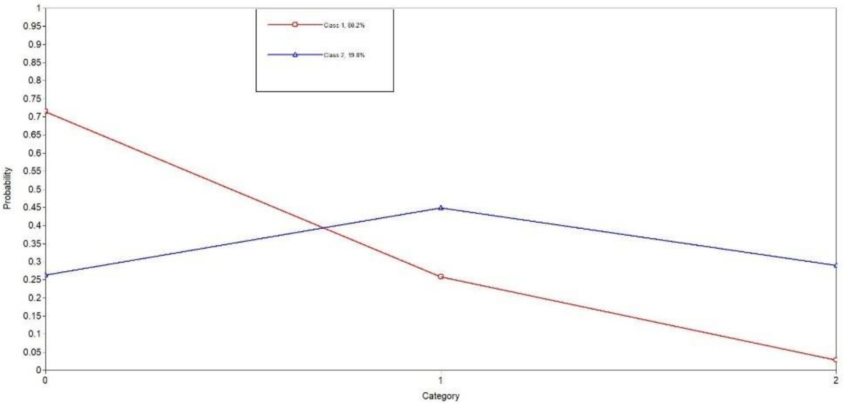
Supplementary Figure 22. Item Probability Plot for CBCL Item 17 “Daydream” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True*.



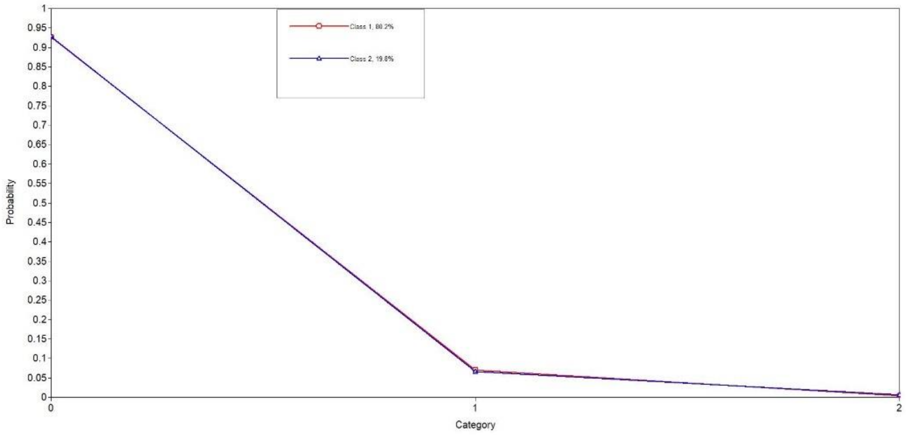
Supplementary Figure 23. Item Probability Plot for CBCL Item 41 “Impulsive” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True.*



Supplementary Figure 24. Item Probability Plot for CBCL Item 61 “Poor School” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True.*



Supplementary Figure 25. Item Probability Plot for CBCL Item 78 “Inattentive” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True.*



Supplementary Figure 26. Item Probability Plot for CBCL Item 80 “Stares” for the Two-Class FMM-3 Model. Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True.*

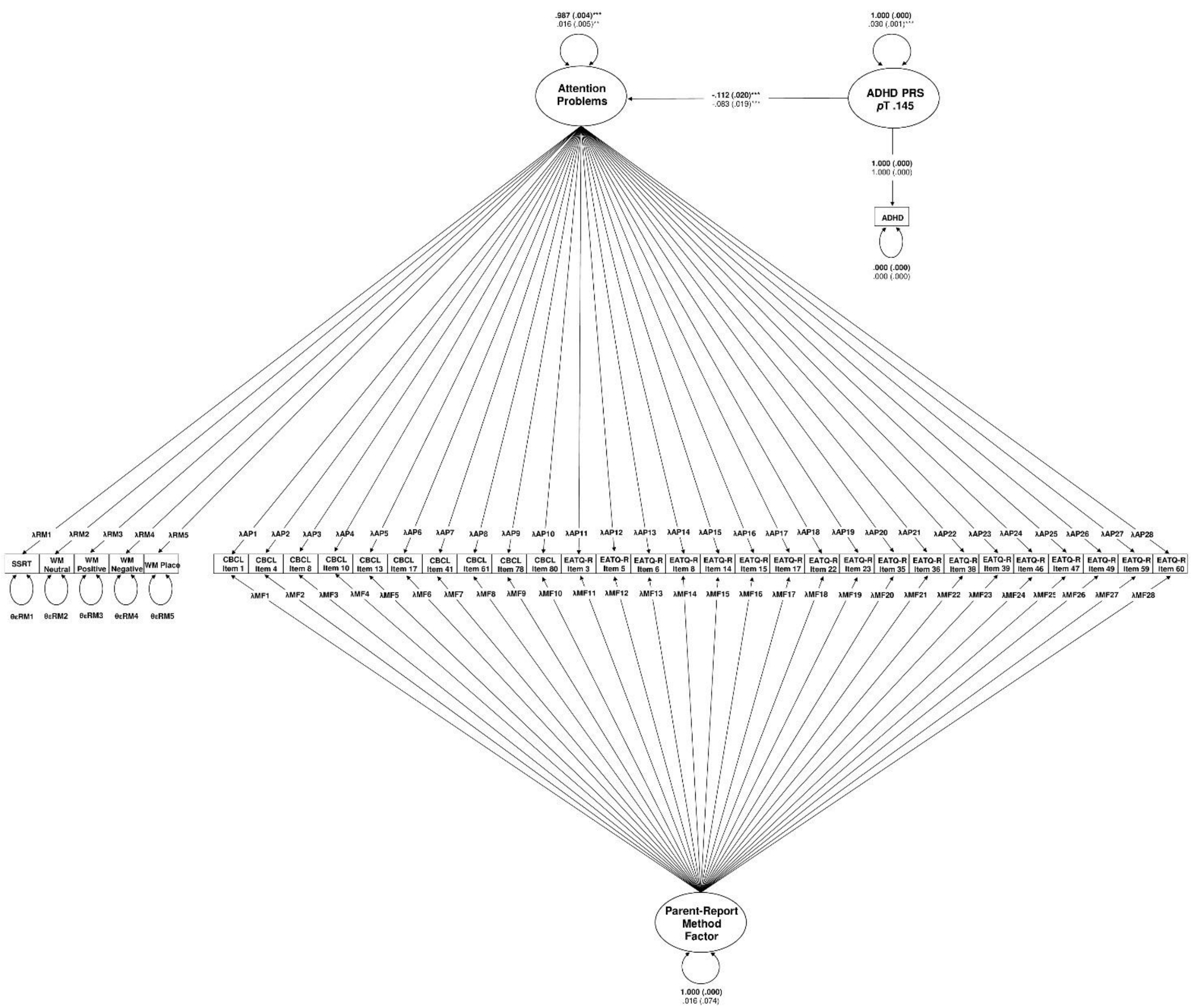
Example 6 – Controlling for Method Variance

To specify a T(M-1) model, one method is chosen as the reference method, which is indistinguishable from the target trait. An important property of this model is that because there is a reference method, there must always be one less method factor than the number of methods used to measure the target psychological attribute (hence the M-1 specification)[65,66] . In other words, it is now understood that method effects are a fundamental element of psychological measurement that cannot be completely excluded from the psychological attribute being measured[65,66]. For this reason, even in multimethod approaches to psychological measurement, one of the methods must be considered the ‘reference method’ and incorporated into the construct as part of the assessment process[65,66]. The advantage of the T(M-1) approach is that the method factor represents the residual variances of the indicators not shared with the trait as measured by the reference method. Thus, the method effect(s) is/are represented as a latent variable(s)[65,66].

As a first step, we sought to increase phenotypic resolution by combining the CBCL attention problems empirical syndrome scale items with the EATQ-R effortful control subscale items, that latter of which represents the adaptive end of the latent trait continuum for ADHD-related problems (example 4). We then incorporated cognitive variables known to be sensitive indicators of ADHD-related problems, response inhihinition[67] and working memory[68-70]. We used stop-signal reaction time as measured on the stop signal task[71] and estimated using the integration method[72] and d-prime[73] as a measure of working memory on four different conditions of a working memory 2-back task: 1) neutral faces; 2) positive faces; 3) negative faces; and 4) places, obtained from the 2-year follow-up wave of data collection of the ABCD study[74]. The stop signal task has been well-described, including in the ABCD cohort[75,76]. For the n-back task, participants had to indicate whether a picture presented on a screen on each trial was a "Match" or "No Match" to stimuli presented two trials prior[74]. Working memory performance was defined as the response accuracy from the two-back condition for each of the four stimulus conditions. We also incorporated polygenic risk scores for ADHD from saliva samples obtained at baseline, at a *p* value threshold (P_T) of .145 (ADHD PRS), which was identified as the optimal threshold for explaining variance in the CBCL attention problems scale in PRSice[77]. ADHD PRS quantifies the cumulative genetic risk for a disorder as a weighted sum of disorder-associated single nucleotide polymorphisms (SNPs) as identified in genome-wide association studies[78-80]. Participants of European ancestry were selected for all further analyses in order to match the genetic ancestry of the discovery genome wide association study (GWAS) for ADHD used to calculate PRSs (n = 2,848)[81,82].

For the purposes of specifying the T(M-1) model, cognitive assessment was selected as the reference method, such that method bias associated with parent-report symptoms and temperament on the CBCL and EATQ-R could be excluded as a method factor from the model[65,66]. We used a listwise approach to case selection to ensure only participants with ADHD PRS and cognitive performance data were included in the analysis. The final T(M-1) model is displayed in Supplementary Figure 27. The attention problems construct was characterized by weak loadings from the cognitive variables (λ = .112 - .176) and modest (λ = .247, $p < .001$) to very strong (λ = .916, $p < .001$) loadings from the parent-report items on the CBCL Attentional Problems and EATQ-R Effortful Control items (Supplementary Table 8). This factor represented the attention problems construct uncontaminated by method variance from parent-report, which was captured by a residual method factor. The residual item loadings on this method factor ranged from very weak (λ = .005, $p = .897$) to moderately strong (λ = .721, $p < .001$) (Supplementary Table 9) and this factor did not have statistically significant variance (φ = .016, $p = .829$), further confirming its status as a junk factor (i.e., representing residual variance related to parent-report not of substantive interest).

We regressed the attention problems factor onto ADHD PRS and found that ADHD PRS explained 1.0% of the variance in the attention problems latent trait factor with cognition as the reference method. In contrast, the method factor was not meaningfully related to ADHD PRS (φ = -.043, *SE* = .026, $p = .101$). Thus, we constrained their association to zero (Supplementary Figure 27). Furthermore, we were unable to get a model without cognition as the reference method and a method factor for the CBCL and EATQ-R items to converge. These results provide evidence that incorporation of multi-method approaches, specified as a T(M-1) model, can yield meaningful results in biology-psychopathology association studies.



Supplementary Figure 27. Trait Method Minus One [T(M-1] model of CBCL attention problems empirical syndrome scale augmented with the EATQ-R effortful control items in the two-year follow-up data wave of the ABCD study (*N* = 2,166). Cognition was the reference method, with parent-report items forming the method factor and its variance excluded from the attention problems latent variable. Note that polygenic risk for ADHD explained variance in the attention problems factor (1.3%), but was unrelated to the parent-report method factor.

Supplementary Table 8

Standardized Parameter Estimates, Standard Errors, and Probability Values of Model Parameter Estimates from the T(M-1) Model of Attention Problems for the Reference Method Variables and the Attention Problems Item Factor Loadings

Parameter	Standardized Estimate (λ)	Standard Error (*SE*)	Probability value (*p*)
λRM1	-0.156	0.030	<.001
λRM2	0.129	0.030	<.001
λRM3	0.156	0.030	<.001
λRM4	0.124	0.031	<.001
λRM5	0.192	0.029	<.001
θεRM1	0.976	0.009	<.001
θεRM2	0.983	0.008	<.001
θεRM3	0.976	0.009	<.001
θεRM4	0.985	0.005	<.001
θεRM5	0.963	0.011	<.001
λAP1	-0.596	0.024	<.001
λAP2	-0.791	0.025	<.001
λAP3	-0.923	0.013	<.001
λAP4	-0.765	0.019	<.001
λAP5	-0.683	0.033	<.001
λAP6	-0.579	0.025	<.001
λAP7	-0.733	0.021	<.001
λAP8	-0.679	0.042	<.001
λAP9	-0.913	0.015	<.001
λAP10	-0.676	0.032	<.001
λAP11	0.724	0.043	<.001
λAP12	0.281	0.029	<.001
λAP13	0.507	0.020	<.001
λAP14	0.414	0.027	<.001
λAP15	0.439	0.049	<.001
λAP16	0.611	0.033	<.001
λAP17	0.462	0.041	<.001
λAP18	0.579	0.019	<.001
λAP19	0.496	0.026	<.001
λAP20	0.664	0.028	<.001
λAP21	0.530	0.062	<.001
λAP22	0.527	0.072	<.001
λAP23	0.614	0.044	<.001
λAP24	0.538	0.067	<.001
λAP25	0.243	0.027	<.001
λAP26	0.693	0.026	<.001
λAP27	0.600	0.038	<.001
λAP28	0.633	0.031	<.001

Note. λ = factor loading; θε = error/residual variance; RM = reference method; AP = attention problems.

Supplementary Table 9

Standardized Parameter Estimates, Standard Errors, and Probability Values of Model Parameter Estimates from the T(M-1) Model of Attention Problems for the Method Factor Item Loadings

Parameter	Standardized Estimate (λ)	Standard Error (*SE*)	Probability value (*p*)
λMF1	0.031	0.020	0.120
λMF2	-0.211	0.078	0.007
λMF3	-0.082	0.092	0.374
λMF4	0.050	0.080	0.529
λMF5	-0.083	0.084	0.322
λMF6	0.008	0.062	0.895
λMF7	-0.007	0.076	0.922
λMF8	-0.409	0.065	<.001
λMF9	-0.088	0.093	0.344
λMF10	0.007	0.077	0.927
λMF11	0.414	0.073	<.001
λMF12	0.190	0.036	<.001
λMF13	-0.038	0.060	0.531
λMF14	0.057	0.049	0.247
λMF15	0.440	0.051	<.001
λMF16	0.295	0.061	<.001
λMF17	0.356	0.052	<.001
λMF18	0.020	0.062	0.749
λMF19	0.095	0.056	0.089
λMF20	0.230	0.069	0.001
λMF21	0.635	0.052	<.001
λMF22	0.744	0.051	<.001
λMF23	0.449	0.058	<.001
λMF24	0.689	0.054	<.001
λMF25	0.064	0.035	0.066
λMF26	0.209	0.072	0.003
λMF27	0.375	0.057	<.001
λMF28	0.266	0.062	<.001

Note. λ = factor loading; MF = method factor.

The Distinction Between the Child Behavior Checklist and the Hierarchical Taxonomy of Psychopathology

The Child Behavior Checklist (CBCL) is dimensional and hierarchical like the Hierarchical Taxonomy of Psychopathology (HiTOP) model and is used widely around the world including in large, consortia-sized datasets (e.g., Adolescent Brain and Cognitive Development study)[83], but has failed to yield robust findings of the neural and genetic correlates of developmental psychopathology (e.g., Marek et al., 2022)[4]. It is also a HiTOP-conformant measure. The use of HiTOP-conformant measures enables broadband dimensional and hierarchical measurement of psychopathology, circumventing issues of arbitrary clinical cut-offs and loss of power, as well as the comorbidity problem. However, the problems of phenotypic complexity and variable-centred heterogeneity can only be resolved when these dimensions are explicitly modelled hierarchically. Common usages of the CBCL rely on subscale raw scores[4,6,25], which do not address the issues of phenotypic complexity and variable-centred heterogeneity. The other limitation of the CBCL is that its development was based on optimising the differentiation of clinically-referred versus non-referred children (i.e., criterion keying)[6,25]. Thus, the CBCL provides high levels of information (i.e., reliability) at the clinical and subclinical end of the psychopathology spectrum, but very low information at the normative end of the continuum (example 2)[19]. Thus, the CBCL has poor phenotypic resolution as we have demonstrated in example 2 and cannot reliably rank-order individuals in the normative range, limiting its utility in biology-psychopathology association studies. In contrast, the broader HiTOP model combines both clinical components and maladaptive traits, the latter of which characterize trait levels across the full spectrum of individual differences[84,85]. Furthermore, some HiTOP conformant measures, including the Computerized Adaptive Assessment of Personality Disorder (CAT-PD) and Externalizing Spectrum Inventory – Brief Form (ESI-BF) have been optimised using techniques such as item response theory to measure individual differences with high precision across the latent trait continuum[84,86]. For these reasons, measures of the HiTOP model are expected to yield more robust findings than the CBCL.

References

1 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).

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

3 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* **57**, 289-300, doi:https://doi.org/10.1111/j.2517-6161.1995.tb02031.x (1995).

4 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).

5 Luciana, M. *et al.* Adolescent neurocognitive development and impacts of substance use: Overview of the adolescent brain cognitive development (ABCD) baseline neurocognition battery. *Dev Cogn Neurosci* **32**, 67-79, doi:https://doi.org/10.1016/j.dcn.2018.02.006 (2018).

6 Achenbach, T. M. & Rescorla, L. A. *Manual for the ASEBA school-age forms & profiles.*, (University of Vermont, Research Center for Children, Youth, & Families, 2001).

7 Lynam, D. Development of a short form of the UPPS-P Impulsive Behavior Scale. *Unpublished technical report* (2013).

8 Cyders, M. A., Littlefield, A. K., Coffey, S. & Karyadi, K. A. Examination of a short English version of the UPPS-P Impulsive Behavior Scale. *Addict Beh* **39**, 1372-1376, doi:https://doi.org/doi.org/10.1016/j.addbeh.2014.02.013 (2014).

9 Whiteside, S. P. & Lynam, D. R. The Five Factor Model and impulsivity: Using a structural model of personality to understand impulsivity. *Pers Individid Differ* **30**, 669-689, doi:https://doi.org/10.1016/S0191-8869(00)00064-7 (2001).

10 Whiteside, S. P., Lynam, D. R., Miller, J. D. & Reynolds, S. K. Validation of the UPPS impulsive behaviour scale: A four-factor model of impulsivity. *Eur J Pers* **19**, 559 - 574, doi:https://doi.org/10.1002/per.556 (2005).

11 Cyders, M. A. *et al.* Integration of impulsivity and positive mood to predict risky behavior: Development and validation of a measure of positive urgency. *Psychol Assess* **19**, 107 - 118, doi:https://doi.org/10.1037/1040-3590.19.1.107 (2007).

12 Pagliaccio, D. *et al.* Revising the BIS/BAS Scale to study development: Measurement invariance and normative effects of age and sex from childhood through adulthood. *Psychol Assess* **28**, 429-442, doi:https://doi.org/10.1037/pas0000186 (2016).

13 Carver, C. S. & White, T. L. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. *J Pers Soc Psychol* **67**, 319, doi:http://dx.doi.org/10.1037/0022-3514.67.2.319 (1994).

14 Loewy, R. L., Therman, S., Manninen, M., Huttunen, M. O. & Cannon, T. D. Prodromal psychosis screening in adolescent psychiatry clinics. *Early Interv Psychiatry* **6**, 69-75, doi:https://doi.org/10.1111/j.1751-7893.2011.00286.x (2012).

15 Loewy, R. L., Bearden, C. E., Johnson, J. K., Raine, A. & Cannon, T. D. The prodromal questionnaire (PQ): Preliminary validation of a self-report screening measure for prodromal and psychotic syndromes. *Schizophr Res* **79**, 117-125, doi:https://doi.org/10.1016/j.schres.2005.03.007 (2005).

16 Funder, D. C. & Ozer, D. J. Evaluating effect size in psychological research: Sense and nonsense. *AMPPS* **2**, 156-168, doi:https://doi.org/10.1177/2515245919847202 (2019).

17 Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav Res Methods* **41**, 1149-1160, doi:https://doi.org/10.3758/BRM.41.4.1149 (2009).

18 McArdle, J. J. Causal-modeling applied to psychonomic systems simulation. *Behav res meth instrum* **12**, 193-209, doi:https://doi.org/10.3758/bf03201598 (1980).

19 Tiego, J. & Fornito, A. Putting behaviour back into brain-behaviour correlation analyses. *Aperture Neuro* ***2*** (2022).

20 Edelen, M. O. & Reeve, B. B. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res* **16(Suppl 1)**, 5-18, doi:10.1007/s11136-007-9198-0 (2007).

21 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).

22 Reise, S. P., Ainsworth, A. T. & Haviland, M. G. Item response theory: Fundamentals, applications, and promise in psychological research. *Curr Dir Psychol Sci* **14**, 95-101, doi:https://doi.org/10.1111/j.0963-7214.2005.00342.x (2005).

23 Toland, M. D. Practical guide to conducting an item response theory analysis. *J Early Adolesc* **34**, 120-151, doi:https://doi.org/10.1177/0272431613511332 (2014).

24 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).

25 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).

26 Achenbach, T. M. & Edelbrock, C. S. *Manual for the Child Behavior Checklist/4-18 and the 1991 Profile.*, (University of Vermont Department of Psychiatry, 1991).

27 Teresi, J. A. Overview of quantitative measurement methods. Equivalence, invariance, and differential item functioning in health applications. *Med Care* **44**, S39-49, doi:https://doi.org/10.1097/01.mlr.0000245452.48613.45 (2006).

28 Teresi, J. A. & Fleishman, J. A. Differential item functioning and health assessment. *Qual Life Res* **16 Suppl 1**, 33-42, doi:https://doi.org/10.1007/s11136-007-9184-6 (2007).

29 Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M. & Ocepek-Welikson, K. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Application to the Mini-Mental State Examination. *Med Care* **44**, S134-S142, doi:https://doi.org/10.1097/01.mlr.0000245251.83359.8c (2006).

30 Cohen, A. S. & Bolt, D. M. A mixture model analysis of differential item functioning. *J Educ Meas* **42**, 133-148, doi:https://doi.org/doi:10.1111/j.1745-3984.2005.00007 (2005).

31 Muthen, B. & Asparouhov, T. Item response mixture modeling: Application to tobacco dependence criteria. *Addict Behav* **31**, 1050-1066, doi:https://doi.org/10.1016/j.addbeh.2006.03.026 (2006).

32 De Ayala, R. J. & Santiago, S. Y. An introduction to mixture item response theory models. *J Sch Psychol* **60**, 25-40, doi:https://doi.org/10.1016/j.jsp.2016.01.002 (2017).

33 Walker, C. M. What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *J Psychoeduc Assess* **29**, 364-376, doi:https://doi.org/10.1177/0734282911406666 (2011).

34 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).

35 Tay, L., Meade, A. W. & Cao, M. An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods* **18**, 3-46, doi:10.1177/1094428114553062 (2015).

36 Essen, C. B., Idaka, I. E. & Metibemu, M. A. Item level diagnostics and model - data fit in item response theory (IRT) using BILOG - MG V3.0 and IRTPRO V3.0 programmes. *Global Journal of Educational Research* **16**, 87-94, doi:http://dx.doi.org/10.4314/gjedr.v16i2.2 (2017).

37 Orlando, M. & Thissen, D. Further investigation of the performance of S - X2: An item fit index for use with dichotomous item response theory models. *Appl Psychol Meas* **27**, 289-298, doi:https://doi.org/10.1177/0146621603027004004 (2003).

38 Cai, L., du Toit, S. H. C. & Thissen, D. *IRTPRO: User guide.*, (Scientific Software International, 2011).

39 Savalei, V. What to do about zero frequency cells when estimating polychoric correlations. *Struct Equ Modeling* **18**, 253 - 273, doi:https://doi.org/10.1080/10705511.2011.557339 (2011).

40 Embretson, S. E. & Reise, S. P. *Item response theory for psychologists*. (Lawrence Erlbaum Associates, 2000).

41 Coghill, D. & Sonuga-Barke, E. J. Annual research review: categories versus dimensions in the classification and conceptualisation of child and adolescent mental disorders--implications of recent empirical study. *J Child Psychol Psychiatry* **53**, 469-489, doi:https://doi.org/10.1111/j.1469-7610.2011.02511.x (2012).

42 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).

43 Ellis, L. K. & Rothbart, M. K. in *Biennial Meeting of the Society for Research in Child Development* (Minneapolis, Minnesota, 2001).

44 Oldehinkel, A. J., Hartman, C. A., Ferdinand, R. F., Verhulst, F. C. & Ormel, J. Effortful control as modifier of the association between negative emotionality and adolescents' mental health problems. *Dev Psychopathol* **19**, 523 - 539, doi:https://doi.org/10.1017/s0954579407070253 (2007).

45 Tackett, J. L. Evaluating models of the personality-psychopathology relationship in children and adolescents. *Clin Psychol Rev* **26**, 584 - 599, doi:https://doi.org/10.1016/j.cpr.2006.04.003 (2006).

46 Krueger, R. F. & Tackett, J. L. Personality and psychopathology: Working toward the bigger picture. *J Pers Disord* **17**, 109 - 128, doi:https://doi.org/10.1521/pedi.17.2.109.23986 (2003).

47 Eisenberg, N., Hofer, C. & Vaughan, J. in *Hanbook of emotion regulation.* (ed J. J. Gross) Ch. 14, 287 - 306 (The Guilford Press, 2007).

48 Nigg, J. T., Sibley, M. H., Thapar, A. & Karalunas, S. L. Development of ADHD: Etiology, heterogeneity, and early life course. *Annu Rev Dev Psychol* **2**, 559-583, doi:https://doi.org/10.1146/annurev-devpsych-060320-093413 (2020).

49 Nigg, J. T. Attention-deficit/hyperactivity disorder: Endophenotypes, structure, and etiological pathways. *Curr Dir Psychol Sci* **19**, 24-29, doi:https://doi.org/10.1177/0963721409359282 (2010).

50 Nigg, J. T., Karalunas, S. L., Feczko, E. & Fair, D. A. Toward a revised nosology for attention-deficit/hyperactivity disorder heterogeneity. *Biol Psychiatry: Cogn Neurosci Neuroimaging* **5**, 726-737, doi:https://doi.org/10.1016/j.bpsc.2020.02.005 (2020).

51 Sonuga-Barke, E. J. S. The dual pathway model of AD/HD: An elaboration of neuro-developmental characteristics. *Neurosci Biobehav Rev* **27**, 593-604, doi:https://doi.org/10.1016/j.neubiorev.2003.08.005 (2003).

52 Costa Dias, T. G. *et al.* Characterizing heterogeneity in children with and without ADHD based on reward system connectivity. *Dev Cogn Neurosci* **11**, 155-174, doi:https://doi.org/10.1016/j.dcn.2014.12.005 (2015).

53 Loo, S. K., McGough, J. J., McCracken, J. T. & Smalley, S. L. Parsing heterogeneity in attention-deficit hyperactivity disorder using EEG-based subgroups. *J Child Psychol Psychiatry* **59**, 223-231, doi:https://doi.org/10.1111/jcpp.12814 (2018).

54 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).

55 Lubke, G. H. & Muthén, B. Investigating population heterogeneity with factor mixture models. *Psychol Methods* **10**, 21-39, doi:https://doi.org/10.1037/1082-989X.10.1.21 (2005).

56 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).

57 Wall, M. M., Park, J. Y. & Moustaki, I. IRT modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Appl Psychol Meas* **39**, 583-597, doi:https://doi.org/10.1177/0146621615588184 (2015).

58 Magnus, B. E. & Thissen, D. Item response modeling of multivariate count data with zero inflation, maximum inflation, and heaping. *J Educ Behav Stat* **42**, 531-558, doi:https://doi.org/10.3102/1076998617694878 (2017).

59 Muthen, B. & Asparouhov, T. Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychol Methods* **17**, 313-335, doi:https://doi.org/10.1037/a0026802 (2012).

60 Appelbaum, M. *et al.* Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *Am Psychol* **73**, 3-25, doi:https://doi.org/10.1037/amp0000191 (2018).

61 Muthén, L. K. & Muthén, B. O. *Mplus User's Guide.* . Eighth edn, (Muthén & Muthén, 1998 - 2017).

62 Lanza, S. T., Tan, X. & Bray, B. C. Latent class analysis with distal outcomes: a flexible model-based approach. *Struct Equ Modeling* **20**, 1-26, doi:https://doi.org/10.1080/10705511.2013.742377 (2013).

63 Bakk, Z. & Vermunt, J. K. Robustness of stepwise latent class modeling with continuous distal outcomes. *Struct Equ Modeling* **23**, 20-31, doi:https://doi.org/10.1080/10705511.2014.955104 (2016).

64 Asparouhov, T. & Muthén, B. Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Struct Equ Modeling* **21**, 329-341, doi:https://doi.org/10.1080/10705511.2014.915181 (2014).

65 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).

66 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).

67 Lipszyc, J. & Schachar, R. Inhibitory control and psychopathology: A meta-analysis of studies using the stop signal task. *J Int Neuropsychol Soc* **16**, 1064 - 1076, doi:https://doi.org/10.1017/S1355617710000895 (2010).

68 Alloway, T. P. *Improving working memory: Supporting students' learning*. (Sage, 2010).

69 Alloway, T. P., Gathercole, S. E., Kirkwood, H. & Elliott, J. The cognitive and behavioral characteristics of children with low working memory. *Child Dev* **80**, 606 - 621, doi:https://doi.org/10.1111/j.1467-8624.2009.01282.x (2009).

70 Gathercole, S. E. & Alloway, T. P. *Working memory and learning: A practical guide for teachers*. (Sage, 2008).

71 Logan, G. D. in *Inhibitory processes in attention, memory, and language.* (eds T. H. Carr & D. Dagenbach) Ch. 5, 189 - 239 (Academic Press, 1994).

72 Verbruggen, F. *et al.* A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. *Elife* **8**, doi:https://doi.org/10.7554/eLife.46323 (2019).

73 Haatveit, B. C. *et al.* The validity of d prime as a working memory index: Results from the “Bergen n-back task”. *J Clin Exp Neuropsychol* **32**, 871-880, doi:https://doi.org/10.1080/13803391003596421 (2010).

74 Casey, B. J. *et al.* The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev Cogn Neurosci.* **32**, 43-54, doi:https://doi.org/10.1016/j.dcn.2018.03.001 (2018).

75 Bissett, P. G., Hagen, M. P., Jones, H. M. & Poldrack, R. A. Design issues and solutions for stop-signal data from the Adolescent Brain Cognitive Development (ABCD) study. *ELife* **10**, e60185, doi:https://doi.org/10.7554/eLife.60185 (2021).

76 Verbruggen, F. & Logan, G. D. Response inhibition in the stop-signal paradigm. *TiCS* **12**, 418 - 424, doi:https://doi.org/10.1016/j.tics.2008.07.005 (2008).

77 Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic risk score software. *Bioinformatics* **31**, 1466-1468, doi:https://doi.org/10.1093/bioinformatics/btu848 (2014).

78 Wray, N. R. *et al.* Research review: Polygenic methods and their application to psychiatric traits. *J Child Psychol Psychiatry* **55**, 1068-1087, doi:https://doi.org/10.1111/jcpp.12295 (2014).

79 Wray, N. R. *et al.* From basic science to clinical application of polygenic risk scores: A primer *JAMA Psychiatry* **78**, 101-109, doi:https://doi.org/10.1001/jamapsychiatry.2020.3049 (2021).

80 Bogdan, R., Baranger, D. A. A. & Agrawal, A. Polygenic risk scores in clinical psychology: Bridging genomic risk to individual differences. *Annu Rev Clin Psychol* **14**, 119-157, doi:https://doi.org/10.1146/annurev-clinpsy-050817-084847 (2018).

81 Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine* **12**, 44, doi:10.1186/s13073-020-00742-5 (2020).

82 Demontis, D. *et al.* Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. *Nat Genet*, doi:https://doi.org/10.1038/s41588-022-01285-8 (2023).

83 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).

84 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).

85 DeYoung, C. G. *et al.* The distinction between symptoms and traits in the Hierarchical Taxonomy of Psychopathology (HiTOP). *J Pers*, doi:https://doi.org/10.1111/jopy.12593 (2020).

86 Patrick, C. J., Kramer, M. D., Krueger, R. F. & Markon, K. E. Optimizing efficiency of psychopathology assessment through quantitative modeling: Development of a brief form of the Externalizing Spectrum Inventory. *Psychol Assess* **25**, 1332-1348, doi:https://doi.org/10.1037/a0034864 (2013).