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

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 values (p); $p > .05$ indicates that the null hypothesis of exact fit of the model to the data cannot be rejected[2]. 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]. We considered freely estimating residual covariances where indicated by modification indices, theoretically plausible, and when statistically significant after controlling for Type I error using the Benjamini-Hochberg procedure[3].

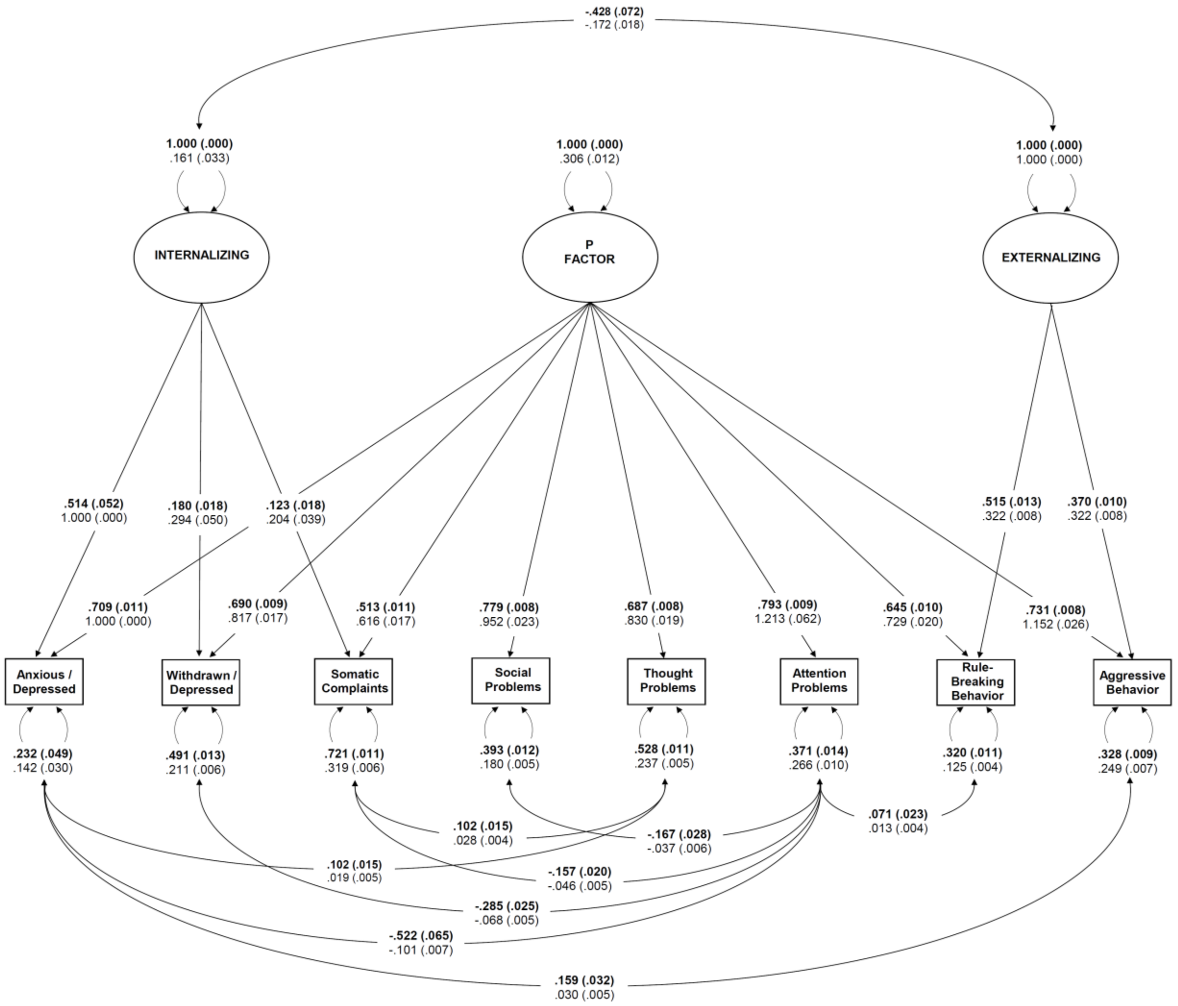
The model is displayed in Supplementary Figure 1. The proportions of phenotypic variance for the eight syndrome and three composite scales attributable to different levels of the psychopathology hierarchy are shown in Supplementary Figure 2. As can be seen from Supplementary Figure 2, almost half (48.8%) of the variance in the eight syndrome scales is common and attributable to the overarching *p*-factor. Common variance will substantially attenuate correlations with external variables, under the assumption that common variance limits our ability to identify associations with specific psychopathology phenotypes. Variance unique to each of the scales ranged from as little as 23.2% for Withdrawn/Depressed to 72.1% for Somatic Complaints, but averaged less than half (42.3%) across the eight scales. Conversely, less than 49% of the sample variance in the Total Problems scale is attributable to the general *p*-factor, with approximately 42% unique to the eight syndrome scales, and around 9% attributable to the Internalizing and Externalizing Problems group factors.

Variance unique to each subscale and the Internalizing and Externalizing factors contained in the Total Problems scale will attenuate relationships between the common variance component (i.e., *p*-factor) and criterion variables (e.g., genetic markers and imaging-derived phenotypes).

By way of example, 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 the unreliability of the CBCL syndrome scales due to contamination by general variance (i.e., *p*-factor) 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,

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

which indicates that, even if we assume zero error in the imaging-derived phenotypes, by taking into account the other sources of variance in each subscale, the true correlations could be considerably higher than those observed and reported (e.g., for an observed effect of *r* = 0.10, the true effects would be Anxious/Depressed *r* = .208; Withdrawn/Depressed *r* = .143; Somatic Complaints *r* = .117; Social Problems *r* = .160; Thought Problems *r* = .138; Attention Problems *r* = .164; Rule-Breaking Behavior *r* = .177; Aggressive Behavior *r* = .175; Internalizing *r* = .310; Externalizing *r* = .223; Total Problems *r* = .143). Considering that r_{es} = .10, .2, and .30 correspond with small, medium and large effects sizes respectively[16], the true effect sizes are 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. For example, a sample size of 614 would be required to achieve 80% power to detect an attenuated effect of r = .10 for the Internalizing scale, but this requirement would decrease to 60 for a disattenuated effect of r = .310[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. In combination, our results demonstrate that attenuation of biology-behavior associations can be substantial when high phenotypic complexity (and low phenotypic resolution) is not considered.

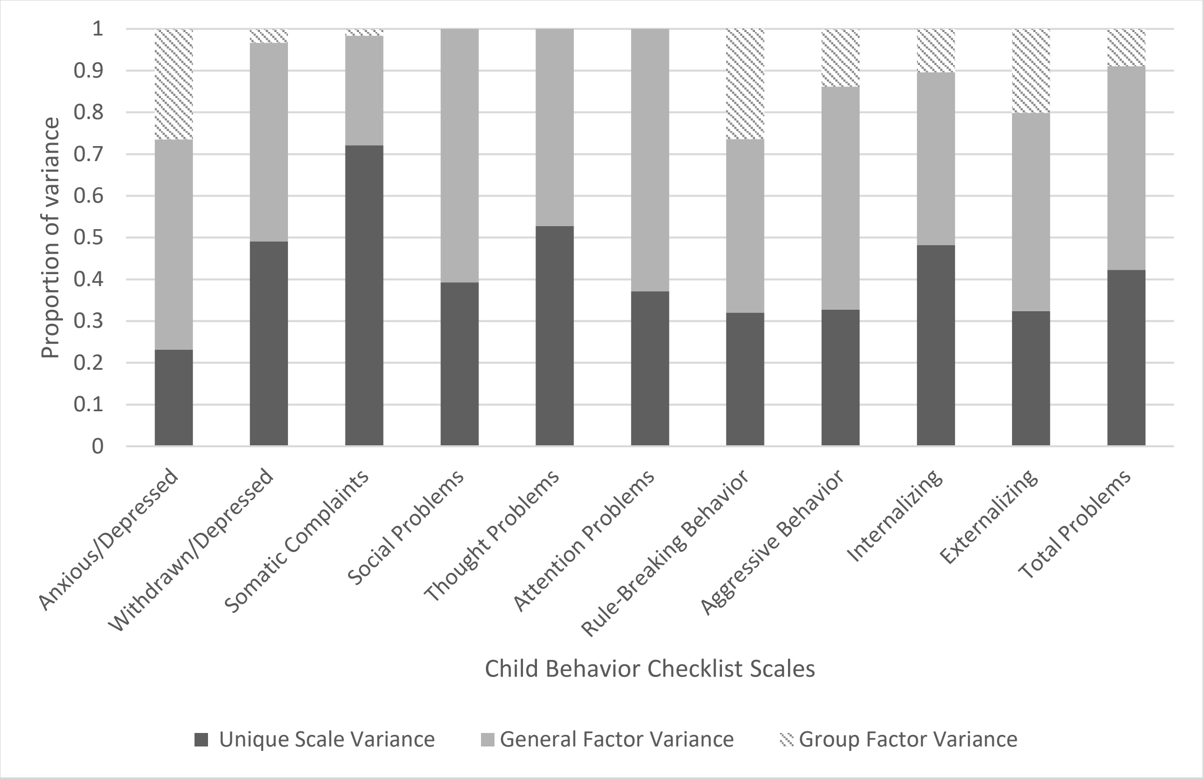


Supplementary Figure 1. Bifactor model of the CBCL data obtained from the two-year follow-up data collection wave of the ABCD study cohort.

Note. Model fit statistics were χ^2 (7) = 8.351, *p* = .303, RMSEA = .006, [95%*CI* = .000, .018], CFI = 1.000, SRMR = .003. All θε < 0.01.

All eight error covariances were statistically significant (*p* < .05) after correction for multiple post hoc comparisons using the Benjamini-Hochberg procedure (Benjamini-Hochberg *p* = .003).

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 symbolised 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. Curved, double-headed arrows between large ellipses are factor (i.e., latent) intercorrelations. Curved, double-headed arrows between error variances (small double-headed arrows pointing to the rectangles) are error covariances.



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.

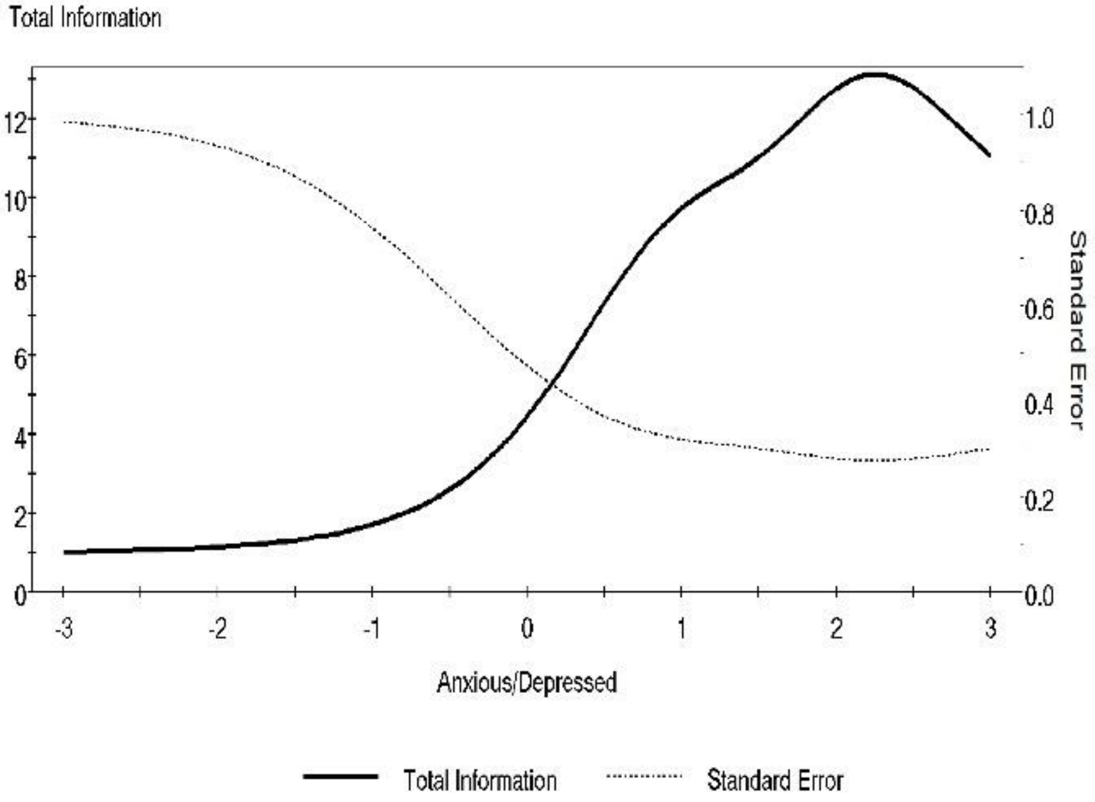
Supplementary Table 1

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

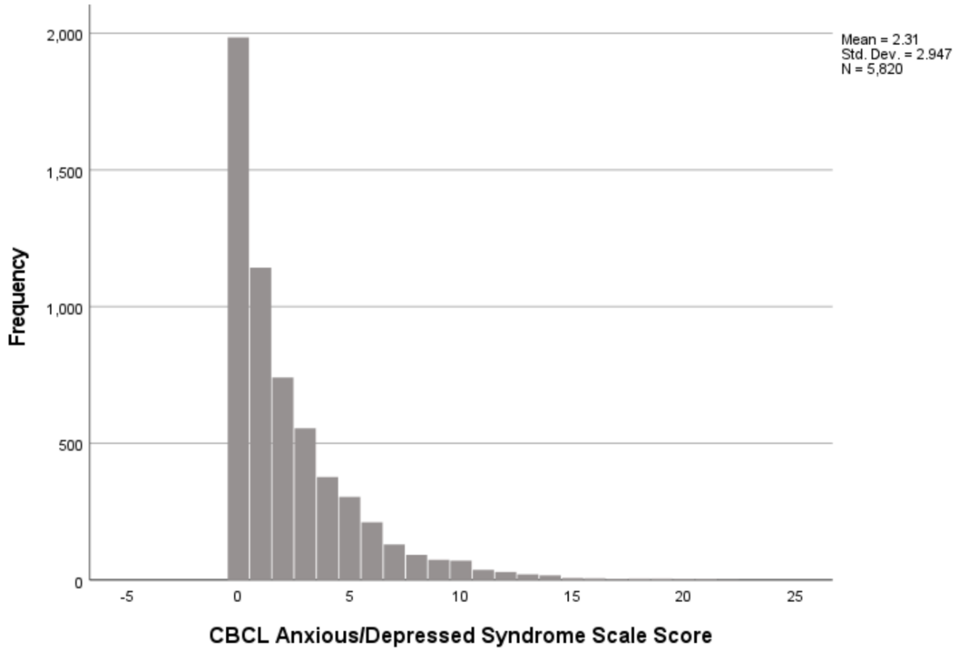
			Reliability r_{xx} (*I*) Across Latent Trait Continuum (θ)												
CBCL Scale	Number of Items	α	-3.0 *SD*	-2.5 *SD*	-2.0 *SD*	-1.5 *SD*	-1.0 *SD*	-0.5 *SD*	*M*	+0.5 *SD*	+1.0 *SD*	+1.5 *SD*	+2.0 *SD*	+2.5 *SD*	+3.0 *SD*
Anxious/Depressed	13	.813	.030	.061	.125	.241	.417	.616	.775	.863	.900	.911	.922	.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)

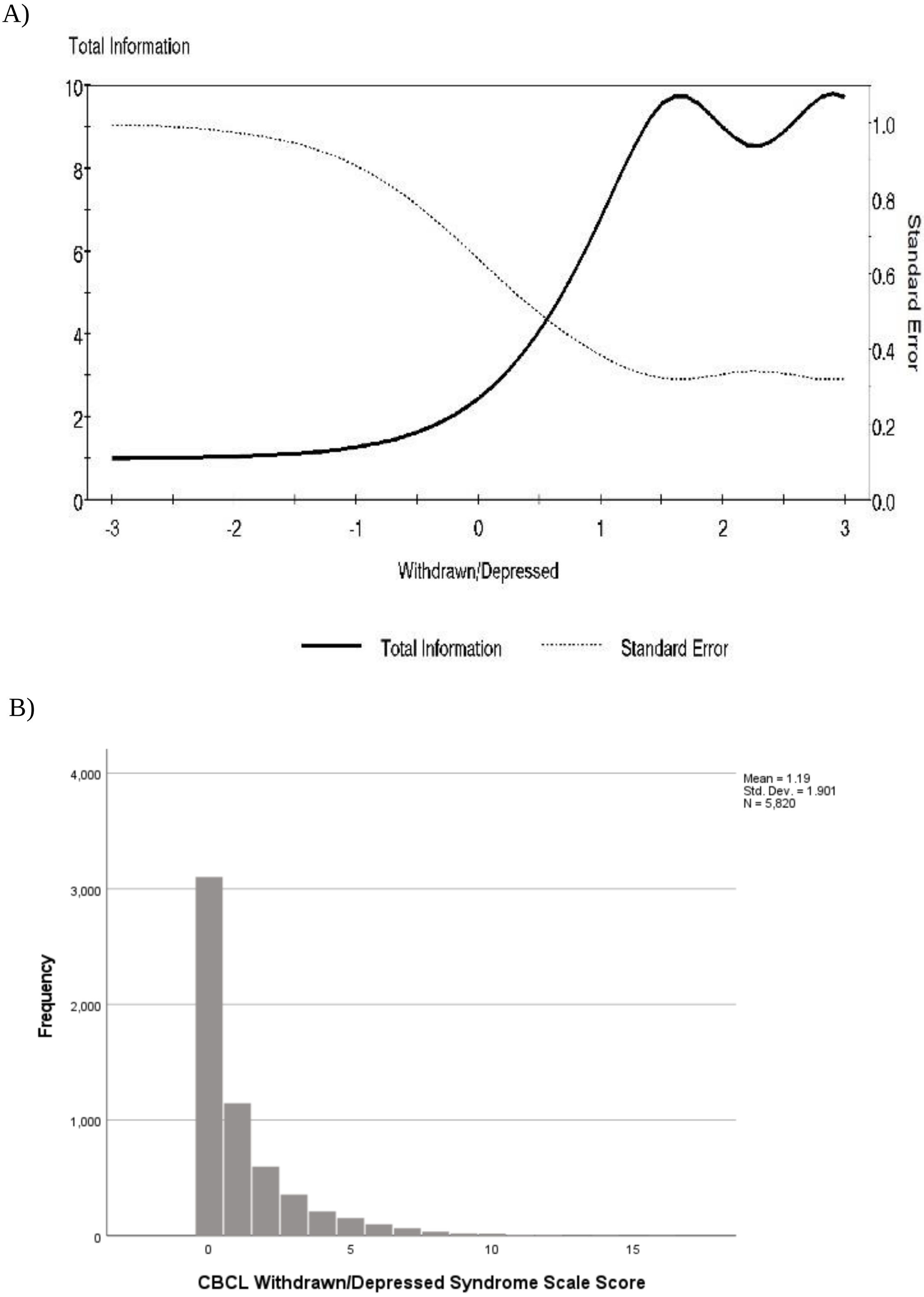


B)



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.

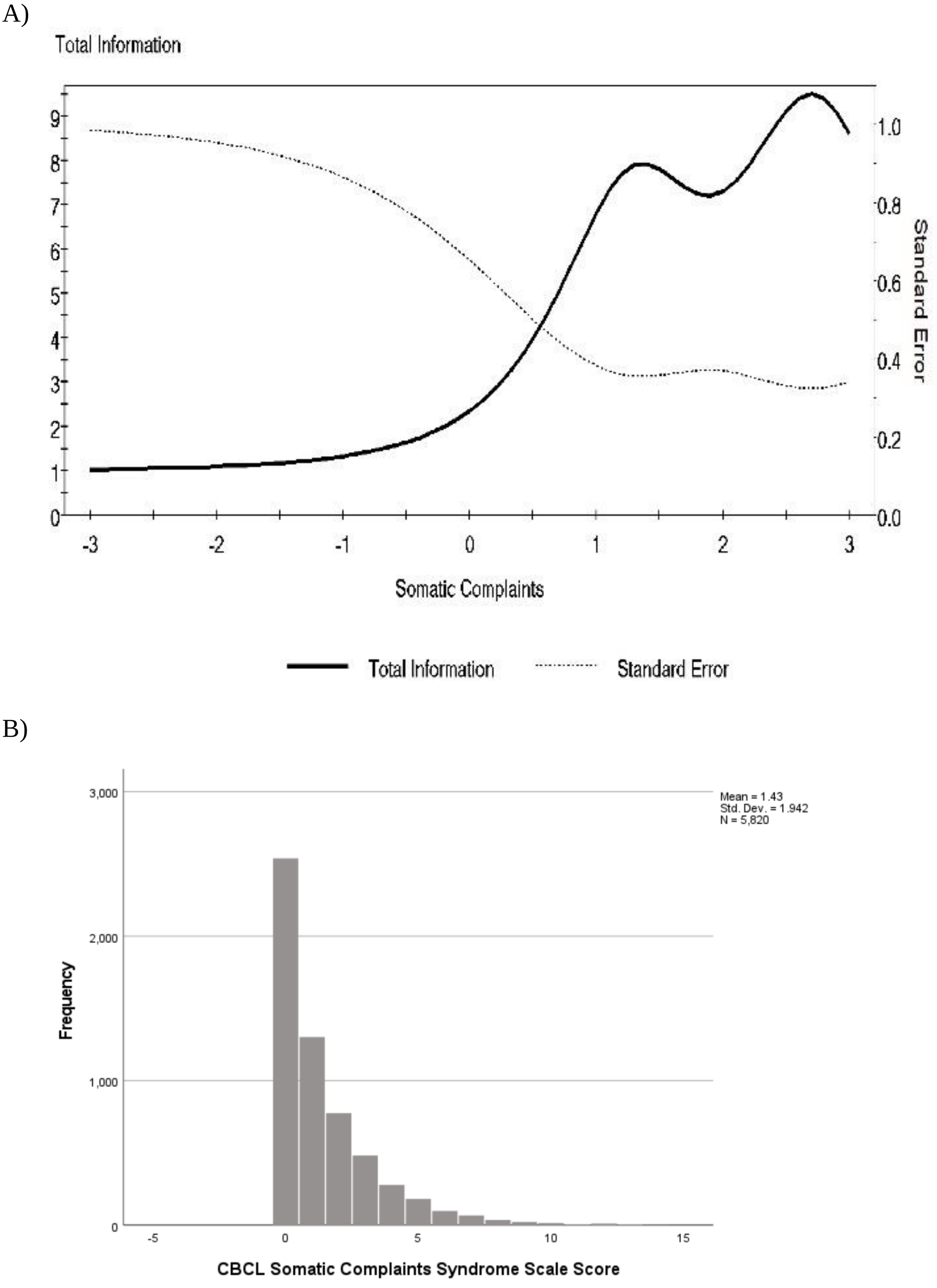
Note. N = 5,820. $r_{xx} = 1 - \left(1/_I\right)$. Standard error of the estimate (*SEE*) = $1/\sqrt{I}$.


A)

B)

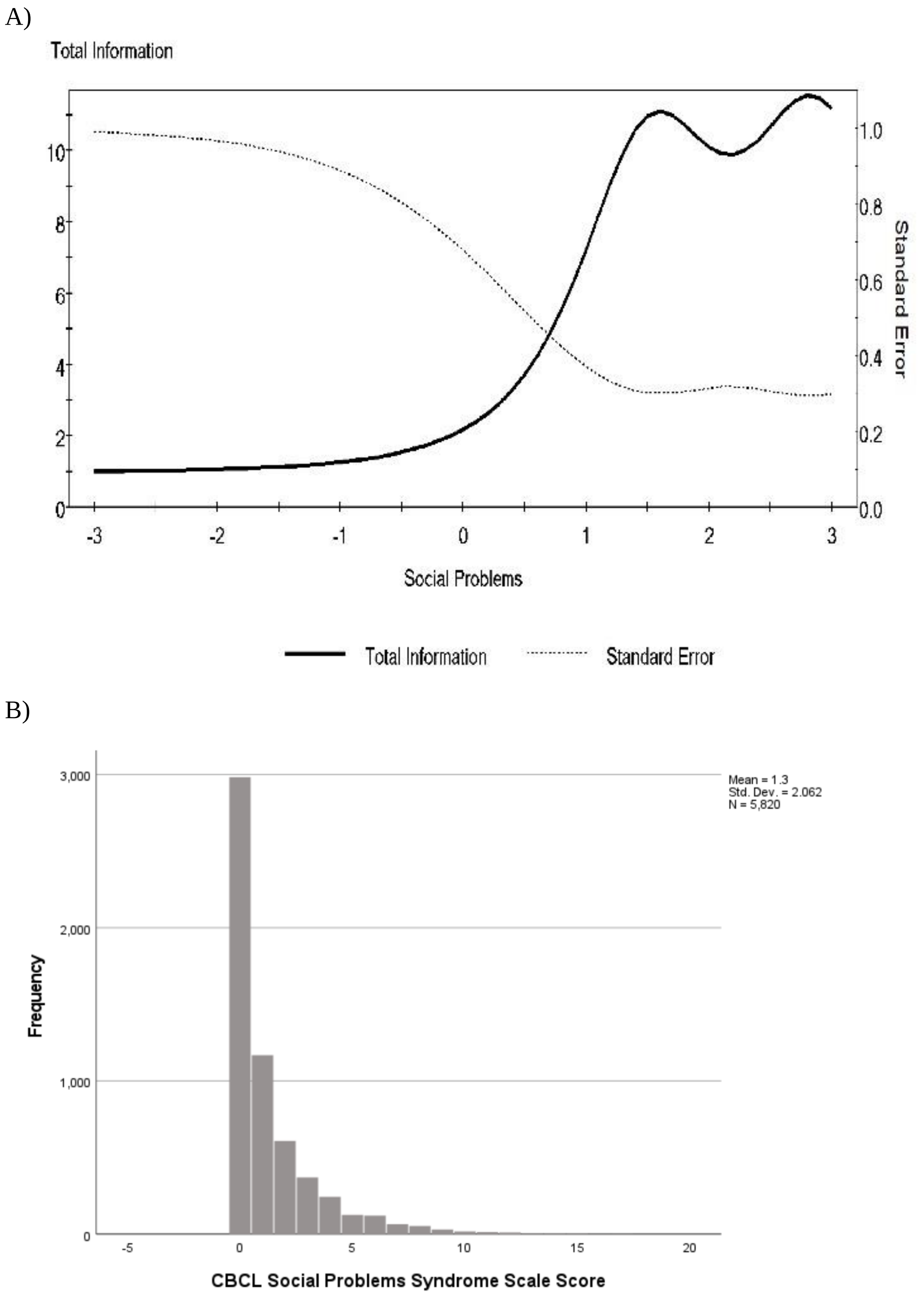
Supplementary Figure 4. A) Total information function / curve (TIF/TIC) for the child behavior checklist Withdrawn/Depressed syndrome scale. Taken from Tiego and Fornito (2022)[19]. Reprinted with permission. B) Histogram of sum scale scores on the Withdrawn/Depressed syndrome scale.

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



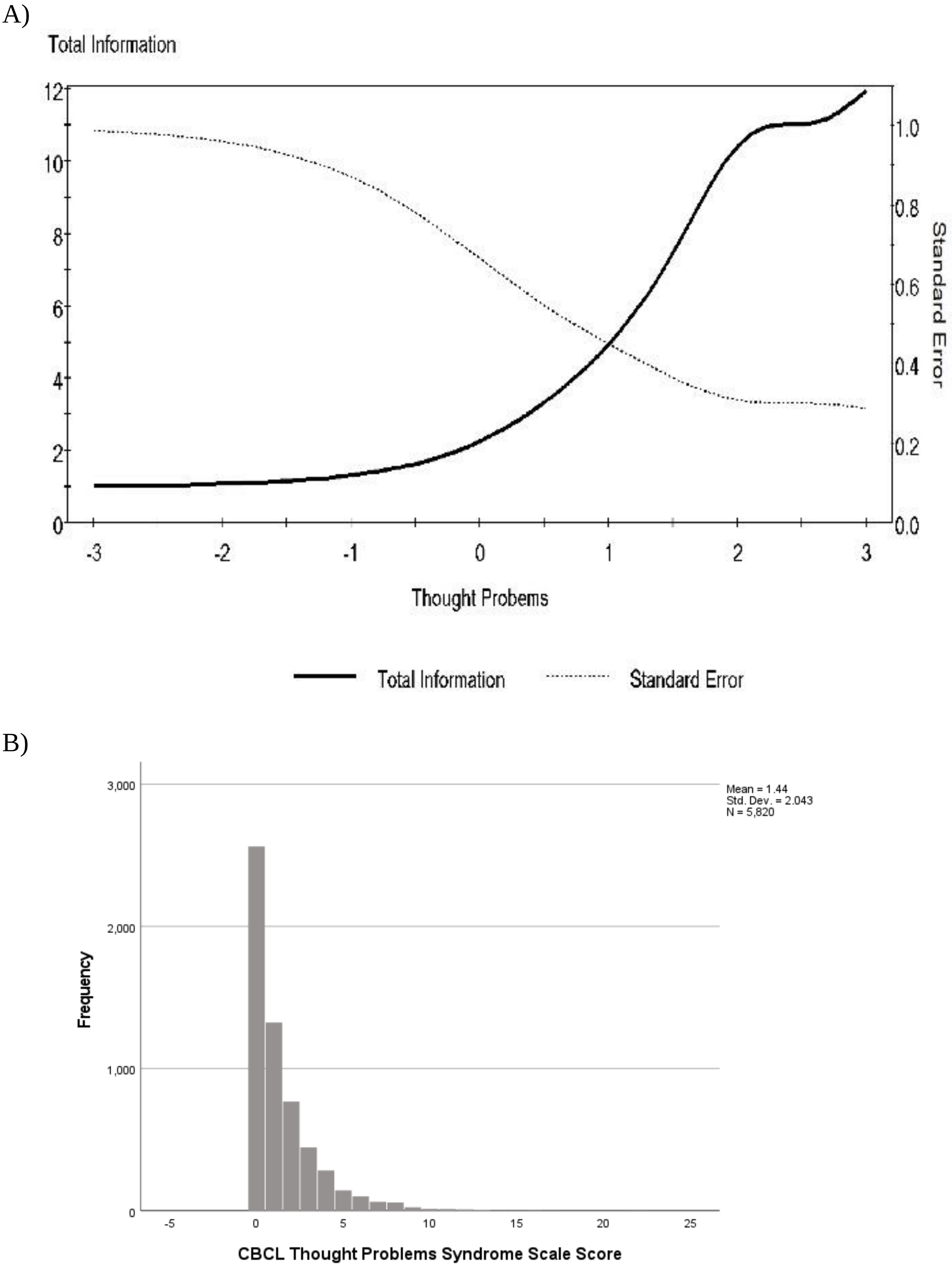
Supplementary Figure 5. A) Total information function / curve (TIF/TIC) for the child behavior checklist Somatic Complaints syndrome scale. B) Histogram of sum scale scores on the Somatic Complaints syndrome scale.

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



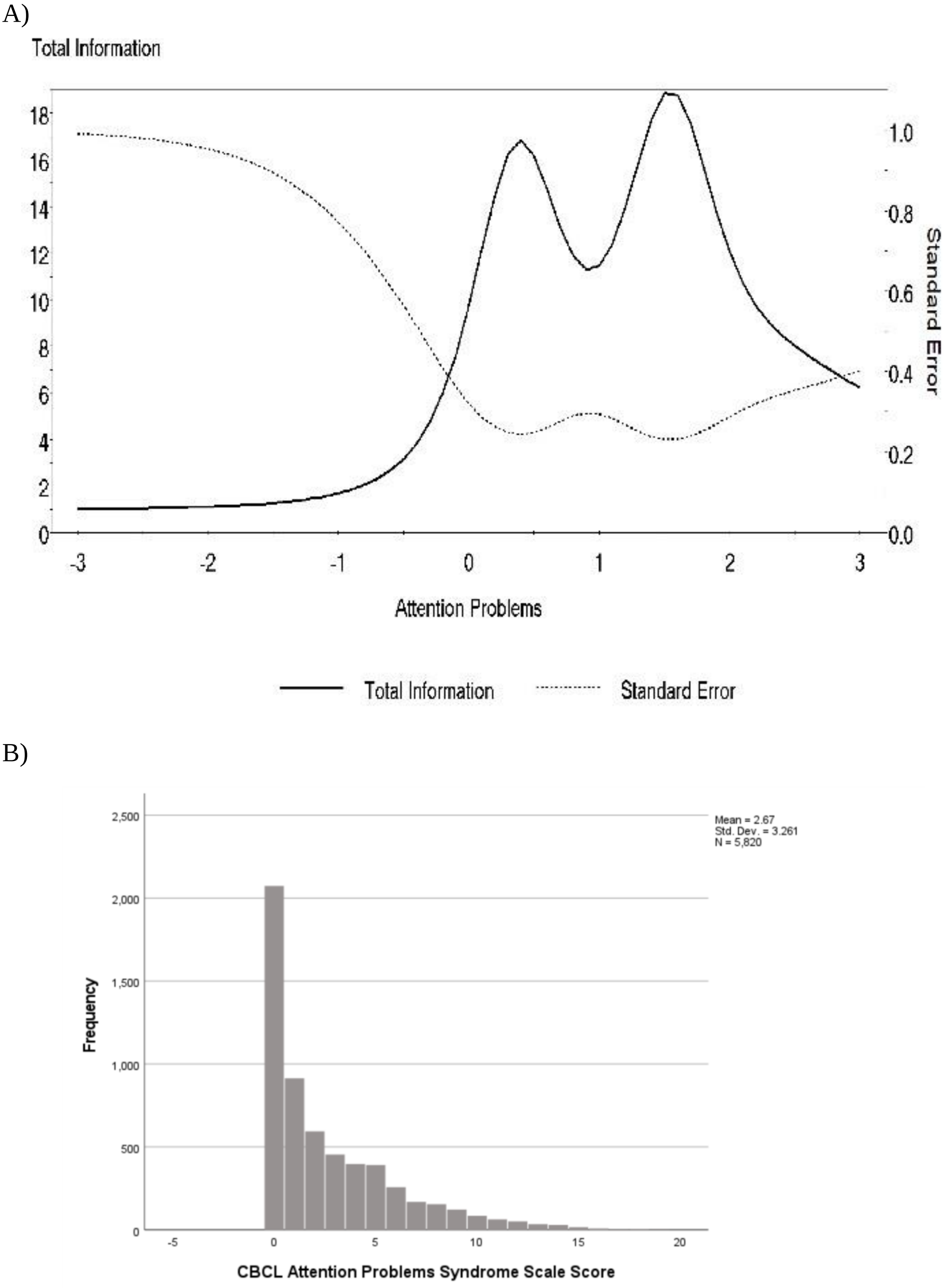
Supplementary Figure 6. A) Total information function / curve (TIF/TIC) for the child behavior checklist Social Problems syndrome scale. B) Histogram of sum scale scores on the Social Problems syndrome scale.

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



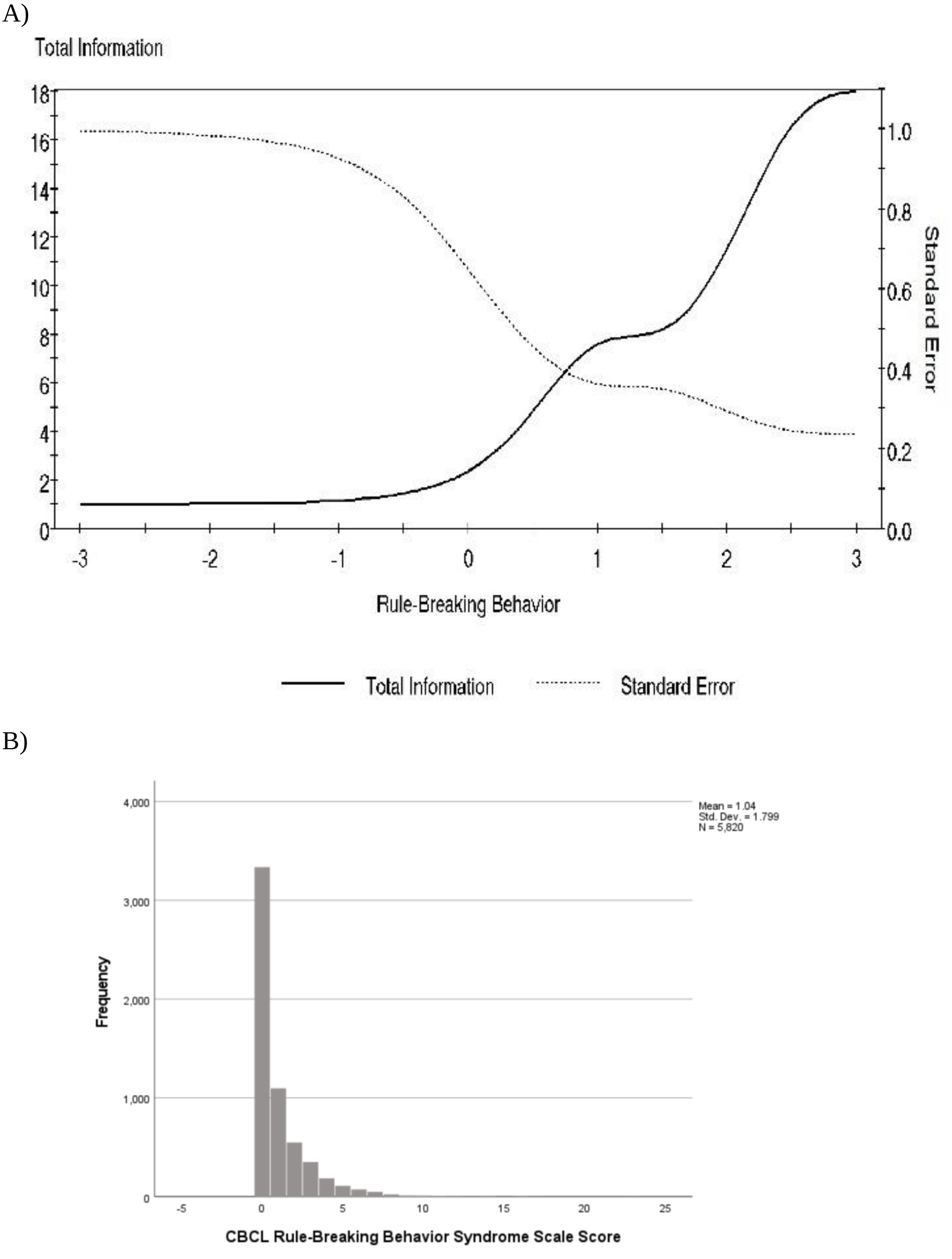
Supplementary Figure 7. A) Total information function / curve (TIF/TIC) for the child behavior checklist Thought Problems syndrome scale. B) Histogram of sum scale scores on the Thought Problems syndrome scale.

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



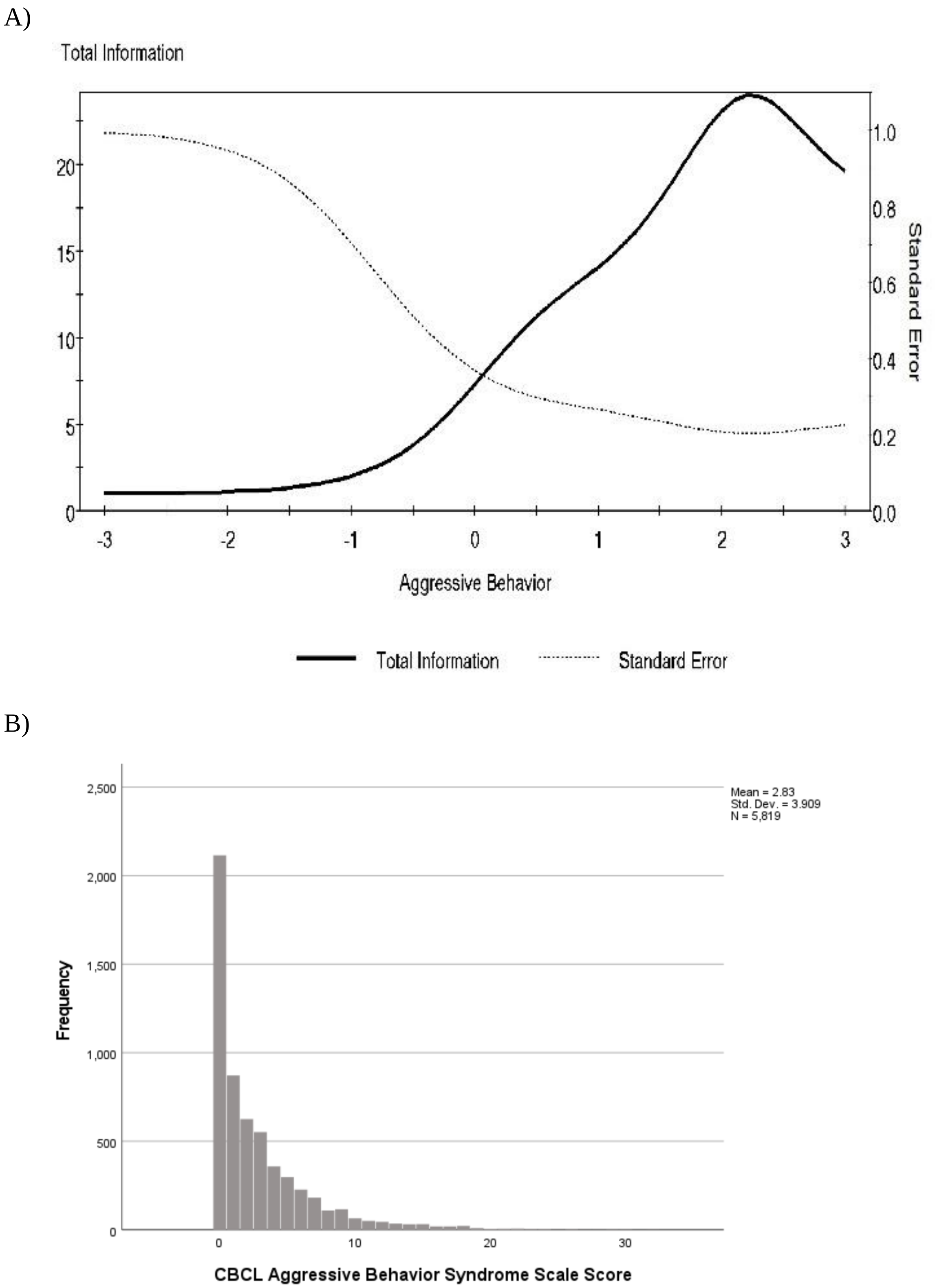
Supplementary Figure 8. A) Total information function / curve (TIF/TIC) for the child behavior checklist Attention Problems syndrome scale. B) Histogram of sum scale scores on the Attention Problems syndrome scale.

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



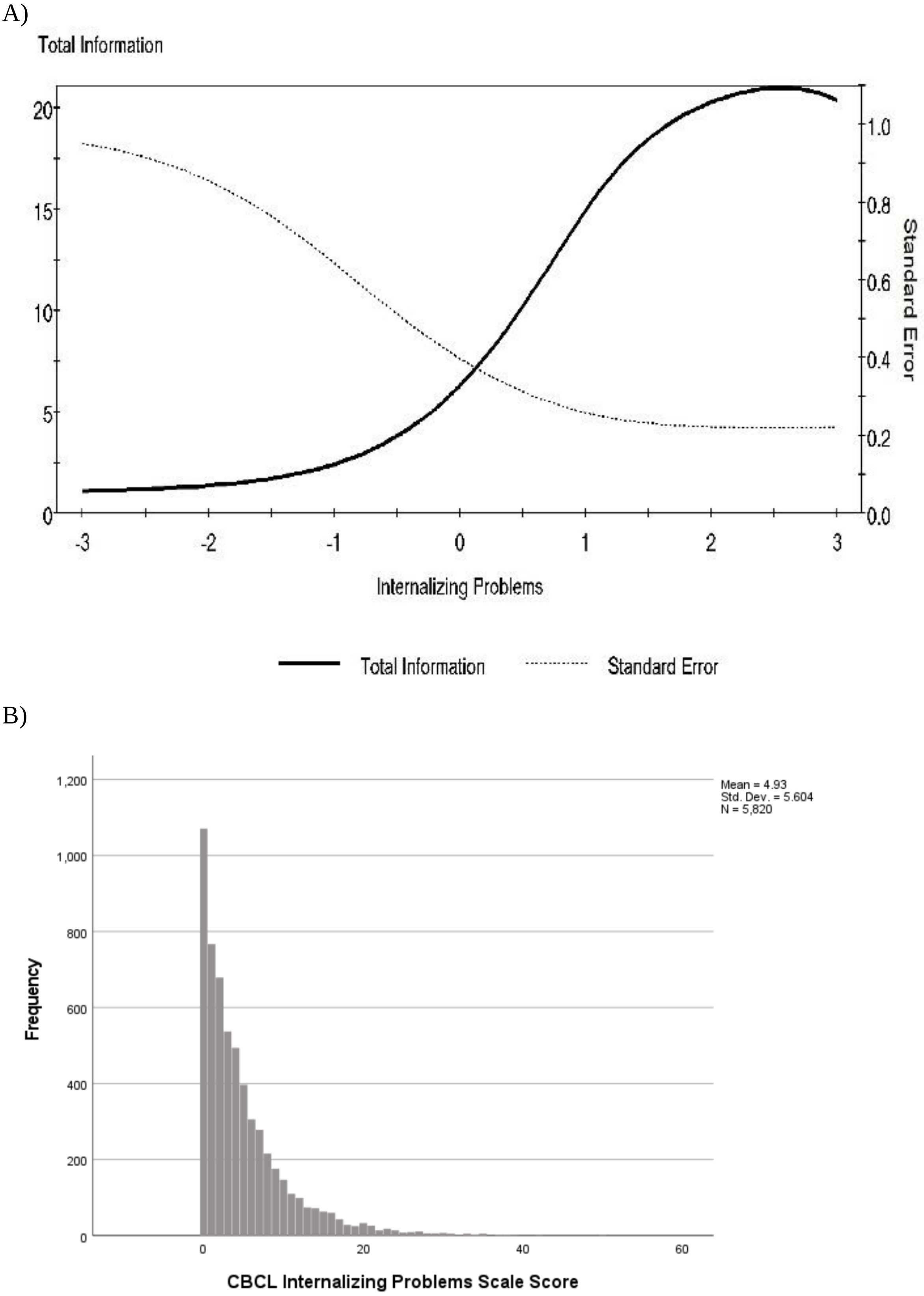
Supplementary Figure 9. A) Total information function / curve (TIF/TIC) for the child behavior checklist Rule-Breaking Behavior syndrome scale. B) Histogram of sum scale scores on the Rule-Breaking Behavior syndrome scale.

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



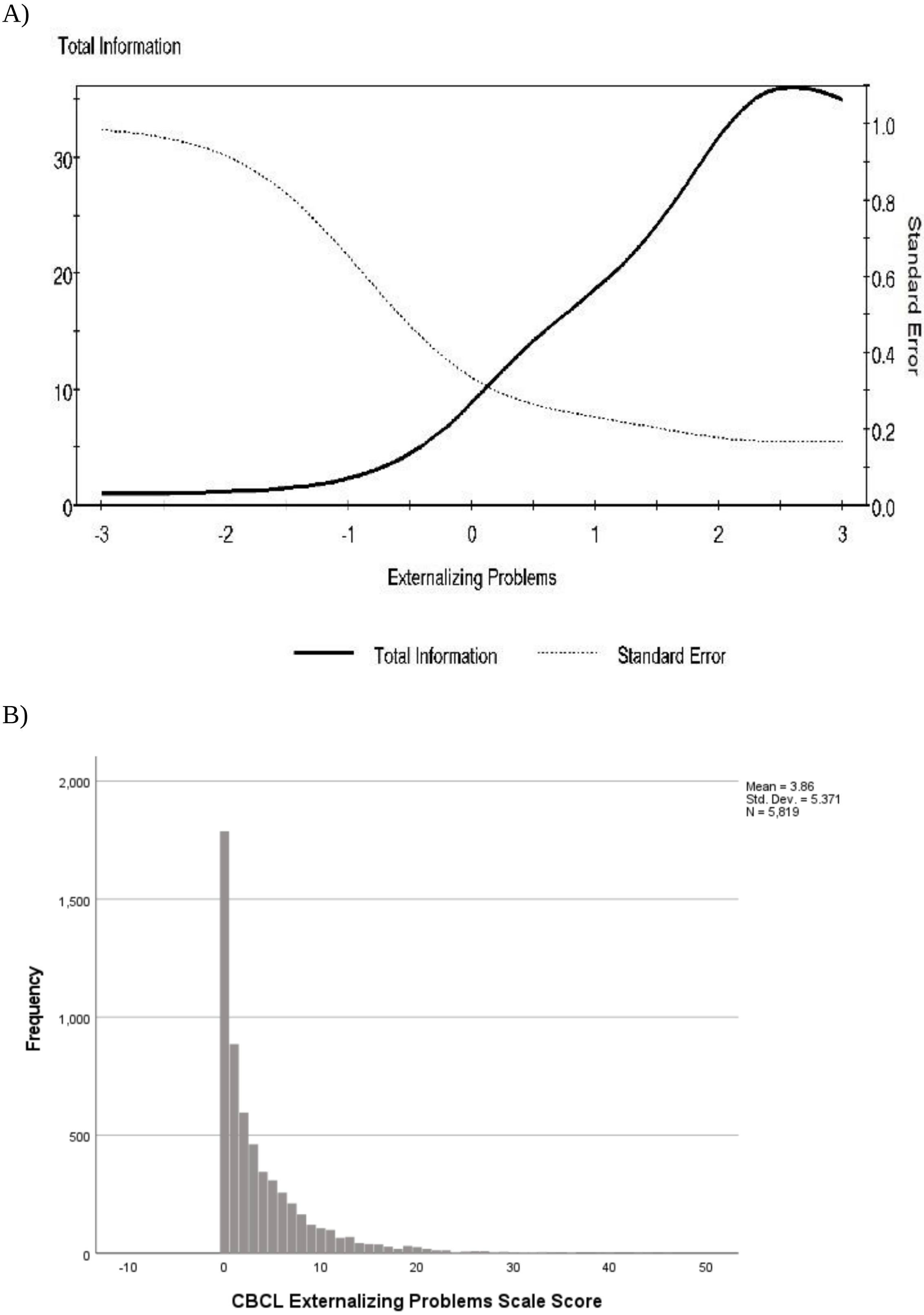
Supplementary Figure 10. A) Total information function / curve (TIF/TIC) for the child behavior checklist Aggressive Behavior syndrome scale. B) Histogram of sum scale scores on the Aggressive Behavior syndrome scale.

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.

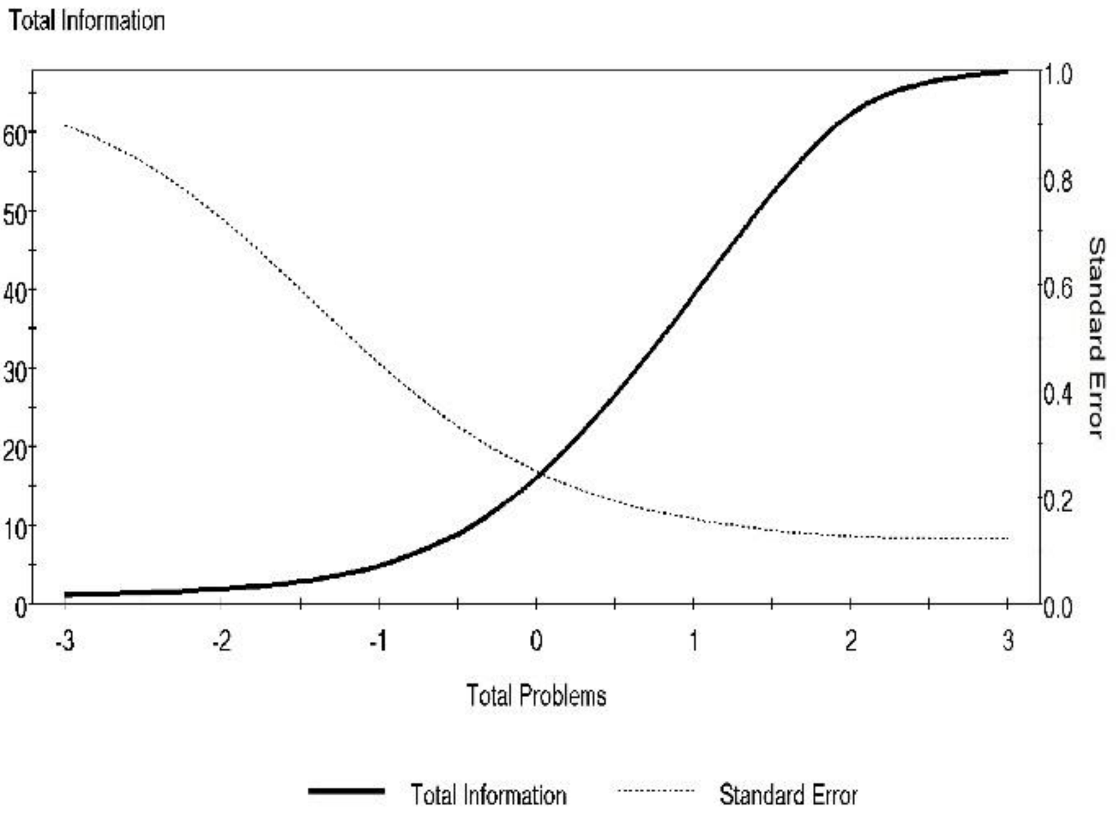
Note. *N* = 5,820. $r_{xx} = 1 - \left({}^{1}/_{I}\right)$. Standard error of the estimate (*SEE*) = $1/\sqrt{I}$.



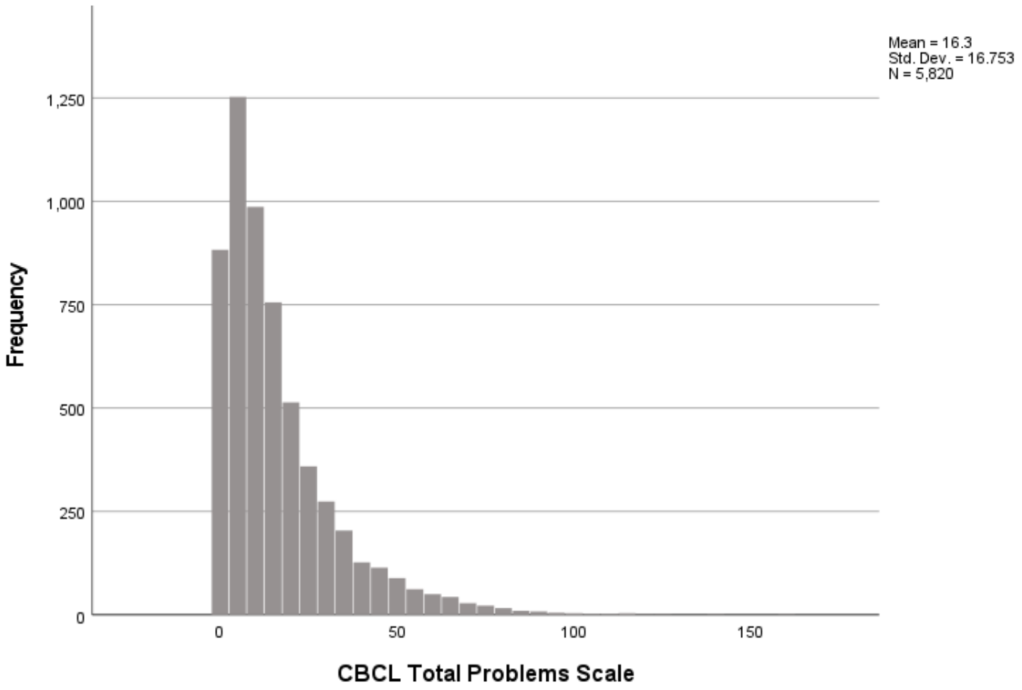
Supplementary Figure 12. A) Total information function / curve (TIF/TIC) for the child behavior checklist Externalizing Problems scale. B) Histogram of sum scale scores on the Externalizing Problems scale.

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

A)



B)



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. $r_{xx} = 1 - (1/I)$. 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].

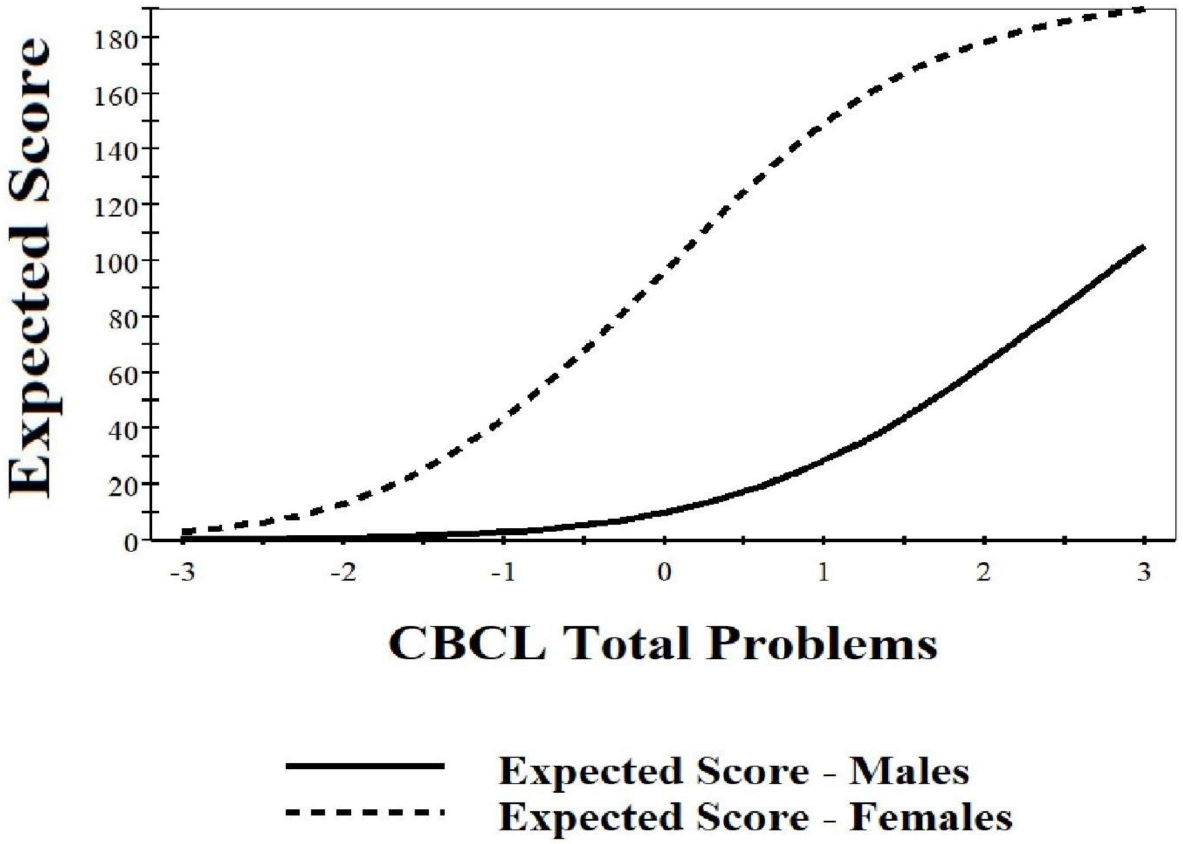
By way of example, we tested for DIF in the Total Problems scale of the CBCL for male and female ABCD participants using the two-year follow-up data. We focused on the Total Problems scale because it has the highest reliability of all the CBCL scales as indexed by Cronbach's α and information values 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 and their Interpretation within a Factor Analytic Framework

Level of invariance	Definition	Interpretation of Invariance	Interpretation of Non-invariance
1. Weak (configural)	equality of factor loadings across the same number of factors	factors have the same substantive interpretation across groups, enabling meaningful comparison of factor variances and covariances	scale items (indicators) are differentially weighted to determine the factors, which are a linear combination of the indicators, meaning that the factors and corresponding raw scores have different substantive interpretations between groups
2. Strong (metric)	equality of factor loadings and intercepts	enables direct comparison of factor means between groups	different response styles operating within each group affect endorsement of item response categories independently of individuals' standing on the underlying factor; sum scale and subscale scores will reflect both true individual/group differences in the underlying construct being measured, but also systematic differences in response styles unrelated to the factor, rendering direct comparisons of raw scores meaningless
3. Strict (scalar)	equality of factor loadings, observed variable intercepts, and error variances	enables meaningful comparison of observed/raw scores between groups	latent variables are not being measured with equivalent precision across groups, with different levels of error variance aggregation, which invalidates comparison of variances and covariances of observed scores



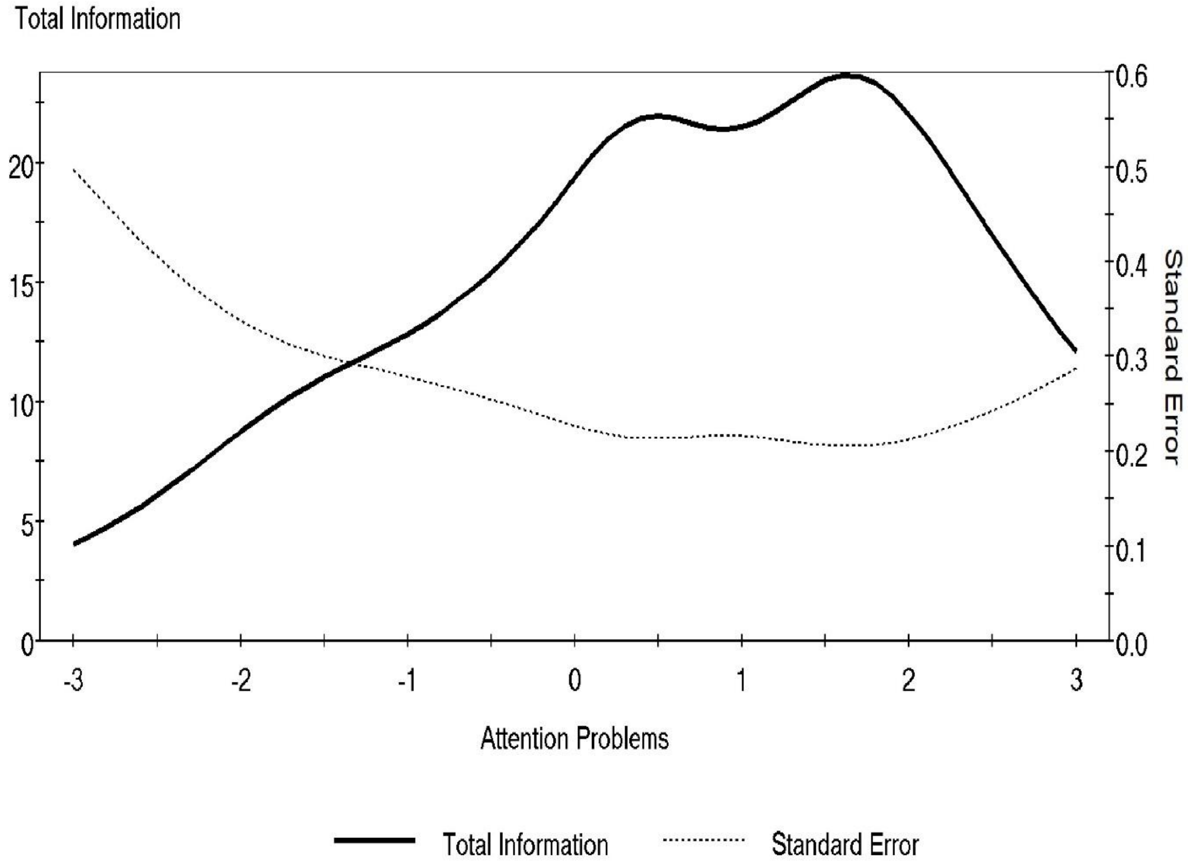
Supplementary Figure 14. Test characteristic curves showing the relationship of expected raw score (*y* axis) as a function of a participants' standing on the CBCL Total Problems latent trait continuum (*x* axis) for males (*n* = 3,025) and females (*n* = 2,795).

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).

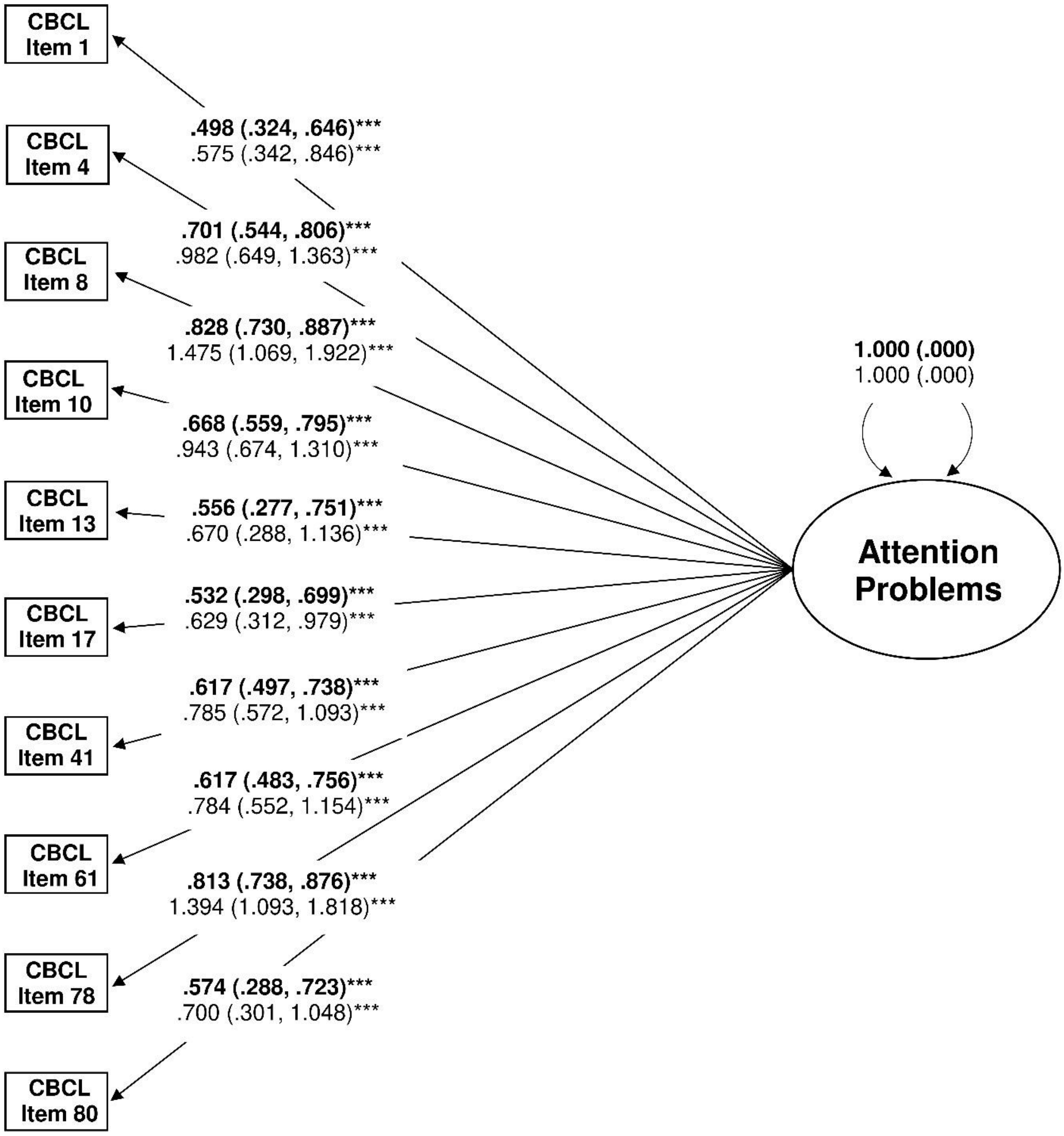
We then began testing FMMs, beginning with the simplest, a one-factor one-class model[54], before moving to one-factor two-class models using the most restrictive and parsimonious FMM (i.e., FMM-1, different latent means only) before progressively relaxing equality constraints on the factor variance-covariance 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 $\geq$ 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 *ΔLL	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*ΔLL = 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.

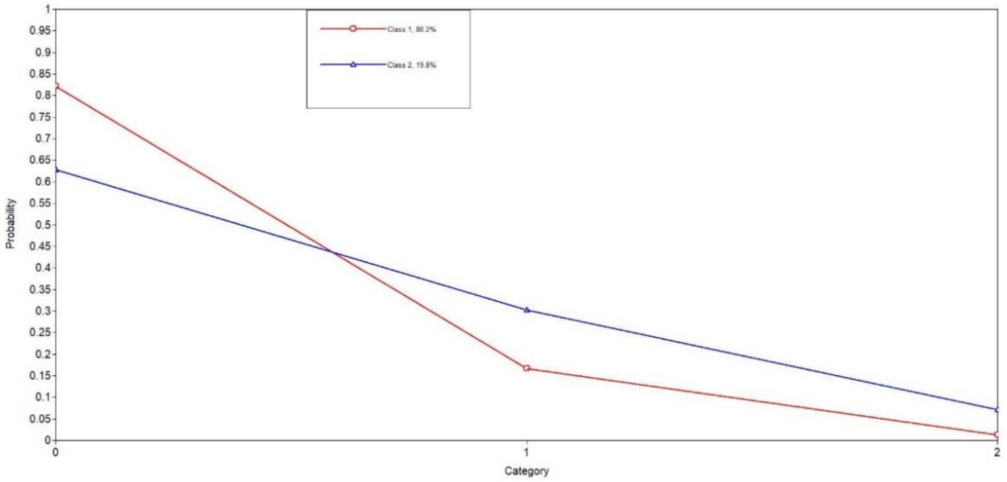
[2] Best loglikelihood values initially obtained using 160 and 32, then replicated using 320 and 64 random starting value perturbations and final stage optimizations.

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

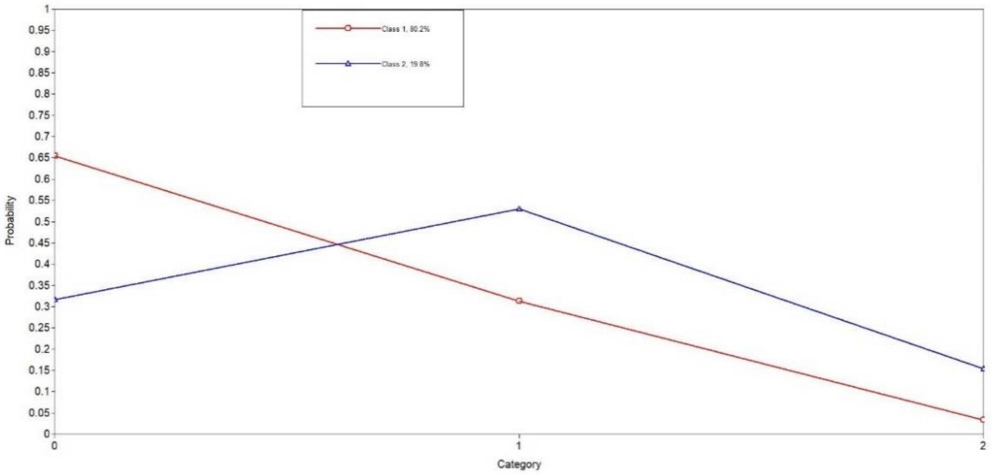
[4] The best log likelihood was not replicated across runs.

Bold typeface indicates preferred model based on fit statistics.

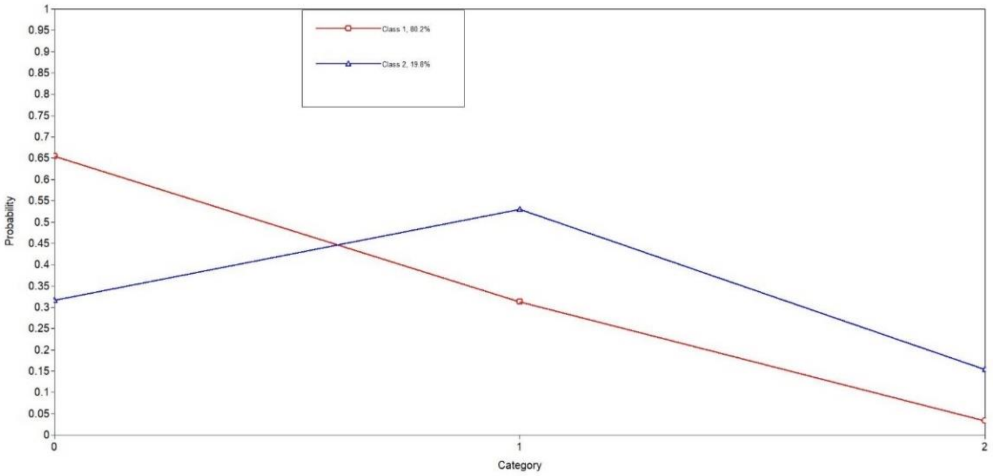
The following models were misspecified and did not converge on trustworthy estimates and therefore the results were not reported for these models: 2C FMM-4; 2C ZI (converged, but had zero cases in the zero-inflated class); 3C FMM-3; 3C FMM-4; 3C ZI FMM-2; 3C ZI FMM-4; 4C FMM-3; 4C FMM-4; 4C ZI FMM-3;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



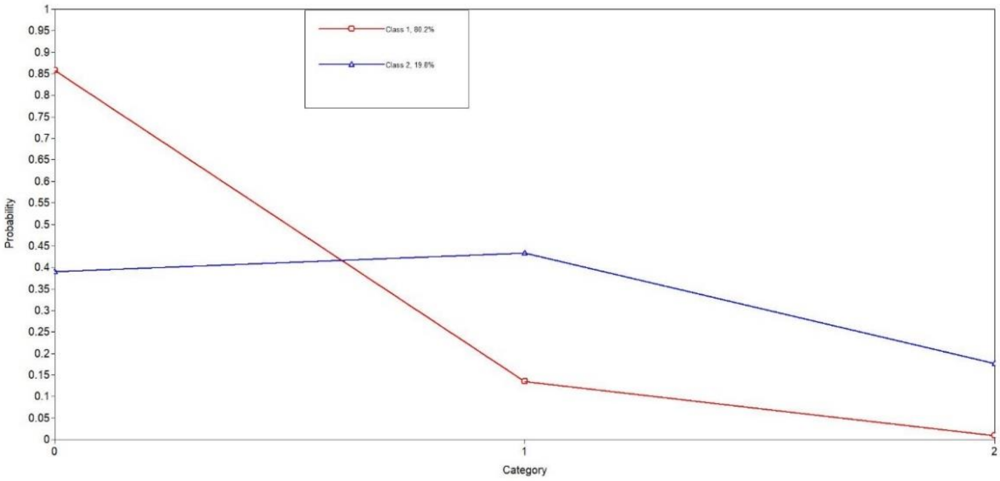
Supplementary Figure 17. Item Probability Plot for CBCL Item 1 “Acts Young” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True*.



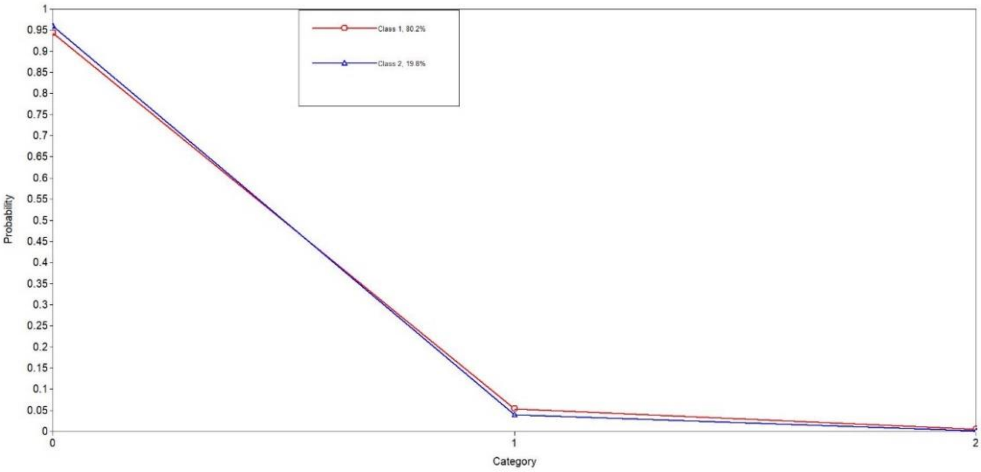
Supplementary Figure 18. Item Probability Plot for CBCL Item 4 “Fails to Finish” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True*.



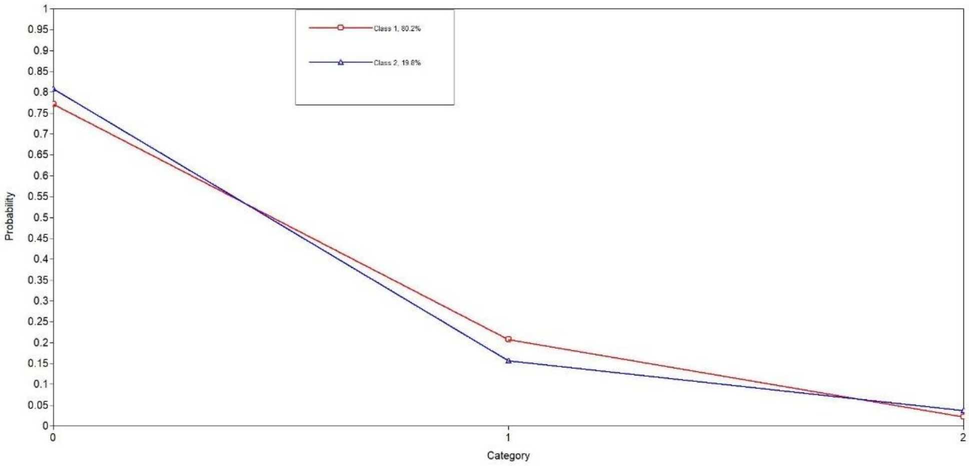
Supplementary Figure 19. Item Probability Plot for CBCL Item 8 “Concentrate” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True*.



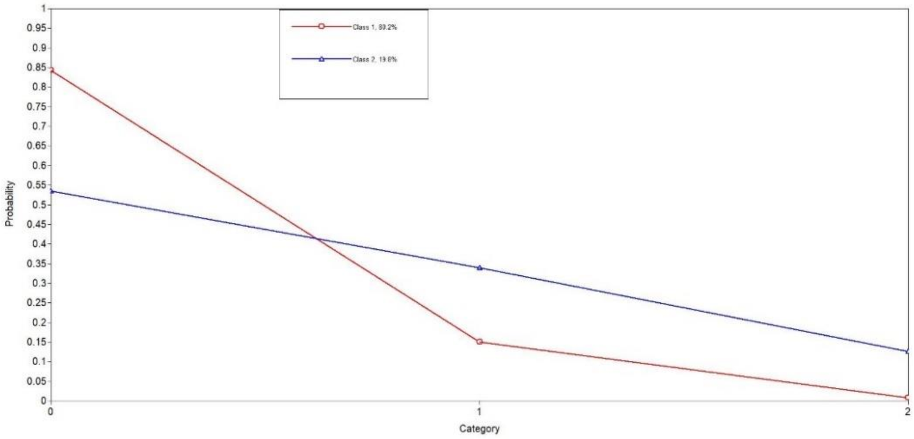
Supplementary Figure 20. Item Probability Plot for CBCL Item 10 “Sit Still” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True*.



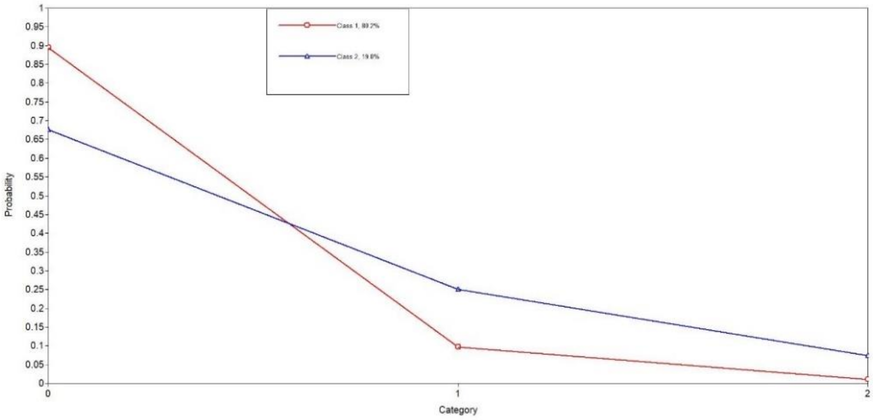
Supplementary Figure 21. Item Probability Plot for CBCL Item 13 “Confused” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True*.



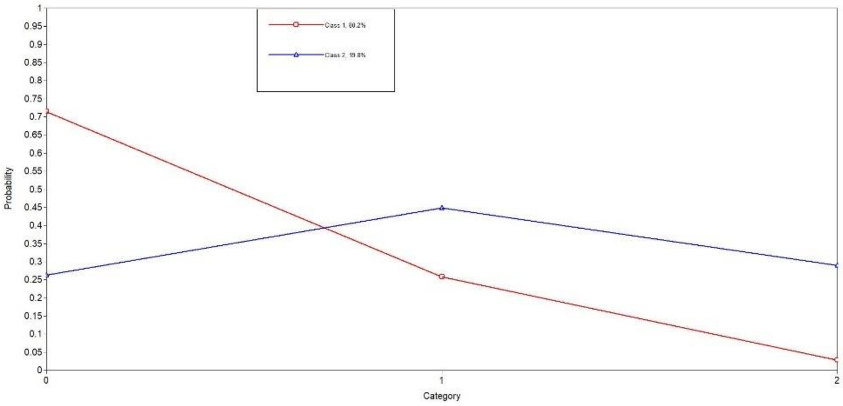
Supplementary Figure 22. Item Probability Plot for CBCL Item 17 “Daydream” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True*.



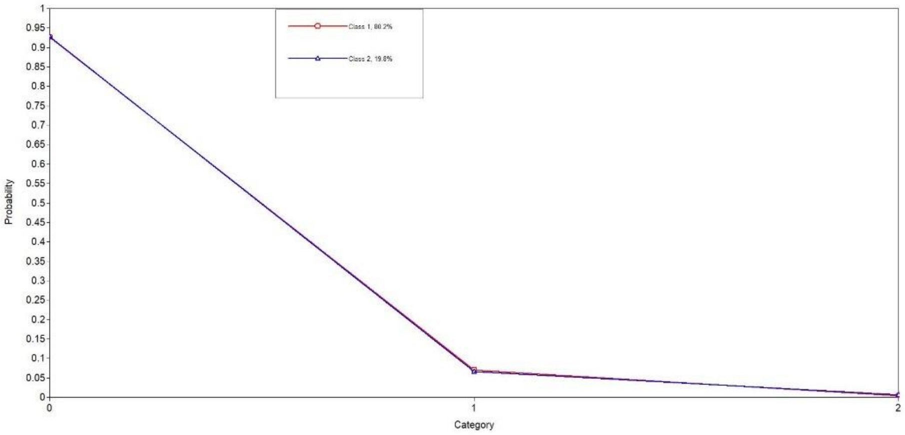
Supplementary Figure 23. Item Probability Plot for CBCL Item 41 “Impulsive” for the Two-Class FMM-3 Model.
Note. *0 = Not True, 1 = Somewhat or Sometimes True, 2 = Very True or Often True*.



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*.

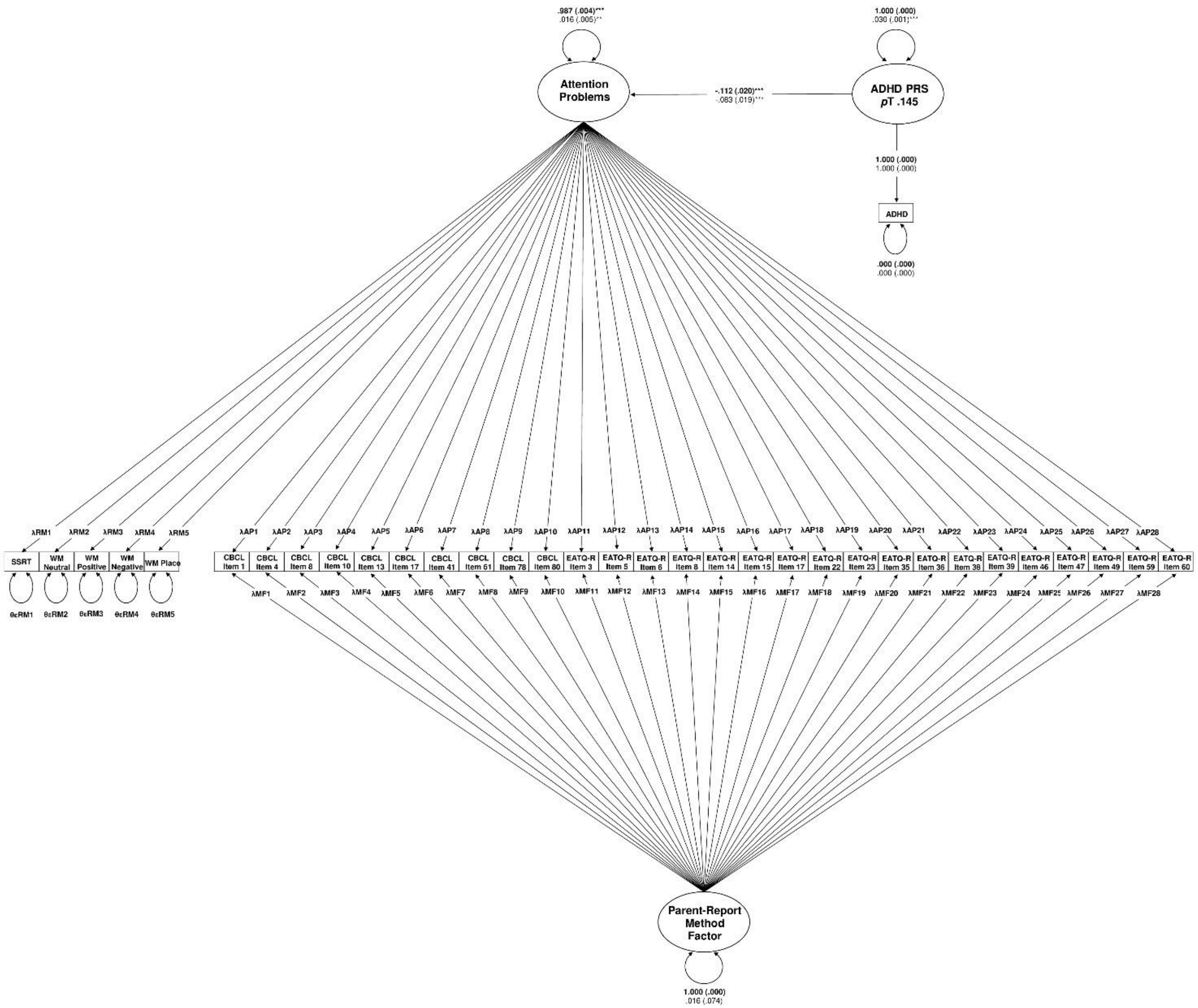
Example 6 – Controlling for Method Variance

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 ($\lambda = .112 - .176$) and modest ($\lambda = .247$, $p < .001$) to very strong ($\lambda = .916$, $p < .001$) 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 ($\lambda = .005$, $p = .897$) to moderately strong ($\lambda = .721$, $p < .001$) (Supplementary Table 9) and this factor did not have statistically significant variance ($\varphi = .016$, $p = .829$), 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 ($\varphi = -.043$, $SE = .026$, $p = .101$). 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 Individ 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).