Article:
Luo, J., Shao, W., Zavlis, O. et al. (2026) Less is more? A hybrid machine learning and psychometric approach to identifying clinically relevant psychopathology in Chinese youth using the Child Behaviour Checklist. Child and Adolescent Psychiatry and Mental Health, 20. 91. ISSN: 1753-2000

https://doi.org/10.1186/s13034-026-01087-4

Child and Adolescent Psychiatry and Mental Health

RESEARCH

Open Access

Less is more? A hybrid machine learning and psychometric approach to identifying clinically relevant psychopathology in Chinese youth using the Child Behaviour Checklist

Jie Luo[1], Wen Shao[2*], Orestis Zavlis[3], Yuanzhen Wu[1], Huanhuan Huang[1], Gaoyang Xu[1], Yanjie Qi[1], Yi Zheng[1] and Fan He[1*]

Abstract

Objective Screening for psychiatric risk in youth at the population level is often constrained by resource limitations and lengthy assessment tools. This study aimed to develop a reduced, psychometrically robust subset of Child Behavior Checklist (CBCL) items to effectively predict transdiagnostic psychiatric morbidity in the youth population.

Methods Data were drawn from a nationally representative sample of 72,109 Chinese youth aged 6–16 years. Initial item screening on the full sample employed unique variance analysis, item–rest correlation tests, and conceptual redundancy checks. A nested case–control subset (approximately 4,500 with diagnoses and 5,000 without) was used for feature selection. Recursive feature elimination with repeated cross-validation was then applied to the nested subset to derive three item sets (n = 35, 69, 98). These were psychometrically evaluated using exploratory graph analysis and confirmatory factor analysis in two age- and gender-stratified samples from the full dataset. Predictive performance was assessed using five machine learning algorithms, trained and tested on a 70/30 split of the nested case–control data.

Results The 35-item and 60-item subsets achieved high diagnostic accuracy (AUC = 0.88–0.89), with performance comparable to the best-performing larger subsets. Items captured transdiagnostic dimensions including Functional Somatic Symptoms, Neurodevelopmental Dysregulation, Affective–Social Withdrawal, Threat Sensitivity and Cognitive–Perceptual Disturbance, and Disinhibited–Irritable Externalising.

Conclusions The reduced CBCL sets demonstrated strong diagnostic utility and psychometric soundness. This scalable tool supports transdiagnostic, data-driven screening of youth psychiatric risk at the population level.

Keywords CBCL, Machine learning , Psychiatric diagnosis, Transdiagnostic, Youth mental health

*Correspondence:
Wen Shao
res.wshao@gmail.com
Fan He
hf981207@163.com
[1]The National Clinical Research Center for Mental Disorders & Beijing Key Laboratory of Mental Disorders, Beijing Anding Hospital & the Advanced Innovation Center for Human Brain Protection, Capital Medical University, De Sheng Men Wai An Kang Hu Tong 5 Hao, Xi Cheng Qu, Beijing 100088, China
[2]Department of Psychology, University of Sheffield, Cathedral Court, 1 Vicar Lane, Sheffield S1 2LT, UK
[3]Department of Psychology and Language Sciences, University College London, Gower Street, London WC1E 6BT, UK

Introduction

Child and adolescent mental health disorders are a growing public health concern worldwide. Globally, roughly 10–20% of youth are affected by mental disorders [1, 2]. In China, the burden of youth psychopathology has risen sharply in recent decades amid rapid social change [2, 3]. These trends underscore an urgent need for early identification and intervention, as untreated childhood mental health problems often persist into adulthood and lead to significant long-term impairment [4].

However, the accurate diagnosis and management of child psychopathology in China face significant barriers. There is a severe shortage of qualified mental health professionals, with fewer than 500 full-time child psychiatrists nationwide as of 2019, most based in major cities [5]. This uneven distribution leaves many families—particularly in rural or under-resourced areas—without access to mental health care. Additionally, structured psychiatric interviews are time-consuming, often requiring hours per child, which is unfeasible in overburdened clinical settings [6, 7]. As a result, many children remain undiagnosed or are only identified once symptoms become severe. Efficient, scalable screening tools are urgently needed to bridge the gap between widespread community needs and limited specialist capacity.

Alternatively, standardised behavioural screening tools can employ threshold-based cut-offs to identify children at risk for psychopathology. The Child Behavior Checklist (CBCL), part of the Achenbach System of Empirically Based Assessment, is among the most widely used parent-report instruments for assessing child and adolescent mental health. The 113-item CBCL assesses broad internalizing and externalizing problems as well as specific psychopathology (e.g., anxious/depressed symptoms). It demonstrates robust psychometric properties across cultures [8–10]. Compared to structured clinical interview, the CBCL is less resource-intensive and suitable for large-scale screening in settings such as schools or primary care, where mental health specialists are limited.

However, the full-item assessment remains resource-intensive; this lack of brevity limits the tool's utility in settings requiring rapid or repetitive psychiatric screening [11]. The broad content introduces redundancy, as many items are correlated and provide conceptually similar information [12]. Additionally, the CBCL's scoring method assumes equal weight for all items, though some symptoms may hold greater clinical significance. This uniform weighting could dilute the predictive value of key indicators, suggesting a more streamlined or weighted approach may enhance efficiency and accuracy [13].

Prior efforts have primarily focused on streamlining screening tools, often with limited evaluation of their clinical utility. One notable effort is the Brief Problem Monitor (BPM), a 19-item version designed for quick tracking of behavioural progress. It yields scores aligned with the CBCL's broad domains and takes only 1–2 min to complete [14]. However, the BPM is intended for ongoing monitoring rather than initial screening, and its brevity omits many symptom areas, limiting its use for broad identification [15]. Other efforts have used psychometric methods—such as factor analysis and item response theory—to create static short forms by selecting items that best represent syndrome constructs or maximise precision [16]. Although the length of these assessments has been successfully reduced, their clinical utility remains limited [17–19]. Meanwhile, screening efforts incorporating clinical outcomes often target specific mental conditions, ignoring the nature of mental illness, which is characterised by high comorbidity and overlapping symptoms [20]. Empirical evidence increasingly supports the assumption that a latent general psychopathology, the '*p*' factor, underlies all mental conditions [21, 22]. Consequently, there is a critical need for brief, transdiagnostic screening tools designed to detect general psychiatric morbidity.

Our study addresses this need by employing a hybrid approach that integrates machine learning (ML) and classical psychometrics to refine the CBCL. ML has been extensively used in the field of computational psychiatry [23], excelling at classification-based feature selection and maximizing predictive accuracy for psychiatric diagnosis [24]. By integrating classical psychometric methods, our approach ensures high diagnostic accuracy with demonstrated structural stability and clinical validity. Applying this approach to a nationally representative Chinese sample, we derived and evaluated shortened, clinically relevant CBCL subsets that maximize brevity while minimizing information loss. Finally, we propose preliminary clinical guidelines to support screening and referral decisions in applied settings. As the analyses are exploratory in nature, no formal hypotheses were specified. Ultimately, this yields a screening instrument that successfully balances the practical demand for brevity with the clinical necessity of psychometric soundness.

Methods

Statistical analyses

Analyses were conducted in R (version 4.4.3) using the R packages psych (2.5.3), EGAnet (2.3.0), caret (7.0-1), caretEnsemble (4.0.1), randomForest (4.7-1.2), xgboost (1.7.9.1), C50 (0.2.0), e1071 (1.7-16), tidyverse (2.0.0), ggplot2 (3.5.2), and qgraph (1.9.8).

Assessment tool

Child behaviour checklist (CBCL)

The CBCL is a carer rating questionnaire assessing a range of behavioural and mental anomalies [8].

It includes 120 items covering eight areas: aggressive behaviour, anxiety and depression, attention problems, rule breaking behaviour, somatic complaints, social problems, thought problems, withdrawal, and depression. The CBCL has been revised to align with DSM-oriented subscales, including depressive problems (e.g., item 35: 'Feels worthless'), anxiety problems (e.g., item 112: 'Worries a lot'), somatic problems (e.g., item 56b: 'Headaches'), ADHD (e.g., item 8: 'Can't concentrate'), oppositional defiant problems (e.g., item 86: 'Stubborn'), and conduct problems (e.g., item 15: 'Cruel to animals'). The item 113 was excluded for the non-specificity. The items were rated on three ordinal levels from 0 absent to 2 very often. Its validity has been examined and supported in Chinese youth [25].

Feature reduction and model evaluation

The initial item pool comprised 119 CBCL items. A complete analysis flowchart presented in Fig. 1.

This study utilised a secondary dataset obtained from a national mental health survey conducted in five major regions of China. Participants were recruited using stratified random sampling, which has been described elsewhere [2]. After excluding cases with missing data, 72,107 participants remained for analysis. Psychometric examinations were conducted on the full sample using the selected features. The age distribution was 11.47 years (SD = 2.82), with 880 cases missing age information. The proportion of female participants in the machine learning sample was 41.50% across both patient and control groups, and 49.51% in the psychometric analysis sample, with 100 missing values. Psychiatric diagnoses were made by trained clinicians using the MINI-KID (Mini International Neuropsychiatric Interview for Children and Adolescents), see the distribution of diagnoses on Supplementary Fig. 1. Supervised machine learning analyses were conducted on a subsample of 4,572 youths with psychiatric diagnoses and 5,000 age- and gender-matched controls from two southern regions of China. This regional focus was based on the administrative structure of the parent project's data availability; at the time of this targeted machine learning analysis, the fully

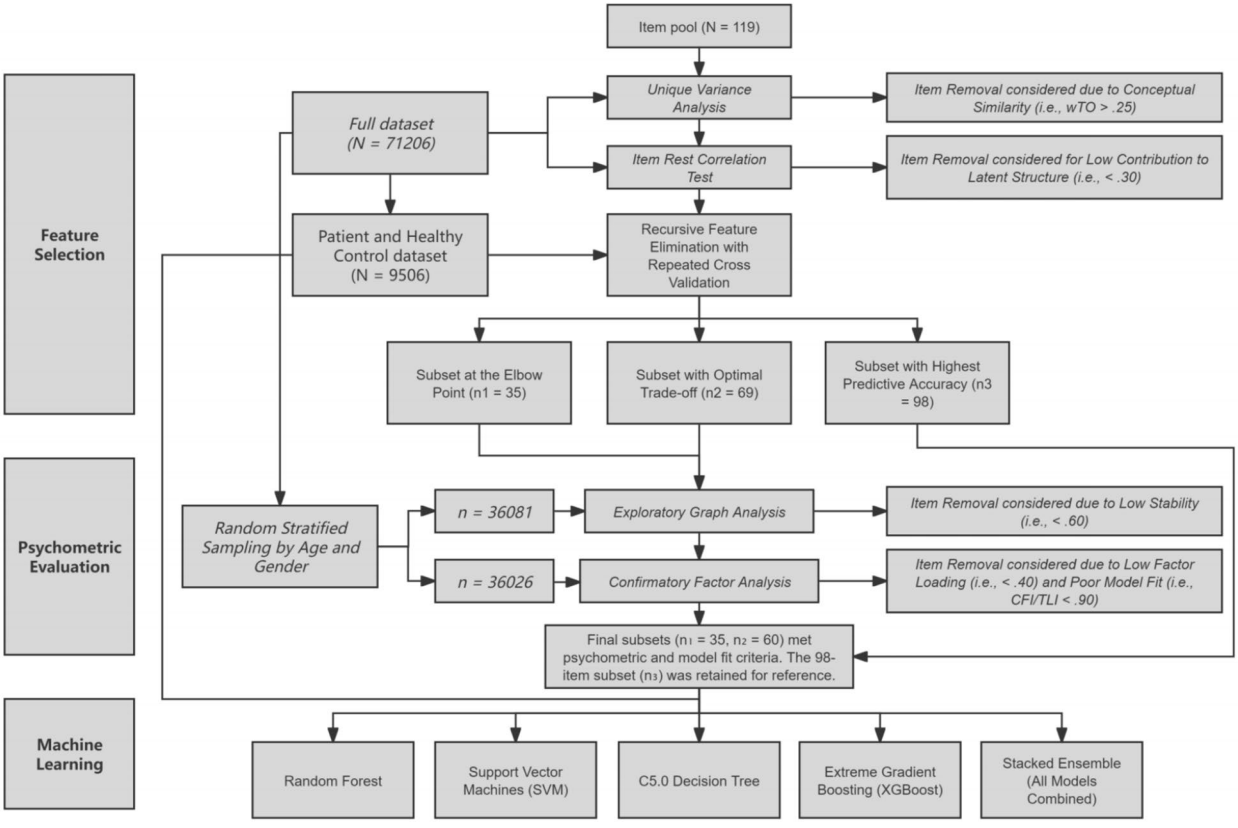

Fig. 1 Technique flow chart. This figure illustrates the sequential feature selection, psychometric evaluation, and machine learning workflow. Starting from an item pool of 119 CBCL items, three subsets were derived using recursive feature elimination (RFE) with repeated cross-validation: the elbow-point subset ($n_1 = 35$), the optimal trade-off subset ($n_2 = 69$), and the subset with the highest predictive accuracy ($n_3 = 98$). The 69-item subset underwent psychometric screening, and 9 items were removed due to low structural stability in exploratory graph analysis (EGA). An additional removal was based on low factor loadings and poor fit in confirmatory factor analysis (CFA). The final retained subsets ($n_1 = 35$, $n_2 = 60$) demonstrated acceptable psychometric properties and model fit (CFI/TLI > 0.90). All subsets were subsequently evaluated using five machine learning models

verified clinical diagnostic data required for the nested case–control matching (i.e., the 'ground truth' labels essential for algorithm training) were systematically complete and accessible for these specific administrative regions. The age distribution for this subsample was as follows: patients (M = 11.8, SD = 2.89) and healthy controls (M = 11.4, SD = 2.95).

Dimensionality reduction

Our first set of analyses aimed to reduce the 119-item CBCL instrument to a set of less items using a multi-step feature selection approach that included both clinical decision-making (by two clinicians) and four statistical tests, which are described in turn. First and foremost, a Unique Variable Analysis (UVA) was conducted on the full sample to detect redundant items. UVA calculates the so-called 'weighted topological overlap (wTO)' scores, which indicate the extent to which two variables exhibit similar associations with other variables (in which case they are considered redundant as they carry the same information). As per simulation studies, variables were considered when they scored: wTO > 0.25 (reference) [26]. Second, item-rest correlation (IRC) analyses were conducted in parallel to identify items weakly correlated (r < 0.30) with the latent structure [27]. Third, near-zero variance detection eliminated items with minimal variability across all sample [28]. Finally, Recursive Feature Elimination with Repeated Cross-Validation (RFERCV) was used via a Random Forest (RF) model, selected for its robustness to overfitting and multicollinearity. The main purpose of the RFERCV analysis was to identify the optimal item subsets from the pool of items that survived the prior IRC analyses, aiming to maximize predictive performance while minimizing information loss. The outcome variable used to compute the model's accuracy was the clinical status of the youth (i.e., the presence versus absence of at least one formal clinical diagnosis). RF model was optimized through simulations (ntree: 300–2000; mtry: 5–30), with final parameters (ntree = 1000, mtry = 9) chosen based on minimal cross-validated error. Our RFERCV utilized five-fold cross-validation repeated 20 times, evaluating subsets across the entire range of items that survived the IRC analyses. This yielded three subsets: a concise 35-item set at the elbow point on the accuracy curve, an intermediate 69-item set with minimal accuracy loss (< 0.005), and a 98-item set with the highest accuracy. The 35- and 69-item subsets were retained for psychometric analysis, as the 98-item set offered marginal gains and could be further reduced via exploratory graph analysis (see Supplementary Fig. 2). Two clinicians independently reviewed the statistical results, prioritizing theoretical and clinical relevance for item retention or removal.

Psychometric validation

Our second set of analyses aimed to validate the ensuing reduced questionnaire using several psychometric methods. The factor exploration used Exploratory Graph Analysis (EGA), a factor-analytic methodology that can infer latent variables from a set of observed variables using modern graph-based methods [29]. We chose this method over traditional exploratory factor analysis because it automates the dimension retention process without relying on subjective heuristics, demonstrates higher accuracy when latent factors are highly correlated, features a built-in framework for rigorously assessing item stability, and provides clear visual interpretability [29]. This analysis was applied on a stratified random sample (N = 36,081; mean age = 11.49, SD = 2.82; 49.51% female; 453 missing age, 62 missing gender), using polychoric correlations (to account for the ordinal nature of our data). To ensure that sampling variability did not influence our results, we repeated the EGA procedure in 2,000 non-parametric bootstrapped samples, estimating the frequency by which dimensions were replicated and items within them were consistent. Items with low stability (e.g., assigned to the same dimension < 60% of bootstraps) or communities with fewer than three items were iteratively removed [30] until all retained items demonstrated acceptable stability (see Supplementary Table 1).

Confirmatory Factor Analysis (CFA) then validated the EGA-derived structures on a separate stratified random sample (n = 36,026; mean age = 11.49, SD = 2.82; 49.50% female; 427 missing age, 38 missing gender). Three models were tested: a simple factor model with correlated latent factors, a bifactor model with a general factor and uncorrelated specific factors, and a higher-order model where specific factors load onto a general factor. All models used the Weighted Least Squares Mean and Variance Adjusted (WLSMV) estimator for ordinal, non-normal data, with latent factor variances fixed at 1. Model fit was assessed using chi-square (descriptive, due to sample size sensitivity), Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) ≥ 0.90, and Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR) < 0.08, with scaled CFI, TLI, and robust RMSEA/SRMR reported. For the simple and higher-order models, internal consistency and convergent validity were evaluated using Composite Reliability (CR > 0.70) and Average Variance Extracted (AVE > 0.50). For the bifactor model, additional indices included Explained Common Variance (ECV), Percent of Uncontaminated Correlations (PUC), Omega Total, Omega Hierarchical (ω_h), Construct Replicability (H), and Factor Determinacy (FD), with thresholds of ECV > 0.70, ω_h > 0.80, H > 0.70, and FD > 0.90 indicating a strong general factor.

To evaluate measurement invariance (MI) across gender and age cohorts (preadolescence: 6–11 years; adolescence: 12–16 years), multi-group confirmatory factor analyses (MG-CFA) were conducted on the 35- and 60-item subsets. MI testing focused solely on the first-order factor model; establishing equivalence at this primary measurement level is a fundamental prerequisite, whereas complex categorical structures (e.g., higher-order and bifactor models) are highly prone to computational non-convergence and inadmissible estimates. Following citation's approach for categorical indicators, configural, threshold, metric, and scalar invariance were tested sequentially [31]. Given the chi-square test's known oversensitivity in large samples, invariance was evaluated via changes in alternative fit indices ($\Delta CFI \leq 0.010$, $\Delta RMSEA \leq 0.015$).

Machine learning

Finally, our last set of analyses aimed to evaluate the performance of the ML models in predicting the overall clinical status of the youth (i.e., the presence versus absence of any formal psychiatric diagnosis). The rationale for selecting this outcome is twofold. Clinically, a streamlined CBCL prioritised by its ability to predict clinical status can facilitate timely referrals and benefit clinical research. Statistically, predicting specific disorders within a community-based sample introduces severe class imbalance due to low base rates, which risks algorithm overfitting—a problem further exacerbated by the high heterogeneity within and across psychiatric entities. Model validation involved testing the 35- and 69-item subsets using the following machine learning algorithms: Random Forest (RF), Support Vector Machines (SVM), eXtreme Gradient Boosting (XGB), C5.0, and stacked ensemble models. SVMs are margin-based classifiers that construct optimal separating hyperplanes, particularly effective in complex, high-dimensional spaces. XGB is a gradient boosting algorithm that builds sequential decision trees optimised to minimise prediction error, offering high accuracy and computational efficiency. C5.0 is a rule-based decision tree algorithm that is interpretable and well-suited for categorical data. Stacked models combine predictions from multiple base learners using a meta-learner to improve predictive performance. To facilitate interpretation, feature importance scores were computed and ranked using importance scores obtained from RF, SVM, and XGBoost models. Additionally, the median importance scores were reported for each dimension across different item subsets to evaluate dimensional trends in diagnoses.

Models were trained on a 7:3 train-test split using five-fold cross-validation repeated 20 times, with hyperparameters tuned via grid search testing five values per parameter. All models were evaluated using a comprehensive set of performance metrics to assess classification efficacy. The primary metric reported was the area under the receiver operating characteristic curve (AUC), which quantifies the model's ability to distinguish between classes. To provide a more detailed understanding of model performance, additional statistics were reported, including accuracy, precision, recall, F1-score, sensitivity, and specificity. These metrics offer insights into various aspects of the model's predictive capabilities and the impact of variable selection.

The optimal cut-off score for identifying individuals at potential risk was determined through receiver operating characteristic (ROC) analysis. Various classification thresholds were evaluated, with corresponding metrics—area under the curve (AUC), sensitivity, specificity, and accuracy—computed across different item subsets. The optimal threshold was identified using Youden's index, a method that balances sensitivity and specificity by maximising the sum of these two metrics minus one, thereby optimising overall classification performance. Generally, AUC, sensitivity, and specificity values exceeding 0.80 are indicative of strong discriminative capability.

Results

Feature selection

The UVA suggested that the following item-pairs might be redundant: I53 (overeating) and I55 (overweight); I44 (bite fingernails) and I98 (thumb sucking); I20 (destroy own things) and I21 (destroy things from family or others); and I107 (wets self during the day) and I108 (wets the bed), each exhibiting significant redundancy with weighted topological overlap (wTO) values exceeding 0.30. Additionally, item pairs I56C and I56G, as well as I76 and I100, demonstrated moderate redundancy ($wTO > 0.25$). The IRC analysis revealed that 13 items had low loadings onto the latent structure ($r < 0.30$) and were subsequently excluded from further analyses (see Supplementary Table 1 for details). Following the removal of these items, all remaining variables exhibited unique variance, resulting in a refined set of 100 items for the RFERCV analysis.

The RFERCV analysis revealed that the predictive accuracy increased markedly from 1 to 20 items and stabilised beyond 35 items (see Supplementary Fig. 2 for details). The standard deviations of accuracy and kappa stabilised after 15 items. Three subsets were selected based on a balance between screening feasibility and diagnostic accuracy. The 35-item subset yielded an accuracy of 0.826 (SD = 0.0086) and a Kappa of 0.652 (SD = 0.0171), representing a concise and efficient solution. The 69-item subset was chosen as a strong intermediate option, achieving an accuracy of 0.832 (SD = 0.0086) and a Kappa of 0.663 (SD = 0.0172). The 98-item subset demonstrated the highest overall predictive performance, with an accuracy

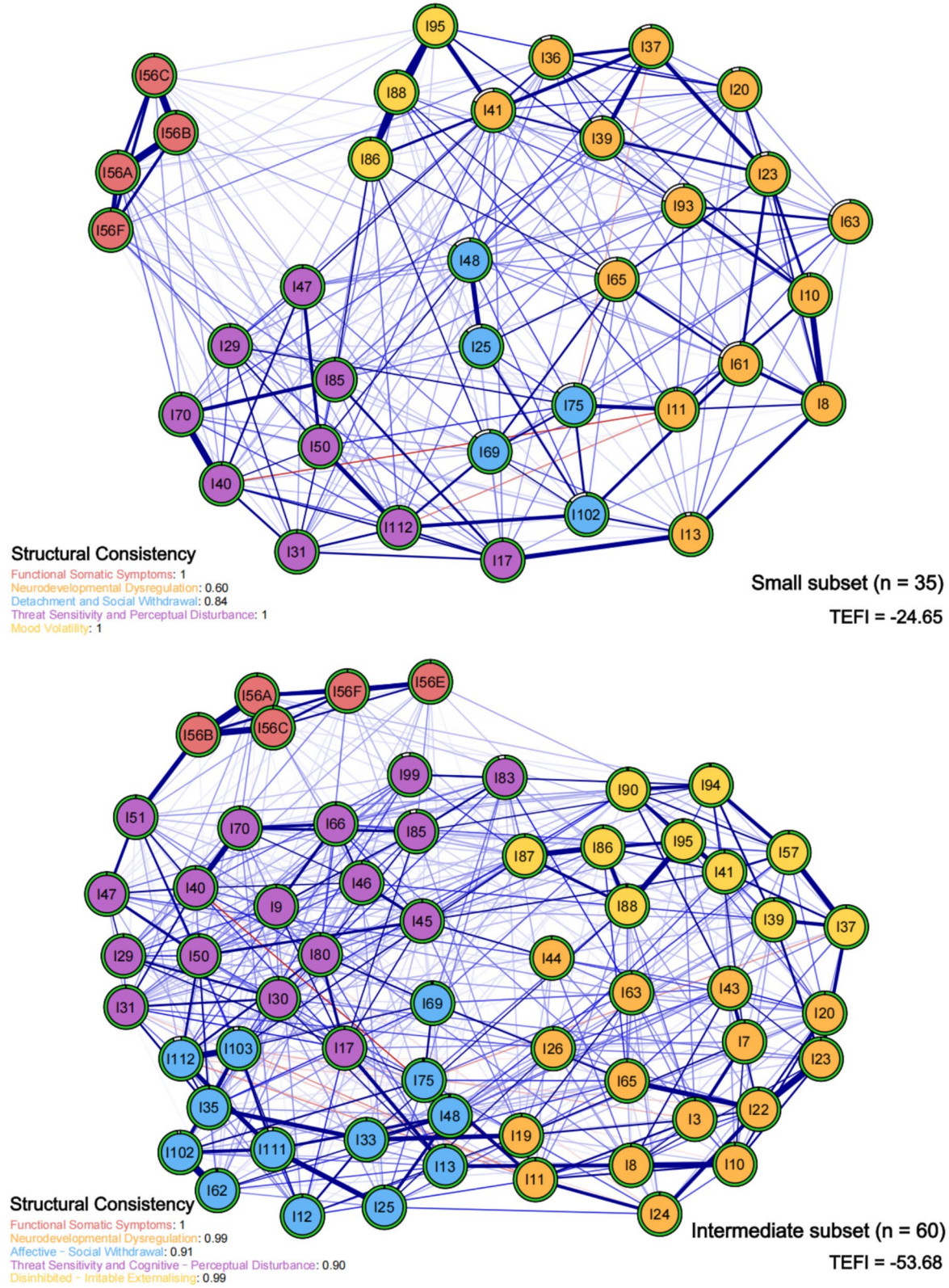

Fig. 2 (See legend on next page.)

(See figure on previous page.)

Fig. 2 Network communities of the small and intermediate item subsets. Node colours were determined by empirical clustering of symptoms. The green outer ring of each node represents item stability, with a complete ring indicating perfect stability across bootstrap replications. The structural consistency value displayed in the lower left denotes the reliability with which the dimensional structure was recovered across bootstrap samples—higher values indicate better consistency. The TEFI index assesses how well the estimated structure fits the observed data—lower values signify better fit. Item codes and descriptions ordered by dimensions: I56A (aches or pains), I56B (headaches), I56C (nausea, feels sick), I56E (rashes or other skin problems), I56F (stomach aches); I3 (argues a lot), I7 (bragging, boasting), I8 (cannot concentrate, cannot pay attention for long), I10 (cannot sit still, restless, or hyperactive), I11 (clings to adults or is too dependent), I19 (demands a lot of attention), I20 (destroys own things), I22 (disobedient at home), I23 (disobedient at school), I24 (does not eat well), I26 (does not seem to feel guilty after misbehaving), I36 (get hurts a lot, accident-prone), I43 (lying or cheating), I44 (bites fingernails), I63 (prefers being with older children), I65 (refuses to talk); I12 (complains of loneliness), I13 (confused or seems to be in a fog), I25 (does not get along with other children), I33 (feels or complains that no one loves them), I35 (feels worthless or inferior), I48 (not liked by other children), I61 (poor school work), I62 (poorly coordinated or clumsy), I69 (secretive, keeps things to self), I75 (too shy or timid), I102 (underactive, slow moving, or lacks energy), I103 (unhappy, sad, or depressed), I111 (withdrawn, does not get involved with others), I112 (worries); I9 (cannot get mind off certain thoughts), I17 (daydreams or gets lost in thoughts), I29 (fears certain animals, situations, or places other than school), I30 (fears going to school), I31 (fears they might think or do something wrong), I40 (hears sounds or voices that are not there), I45 (nervous, high-strung, or tense), I46 (nervous movements or twitching), I47 (nightmares), I50 (too fearful or anxious), I51 (feels dizzy or lightheaded), I66 (repeats certain acts over and over; compulsions), I70 (sees things that are not there), I80 (stares blankly), I83 (stores up too many things they do not need), I85 (strange ideas), I99 (smokes, chews, sniffs tobacco, or uses e-cigarettes); I37 (gets in many fights), I39 (hangs around with others who get in trouble), I41 (impulsive or acts without thinking), I57 (physically attacks people), I86 (stubborn, sullen, or irritable), I87 (sudden changes in mood or feelings), I88 (sulks a lot), I90 (swearing or obscene language), I93 (talks too much), I94 (teases a lot), I95 (temper tantrums or hot temper)

of 0.833 (SD = 0.0085) and a Kappa of 0.666 (SD = 0.0170), although its marginal gain over smaller subsets was negligible. Among the 98 variables, the ten most influential items for predicting psychiatric diagnoses—based on average importance scores—were: physical pain (I56A), hyperactivity (I10), inattention (I8), shyness or timidity (I75), nausea (I56C), poor school performance (I61), frequent fighting (I37), peer rejection (I48), specific fears (I29), and impulsivity (I41).

Psychometric examination

Small subset

EGA identified a five-factor structure within the 35-item subset in 71% of bootstrapped samples (median = 5; 95% CI: 4–6). Three dimensions—Functional Somatic Symptoms (FSS), Threat Sensitivity and Perceptual Disturbance (TSPD), and Mood Volatility (MV)—showed perfect structural consistency and item stability. Neurodevelopmental Dysregulation (ND) and Detachment and Social Withdrawal (DSW) displayed slightly lower, yet acceptable, stability. Items I41 (impulsivity), I48 (not liked by peers), and I75 (shy/timid) showed cross-domain associations in 14–21% of bootstrapped samples (Fig. 2).

All three factor models (five-factor, higher-order, bifactor) demonstrated acceptable fit, though chi-square values were elevated, likely due to sample size. The bifactor model fits the data only marginally better than the others (Table 1).

In the five-factor model, most item loadings exceeded 0.60, with exceptions including I11 (0.46), I63 (0.47), and I75 (0.59). Internal consistency was high (CR = 0.94), though convergent validity was marginal (AVE = 0.49). The five-factor solution remains invariant across gender and partially invariant across preadolescence and adolescence. Specifically, a scalar invariance violation was observed, with a reduction in CFI of 0.012, exceeding the widely accepted threshold of ΔCFI < 0.01. This suggests that age-related differences exist in how parents perceive and report specific symptoms, see Supplementary Tables 2a-b for details.

The higher-order model showed strong loadings (> 0.80) on the general factor, except FSS (0.63), and acceptable validity (CR = 0.96, AVE = 0.51). The bifactor model showed strong general factor saturation (ECV = 0.69, ω_h = 0.82), though DSW and MV had low specific factor reliability (ω_h < 0.20), suggesting they may be weakly defined once the general factor is accounted for. While the general factor captures substantial shared variance, not all specific factors appear robust enough for confident interpretation. Thus, the bifactor model remains informative but should be interpreted with caution. Detailed indices for each factor are presented in Supplementary Tables 3a-d.

Intermediate subset

EGA identified five factors within the 60-item subset, with some dimensional refinement. DSW became Affective Social Withdrawal (ASW), TSPD became Threat Sensitivity and Cognitive Perceptual Disturbance (TSCPD), and MV was relabelled Disinhibited-Irritable Externalising (DI-Ext) (Fig. 2). All communities demonstrated high structural consistency (> 0.90).

All factor solutions again demonstrated acceptable fit, with the bifactor model showing the best overall performance (Table 2). The five-factor solution demonstrated full invariance across gender and age group (Supplementary Tables 2c-d). Inter-factor correlations in the five-factor model ranged from 0.44 to 0.88, with FSS less strongly correlated with others. Loadings were generally strong (> 0.60), with few exceptions (Supplementary Tables 4a-4d). CR was high (0.96), AVE marginal (0.47).

In the higher-order model, all constructs loaded strongly onto the general factor (0.65–0.93), with varying

Table 1 Model fit indices for different factor solutions (35-item subset)

Model	χ^2	CFI	TLI	RMSEA	SRMR
Five-factor model	28152***	0.94	0.93	0.071 (0.070–0.072)	0.050
Bifactor model	24862***	0.95	0.94	0.067 (0.067–0.068)	0.048
Higher-order model	30529***	0.93	0.93	0.072 (0.071–0.073)	0.054

CFI Comparative Fit Index, *TLI* Tucker–Lewis Index, *RMSEA* Root Mean Square Error of Approximation, *SRMR* Standardised Root Mean Square Residual. Due to the ordinal nature and non-normal distribution of the data, scaled statistics were reported for χ^2, CFI, and TLI, and robust estimates were reported for RMSEA

*** $p < 0.001$

Table 2 Model fit indices for different factor solutions (60-item subset)

Model	χ^2	CFI	TLI	RMSEA (90% CI)	SRMR
Five-factor model	69846***	0.92	0.91	0.064 (0.064–0.065)	0.051
Bifactor model	68453***	0.92	0.91	0.061 (0.061–0.062)	0.050
Higher-order model	77980***	0.91	0.90	0.065 (0.065–0.066)	0.055

CFI Comparative Fit Index, *TLI* Tucker–Lewis Index, *RMSEA* Root Mean Square Error of Approximation, *SRMR* Standardised Root Mean Square Residual. Due to the ordinal nature and non-normal distribution of the data, scaled statistics were reported for χ^2, CFI, and TLI, and robust estimates were reported for RMSEA

*** $p < 0.001$

explained variance ($R^2 = 0.42–0.86$), especially low for FSS. CR and AVE were high (0.89 and 0.70, respectively).

In the bifactor model, most items loaded more strongly on the general factor (range: 0.34–0.73). Highest item loadings on specific factors included I56A (FSS, 0.65), I22 (ND, 0.61), I111 (ASW, 0.56), I40 (TSCPD, 0.56), and I88 (DI-Ext, 0.50). Some items loaded more on specific factors than the general factor, and several DI-Ext items showed negative loadings, suggesting limited unique variance. General factor saturation was strong (ECV = 0.73, Ω Total = 0.95, $\omega_h = 0.87$), with high construct replicability (0.97). Specific factor saturation was low for DI-Ext (ECV = 0.17, $\omega_h = 0.06$).

Machine learning models

Model performance

Table 3 presents the classification performance of five machine learning models across three item subsets: 35-item, 60-item, and 98-item sets. The 98-item subset served as the reference group, having been identified through RFERCV as the most accurate model-derived configuration. Across all subsets, model performance was comparable, with accuracy values ranging from 0.81 to 0.83 and area AUC scores between 0.88 and 0.90. The stacked ensemble model consistently demonstrated the strongest or equivalent performance across most metrics, including recall, F1-score, and AUC—particularly in the 60- and 98-item subsets.

In the 35-item subset, the stacked model achieved the highest recall (0.83), F1-score (0.82), and AUC (0.89), with balanced sensitivity and specificity (0.83; 95% BCI = 0.82–0.85). Both the RF and C5.0 models also performed well, whereas the SVM showed slightly lower sensitivity (0.79) and recall (0.79). Model performance remained stable in the 60-item subset. The stacked model again yielded strong results (AUC = 0.89; sensitivity and specificity = 0.83), while RF, C5.0, and XGBoost demonstrated similarly high discriminative ability across

Table 3 Performance of various machine learning models in different item subsets

Models	Accu	Pre	RC	F1	Sens (95% BCI)	Spec (95% BCI)	AUC (95% BCI)
35-item subset							
RF	0.83	0.82	0.81	0.81	0.81 (0.82–0.86)	0.84 (0.82–0.85)	0.89 (0.88–0.90)
C5.0	0.82	0.81	0.82	0.81	0.82 (0.81–0.83)	0.83 (0.81–0.84)	0.89 (0.88–0.90)
SVM	0.81	0.80	0.79	0.80	0.79 (0.79–0.82)	0.83 (0.79–0.82)	0.88 (0.87–0.90)
XGBoost	0.81	0.81	0.77	0.79	0.77 (0.79–0.81)	0.84 (0.79–0.81)	0.88 (0.87–0.89)
Stack	0.82	0.80	0.83	0.82	0.83 (0.82–0.85)	0.82 (0.82–0.85)	0.89 (0.88–0.91)
60-item subset							
RF	0.82	0.80	0.82	0.81	0.82 (0.81–0.85)	0.82 (0.81–0.84)	0.89 (0.88–0.90)
C5.0	0.82	0.80	0.82	0.81	0.82 (0.82–0.85)	0.82 (0.82–0.85)	0.89 (0.88–0.90)
SVM	0.81	0.81	0.79	0.80	0.79 (0.80–0.83)	0.83 (0.80–0.82)	0.88 (0.87–0.90)
XGBoost	0.82	0.81	0.80	0.81	0.80 (0.81–0.84)	0.84 (0.81–0.83)	0.88 (0.87–0.90)
Stack	0.82	0.80	0.83	0.82	0.83 (0.82–0.85)	0.82 (0.82–0.85)	0.89 (0.88–0.91)
98-item subset (Reference group)							
RF	0.82	0.81	0.82	0.82	0.82 (0.82–0.85)	0.83 (0.82–0.86)	0.89 (0.88–0.91)
C5.0	0.83	0.82	0.83	0.82	0.83 (0.83–0.86)	0.84 (0.84–0.86)	0.90 (0.88–0.91)
SVM	0.82	0.82	0.80	0.81	0.80 (0.80–0.83)	0.84 (0.80–0.84)	0.89 (0.88–0.90)
XGBoost	0.82	0.81	0.81	0.81	0.81 (0.82–0.85)	0.83 (0.81–0.84)	0.89 (0.88–0.90)
Stack	0.83	0.82	0.83	0.82	0.83 (0.84–0.86)	0.84 (0.84–0.86)	0.90 (0.89–0.91)

RF Random Forest, *C5.0* C5.0 Decision Tree, *SVM* Support Vector Machine, *XGBoost* Extreme Gradient Boosting, *Stack* Stacked Ensemble Model (meta-learner: logistic regression), *Accu* Accuracy, *Pre* Precision, *Rec* Recall, *F1* F1-score, *Sens* Sensitivity, *Spec* Specificity, *AUC* Area Under the ROC Curve. Values in parentheses represent 95% bootstrap confidence intervals (BCIs) based on 2000 replicates on the test set. As the stacked models demonstrated comparable performance, only the results from the default stacked model are reported for clarity

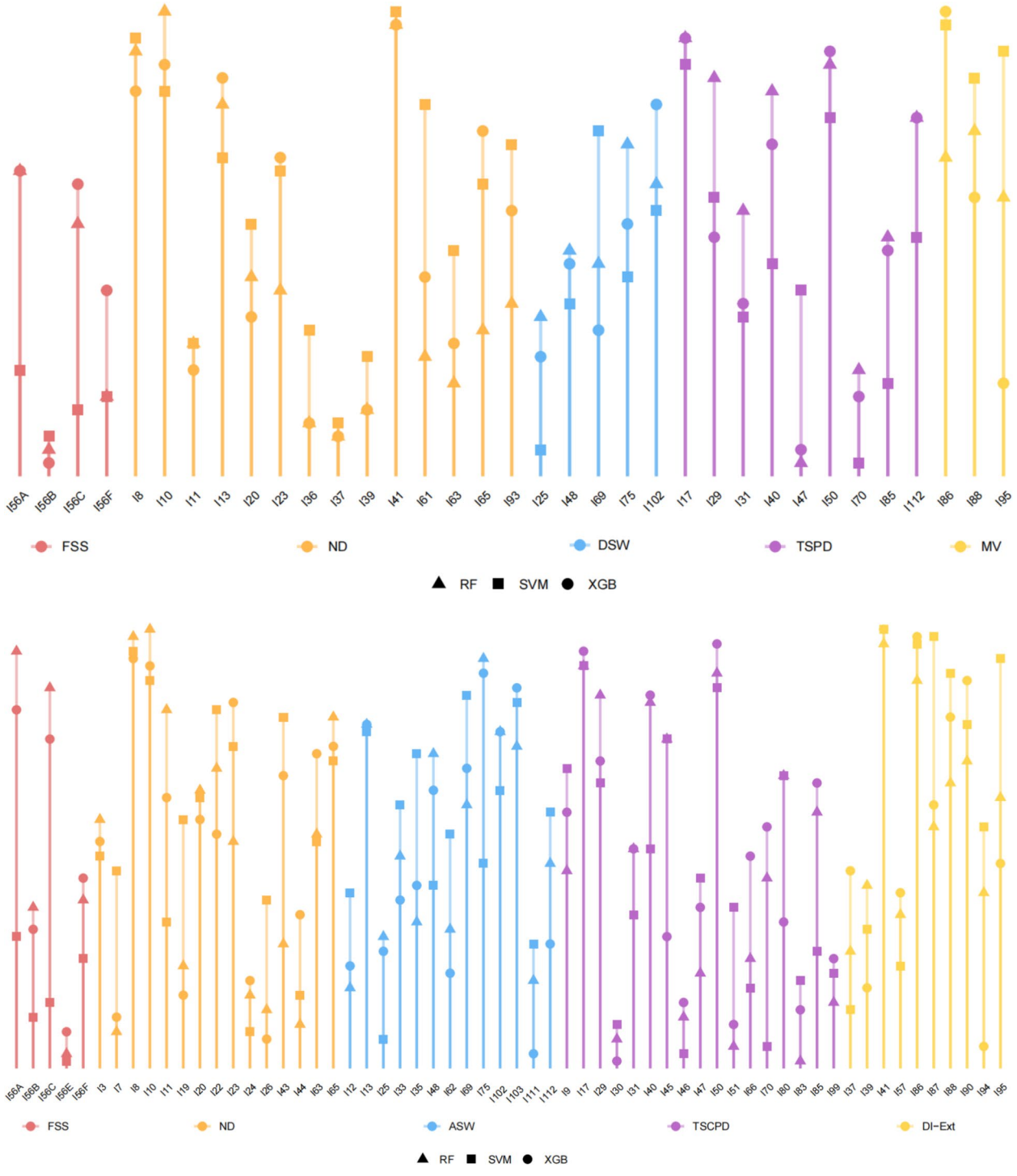

Fig. 3 (See legend on next page.)

all metrics. As expected, the 98-item reference model offered only marginal improvements over the shorter subsets. The stacked model showed the highest AUC (0.90), recall (0.83), and specificity (0.84). However, the overall differences between subsets were minor, suggesting that reduced item sets—such as the 35- or 60-item configurations—may retain predictive accuracy comparable to that of the full RFE-optimised model.

Item contribution

The ranked importance of items' contributions to the models is visualised in Fig. 3. In the 60-item subset, the

(See figure on previous page.)

Fig. 3 Item contributions to classification by model and dimension. Item contributions to classification by model and dimension. Each item was ranked within each model based on its raw importance scores (e.g., from random forest, SVM, or XGBoost), with ranking scores ranging from 1–35 or 1–60 depending on the item set size. Higher values indicate greater importance. Triangle, square, and circle markers denote the importance ranks assigned by RF, SVM, and XGBoost models, respectively. Line colours represent the item's associated dimension. *FSS* Functional Somatic Symptoms, *ND* Neurodevelopmental Dysregulation, *DSW* Detachment/Social Withdrawal, *TSPD* Threat Sensitivity and Perceptual Dysregulation, *MV* Mood Volatility, *ASW* Affective/Social Withdrawal, *TSCPD* Threat Sensitivity and Cognitive–Perceptual Disturbance, *DI-Ext* Disinhibited–Irritable Externalising. Item codes and descriptions ordered by dimension: I56A (aches or pains), I56B (headaches), I56C (nausea, feels sick), I56E (rashes or other skin problems), I56F (stomach aches); I3 (argues a lot), I7 (bragging, boasting), I8 (cannot concentrate, cannot pay attention for long), I10 (cannot sit still, restless, or hyperactive), I11 (clings to adults or is too dependent), I19 (demands a lot of attention), I20 (destroys own things), I22 (disobedient at home), I23 (disobedient at school), I24 (does not eat well), I26 (does not seem to feel guilty after misbehaving), I36 (gets hurt a lot, accident-prone), I43 (lying or cheating), I44 (bites fingernails), I63 (prefers being with older children), I65 (refuses to talk); I12 (complains of loneliness), I13 (confused or seems to be in a fog), I25 (does not get along with other children), I33 (feels or complains that no one loves them), I35 (feels worthless or inferior), I48 (not liked by other children), I61 (poor school work), I62 (poorly coordinated or clumsy), I69 (secretive, keeps things to self), I75 (too shy or timid), I102 (underactive, slow moving, or lacks energy), I103 (unhappy, sad, or depressed), I111 (withdrawn, does not get involved with others), I112 (worries); I9 (cannot get mind off certain thoughts), I17 (daydreams or gets lost in thoughts), I29 (fears certain animals, situations, or places other than school), I30 (fears going to school), I31 (fears they might think or do something wrong), I40 (hears sounds or voices that are not there), I45 (nervous, high-strung, or tense), I46 (nervous movements or twitching), I47 (nightmares), I50 (too fearful or anxious), I51 (feels dizzy or lightheaded), I66 (repeats certain acts over and over; compulsions), I70 (sees things that are not there), I80 (stares blankly), I83 (stores up too many things they do not need), I85 (strange ideas), I99 (smokes, chews, sniffs tobacco, or uses e-cigarettes); I37 (gets in many fights), I39 (hangs around with others who get in trouble), I41 (impulsive or acts without thinking), I57 (physically attacks people), I86 (stubborn, sullen, or irritable), I87 (sudden changes in mood or feelings), I88 (sulks a lot), I90 (swearing or obscene language), I93 (talks too much), I94 (teases a lot), I95 (temper tantrums or hot temper)

RF model identified I10 (hyperactivity) as the most influential item, whereas both the XGBoost and SVM models ranked I41 (impulsivity) as the top predictor. This pattern was partially replicated in the 35-item subset, although the XGBoost model instead identified I86 (irritability) as the most important item.

At a broader level, items from the Neurodevelopmental Dysregulation (ND) and Disinhibited–Irritable Externalising (DI-Ext) dimensions were consistently ranked higher than items from other domains. The DI-Ext dimension had the highest median importance rank in two of the three models, while the ND dimension ranked highest in the XGBoost model. In contrast, dimensions such as Functional Somatic Symptoms (FSS) and Threat Sensitivity and Cognitive–Perceptive Disturbance (TSCPD) exhibited lower median rankings, suggesting a more modest contribution to the classification models. In the 35-item subset, the MV (Mood volatility) dimension (a smaller version of DI-Ext) demonstrated notably high influence in both the SVM and XGBoost models, with a narrow range of importance ranks, indicating internal consistency. Conversely, the RF model assigned the greatest importance to items in the TSCPD dimension, despite the associated rank range was notably wide, reflecting higher variability in item-level influence, see Table 4 for details.

Overall, the influence patterns were similar across models, with the highest correlations observed between the RF and XGBoost models (Spearman's ρ range = 0.83–0.86, $p < 0.001$), and moderate correlations between the SVM and the tree-based models (Spearman's ρ range = 0.61–0.64, $p < 0.001$). Similarly, moderate to high overlap in the top 15 most influential items was observed between the two tree-based models (Jaccard similarity range = 0.58–0.67), while low to moderate overlap was found between the tree-based models and the SVM (Jaccard similarity range = 0.30–0.50). The pattern of item importance across item subsets was also highly robust: among the 32 items common to both the 60-item and 35-item models, rank-order consistency remained high (Spearman's ρ range = 0.80–1.00, $p < 0.001$), and top-item overlap ranged from strong to perfect (Jaccard similarity = 0.67–1.00).

Cutoff scores

For the 35-item subset, the optimal cut-off score was identified as 16 (AUC = 0.88, 95% bias-corrected confidence interval [BCI]: [0.87, 0.88]; accuracy = 0.80 [0.80, 0.81]; sensitivity = 0.80 [0.76, 0.84]; specificity = 0.81 [0.77, 0.85]). This suggests that individuals scoring 16 or above on this measure may warrant further clinical evaluation. Due to the scalar non-invariance observed across age groups, cut-off scores were calculated separately for preadolescence and adolescence. These were 15 for preadolescence (AUC = 0.87 [0.86, 0.88]; accuracy = 0.80 [0.79, 0.81]; sensitivity = 0.82 [0.76, 0.86]; specificity = 0.78 [0.74, 0.83]) and 17 for adolescence (AUC = 0.88 [0.87, 0.89]; accuracy = 0.81 [0.80, 0.82]; sensitivity = 0.79 [0.75, 0.82]; specificity = 0.83 [0.79, 0.86]). For the 60-item subset, the optimal cut-off score was 24 (AUC = 0.87, 95% BCI = [0.86, 0.88]; accuracy = 0.80 [0.80, 0.81]; sensitivity = 0.82 [0.75, 0.83]; specificity = 0.78 [0.79, 0.81]). While both subsets demonstrated comparable overall performance, the 35-item subset exhibited slightly higher specificity and AUC, suggesting a marginally superior capacity for accurate classification.

Table 4 Median and range of ranked importance scores by dimension, model, and item subset

Dimensions	RF	SMV	XGBoost
35-item subset			
FSS (4)	12.5 (2–23)	5.5 (3–8)	18 (1–23)
ND (14)	12 (3–35)	22.5 (4–35)	17.5 (3–34)
DSW (5)	17 (12–25)	15 (2–26)	16 (9–28)
TSPD (9)	27 (1–33)	16 (1–31)	18 (2–33)
MV (3)	24 (21–26)	32 (30–34)	21 (7–35)
60-item subset			
FSS (5)	23 (2–57)	9 (1–18)	26 (5–49)
ND (15)	32 (5–60)	34 (5–57)	34 (4–56)
ASW (13)	29 (11–56)	35 (4–51)	25 (2–54)
TSCPD (17)	27 (1–55)	22 (2–55)	29 (1–58)
DI-Ext (10)	35 (16–58)	50.5 (8–60)	32 (3–60)

FSS Functional Somatic Symptoms, *ND* Neurodevelopmental Dysregulation, *DSW* Detachment/Social Withdrawal, *TSPD* Threat Sensitivity and Perceptual Dysregulation, *MV* Mood Volatility, *ASW* Affective/Social Withdrawal, *TSCPD* Threat Sensitivity and Cognitive–Perceptual Disturbance, *DI-Ext* Disinhibited–Irritable Externalising. Values in brackets of the first column reflect the number of items within each dimension. Values in the columns of RF, SMV, XGBoost represent the median (minimum–maximum) of ranked importance scores for items in each dimension and model. Higher ranks indicate greater importance

Discussion

In this study, we developed 35- and 60-item short forms of the Child Behavior Checklist (CBCL) for Chinese youth, using machine learning and psychometric validation. Both subsets demonstrated high classification performance (AUC ≈ 0.88–0.90) in distinguishing psychiatric cases from healthy controls, closely approximating the full CBCL. At optimal cut-offs, the brief forms achieved balanced sensitivity and specificity (~0.80), effectively identifying most true cases while limiting false positives. These findings support that a reduced item set can retain diagnostic accuracy and offer a practical solution for early mental health screening.

Beyond overall screening performance, the shortened CBCL demonstrated a robust and interpretable internal structure. Exploratory graph analysis consistently identified five factors in both the 35- and 60-item sets, supported by confirmatory and bifactor models. For the 35-item version, factors included Functional Somatic Symptoms (FSS), Neurodevelopmental Dysregulation (ND), Detachment and Social Withdrawal (DSW), Threat Sensitivity and Perceptual Disturbance (TSPD), and Mood Volatility (MV). The 60-item version yielded similar factors, with refinements: Affective Social Withdrawal (ASW), Threat Sensitivity and Cognitive-Perceptual Disturbance (TSCPD), and Disinhibited–Irritable Externalising (DI-Ext). In both versions, all five dimensions loaded strongly onto a general psychopathology factor ('p-factor'), suggesting they reflect a common severity dimension. ND and DI-Ext captured the largest share of general variance, while somatic symptoms were more distinct. This aligns with our machine learning findings that neurodevelopmental dysregulation and disinhibited-irritable externalising behaviours carried the highest feature importance for predicting common psychiatric morbidity in Chinese children. A possible explanation is that the CBCL relies on parent and caregiver reports. Because these problems are highly observable, they often serve as strong proxy indicators for clinical attention [32]. Consequently, these externalising features effectively flag youths requiring comprehensive clinical evaluation. Overall, both short forms retained strong reliability and validity comparable to the full CBCL while being substantially more efficient.

The emergence of a stable five-factor structure in this Chinese youth sample is notable, diverging from the CBCL's traditional eight-syndrome structure to group symptoms into broader, clinically meaningful domains [33]. It reflects a mid-level hierarchy of psychopathology—more detailed than the internalising vs. externalising distinction, yet more parsimonious than numerous narrow syndromes [34]. This finding aligns with modern dimensional models that highlight a few core domains (e.g., neurodevelopmental dysregulation, negative affectivity, disruptive behaviour) in youth [35]. However, applying this five-factor structure across different scale lengths revealed important psychometric and developmental implications. While the 60-item solution achieved full scalar invariance across age cohorts, the 35-item subset exhibited scalar non-invariance. Statistically, this reflects the "buffering effect" of larger models, as simulations indicate that categorical fit indices are highly sensitive to scale length [36]. Beyond statistical explanations, this non-invariance, as observed in the 35-item subset optimised for clinical significance, likely captures meaningful developmental shifts during the transition to adolescence. This aligns with the concept of heterotypic continuity, wherein the manifestation of underlying psychopathology shifts across developmental stages [37]. Therefore, future longitudinal research should elucidate how these age-related measurement differences reflect true psychopathological dynamics across developmental trajectories from childhood to adolescence.

Each identified domain reflects a recognised pattern of child psychopathology. In both the 35- and 60-item subsets, the Functional Somatic Symptoms (FSS) dimension clustered consistently, particularly around item I56 (aches, pains, nausea, etc.), underscoring the relative independence of somatic complaints in youth [38]. This finding aligns with prior research showing that somatic symptoms often form a distinct cluster, separate from emotional or behavioural problems [39]. In our study, the FSS factor had the weakest association with the general psychopathology factor, suggesting these symptoms may occur independently, potentially reflecting cultural preferences for somatic expression of psychological distress

[40]. Although its diagnostic specificity was lower than domains like ND or DI-Ext, the robustness of the FSS factor highlights its clinical relevance [41].

Similarly, both item subsets consistently yielded a Neurodevelopmental Dysregulation (ND) dimension. In the 35-item version, ND primarily encompassed attention deficits, hyperactivity, and impulsivity—core features of ADHD and related conditions—highlighting neurodevelopmental dysregulation as a key contributor to child psychopathology [42]. Highly predictive items such as restlessness and poor concentration underscored the centrality of attentional and regulatory difficulties as early risk indicators [43]. In the 60-item version, ND extended to include externalising behaviours (e.g., arguing, bragging, disobedience, lack of guilt, lying), capturing a broader behavioural profile often comorbid with neurodevelopmental traits [44].

The Detachment and Social Withdrawal (DSW) dimension in the 35-item set and the Affective Social Withdrawal (ASW) dimension in the 60-item set showed strong alignment, both including key indicators such as interpersonal difficulties (I25) and peer rejection (I48), underscoring their central relevance to internalising disorders across different scale lengths. The 60-item version expanded this domain by adding affective items—loneliness (I12), confusion (I13), worthlessness (I35), and depressive symptoms (I103)—broadening its sensitivity to a wider spectrum of emotional distress, including depression and anxiety [45].

Similarly, the Threat Sensitivity and Perceptual Disturbance (TSPD; 35-item) and Threat Sensitivity and Cognitive-Perceptual Disturbance (TSCPD; 60-item) dimensions consistently captured anxiety-related cognitive and perceptual symptoms, including daydreaming (I17), specific fears (I29), and auditory hallucinations (I40). The consistent clustering of these items underscores their relevance as stable features within child psychopathology. The 60-item version further incorporated obsessive–compulsive traits, dissociation, and somatic anxiety (I46, I51, I66, I80, I83, I99), broadening the spectrum to include more severe internalising or trauma-related presentations [46].

Finally, the Mood Volatility (MV; 35-item) and Disinhibited–Irritable Externalising (DI-Ext; 60-item) dimensions both highlighted core features of irritability and mood instability, including stubbornness (I86), sulking (I88), and temper tantrums (I95). The 60-item version expanded this domain to cover more explicit externalising behaviours, such as fighting (I37), peer conflict (I39), impulsivity (I41), physical aggression (I57), and excessive talking (I93), offering a broader assessment of behavioural dysregulation. The centrality of irritability—marked by frequent outbursts—underscores its role as a transdiagnostic indicator of severity, consistently linked to poor mental health outcomes [47].

A broader insight from our results is the prominence of a general psychopathology factor spanning all domains. Bifactor analysis showed that a single "p-factor" accounted for much of the common variance among the 35 items and 60 items, echoing evidence for a general dimension of psychopathology in youth [22]. Items from ND and DI-Ext domains loaded most strongly on this factor, suggesting attentional and behavioural dysregulation often co-occur with diverse symptoms [48]. In contrast, somatic complaints showed more specific variance, indicating relative independence from the general factor [49]. These findings support hierarchical models comprising a general and multiple specific factors and endorse the use of a total problem score as a broad screening tool, with domain scores offering further clinical precision.

The development of these short-form CBCL-based screeners holds several implications for research and practice. The 35-item subset markedly reduces respondent burden, making it ideal for large-scale use in schools, community health settings, and paediatrics without sacrificing diagnostic accuracy [50]. The 60-item version, while still concise, provides richer clinical detail, suitable for settings requiring greater diagnostic precision. Both scales build on the CBCL's widespread use, enabling smooth integration into existing protocols. Empirically derived cut-off scores (e.g., ~ 16 for 35 items; ~ 24 for 60 items) offer clear decision thresholds that balance sensitivity and specificity. In low-resource settings, this enables efficient triage and prioritisation of children most likely in need of further care [51]. By focusing on the most predictive items, the short forms are sensitive to a wide range of psychopathology and comorbidity profiles; a positive screen indicates elevated risk, warranting further evaluation, even when specific diagnoses are unclear [52]. This transdiagnostic approach ensures children with complex or atypical symptoms are not overlooked, supporting comprehensive and equitable identification practices [53].

Limitations

A primary limitation of this study is its constrained generalisability. The machine learning models were trained on a subsample of diagnosed youths and matched controls from two southern Chinese regions, which may not reflect the broader population. Furthermore, reliance on parent-reported CBCL data introduces subjectivity and potential bias, as caregiver perceptions may not fully capture the child's mental health state. In addition, although using a binary "any diagnosis" outcome is appropriate for validating a preliminary screening tool, it limits the specificity of our predictions. Moreover, the absence of external validation dataset restricts our ability to assess

model performance across different settings or populations. Consequently, these findings should be interpreted with caution.

Future research should test the tool across diverse geographic regions and informant types, examine its longitudinal predictive value, and explore digital or adaptive formats to improve accessibility and scalability in child mental health care. Finally, future studies should utilise multi-class classification frameworks and larger, disorder-specific clinical subsamples to thoroughly validate the derived dimensions. For example, future research should investigate whether dimensions such as neurodevelopmental dysregulation possess the discriminant validity to accurately distinguish ADHD from other psychiatric conditions.

Conclusion

We developed a 35-item CBCL short form that achieves efficient yet accurate screening for child psychopathology. Leveraging machine learning to distill the most predictive items and confirming a sound factor structure, we created a brief assessment tool that performs comparably to the full CBCL in identifying at-risk youth. This work demonstrates the promise of integrating data-driven methods with established measures to improve early detection in mental healthcare. Our findings support the feasibility of streamlined evidence-based screening approaches that could broaden the reach of child mental health services.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13034-026-01087-4.

Supplementary material 1.

Acknowledgements
The authors are grateful to all participants for their support.

Author contributions
All authors were involved in data collection. J.L. and W.S. performed the data analysis and wrote the manuscript. O.Z. made revisions to the manuscript. YZ.W, HH.H, GY.X, and YJ.Q were involved throughout the data collection process. F.H. made practical contributions to the conduct and advancement of the trial. Y.Z. led the entire team and directed and supervised the whole trial. The authors have declared no competing or potential conflicts of interest.

Funding
This study was funded by Research on prevention and control of major chronic non- communicable diseases in the Ministry of Science and Technology (No: 2016YFC1306100); Capital's Funds for Health Improvement and Research, code: CFH2024-1-2123;Beijing Hospitals Authority Clinical Medicine Development of Special Funding Support, code: YGLX202538 ; High-level Public Health Technology Talent Construction Project, No: Discipline Backbone-01-029; High-level Public Health Technology Talent Construction Project, No: Academic leader-02-01.

Data availability
The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Declarations

Ethics approval and consent to participate
The project was approved by the Ethics Committee of Beijing Anding Hospital (201743FS-2) in 2017 and performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. All subjects and their parents signed an informed consent form before joining the trial.

Consent for publication
Participants gave consent for data collected to be published in anonymous form in academic journals.

Competing interests
The authors declare no competing interests.

Received: 4 September 2025 / Accepted: 14 April 2026
Published online: 06 May 2026

References

1. Kieling C, Buchweitz C, Caye A, Silvani J, Ameis SH, Brunoni AR, et al. Worldwide prevalence and disability from mental disorders across childhood and adolescence: evidence from the global burden of disease study. JAMA Psychiat. 2024;81(4):347–56. https://doi.org/10.1001/jamapsychiatry.2023.5051.
2. Li F, Cui Y, Li Y, Guo L, Ke X, Liu J, et al. Prevalence of mental disorders in school children and adolescents in China: diagnostic data from detailed clinical assessments of 17,524 individuals. J Child Psychol Psychiatry. 2022;63(1):34–46. https://doi.org/10.1111/jcpp.13445.
3. Wei T, Guangcan Y, Shangzhi X, Jing Z, Junyi P, Xinyi Z, et al. Burden of depressive and anxiety disorders in China and its provinces, 1990–2021: findings from the Global Burden of Disease Study 2021. Br J Psychiatry. 2025. https://doi.org/10.1192/bjp.2024.267.
4. Uhlhaas PJ, Davey CG, Mehta UM, Shah J, Torous J, Allen NB, et al. Towards a youth mental health paradigm: a perspective and roadmap. Mol Psychiatry. 2023;28(8):3171–81. https://doi.org/10.1038/s41380-023-02202-z.
5. Jiang Z, Abbey C, Chen J, Yang Z, Xu H, Zhang A, et al. The shortage of child psychiatrists in mainland China. Nat Ment Health. 2024;2(7):741–4. https://doi.org/10.1038/s44220-024-00273-7.
6. Adrian A, Alaattin E, William C, Robert G, Prudence W F, E Jane C. Psychiatric diagnostic interviews for children and adolescents: a comparative study. J Am Acad Child Adolesc Psychiatry. 2012. https://doi.org/10.1016/j.jaac.2012.02.020.
7. King RA. Practice parameters for the psychiatric assessment of children and adolescents. J Am Acad Child Adolesc Psychiatry. 1997;36:4S-20S. https://doi.org/10.1097/00004583-199710001-00002.
8. Achenbach TM. Child behavior checklist. In: Kreutzer JS, DeLuca J, Caplan B, editors. Encyclopedia of clinical neuropsychology. New York: Springer; 2011. p. 546–52.
9. Jiang Z, Wang D, Xu H, Zhang A, Zhao Q, Yan J, et al. Diagnostic efficiency and psychometric properties of CBCL DSM-oriented scales in a large sample of Chinese school-attending students aged 5–16. Asian J Psychiatry. 2023;88:103724. https://doi.org/10.1016/j.ajp.2023.103724.
10. Nakamura B, Ebesutani C, Bernstein A, Chorpita B. A psychometric analysis of the Child Behavior Checklist DSM-oriented scales. J Psychopathol Behav Assess. 2009;31:178–89. https://doi.org/10.1007/s10862-008-9119-8.
11. Sharma H. How short or long should be a questionnaire for any research? Researchers dilemma in deciding the appropriate questionnaire length. Saudi J Anaesth. 2022;16(1):65.
12. Xu W, Wang M, Zhang X, Liu Y. The Brief Problem Monitor (BPM-Y/BPM-P) among Chinese youth: psychometric properties and measurement invariance. J Psychopathol Behav Assess. 2022;44(2):496–510. https://doi.org/10.1007/s10862-021-09927-7.
13. Zhao C, Huang J. The impact of the scoring system of a large-scale standardized EFL writing assessment on its score variability and reliability: implications

for assessment policy makers. Stud Educ Eval. 2020;67:100911. https://doi.org/10.1016/j.stueduc.2020.100911.
14. Piper BJ, Gray HM, Raber J, Birkett MA. Reliability and validity of Brief Problem Monitor, an abbreviated form of the Child Behavior Checklist. Psychiatry Clin Neurosci. 2014;68(10):759–67. https://doi.org/10.1111/pcn.12188.
15. Achenbach T, Mcconaughy S, Ivanova M, Rescorla L. Manual for the ASEBA brief problem monitor™(BPM). Burlington: ASEBA; 2011.
16. Cappelleri JC, Lundy JJ, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther. 2014;36(5):648–62. https://doi.org/10.1016/j.clinthera.2014.04.006.
17. Kendra LR, Cara AS, Jeremy P, Philip CK, Scott C, John P, et al. Predicting anxiety diagnoses and severity with the CBCL-A: improvement relative to other CBCL scales? J Psychopathol Behav Assess. 2015. https://doi.org/10.1007/s10862-014-9439-9.
18. Somer DR, Boris B, David A, Ben G, MaryKay G, Michael S, et al. The Child Behavior Checklist (CBCL) and the CBCL-bipolar phenotype are not useful in diagnosing pediatric bipolar disorder. J Child Adolesc Psychopharmacol. 2009. https://doi.org/10.1089/cap.2008.067.
19. Reynolds CR, Altmann RA, Allen DN. Item analysis: methods for fitting the right items to the right test. In: Reynolds CR, Altmann RA, Allen DN, editors. Mastering modern psychological testing: theory and methods. Cham: Springer International Publishing; 2021. p. 263–89.
20. Dalgleish T, Black M, Johnston D, Bevan A. Transdiagnostic approaches to mental health problems: current status and future directions. J Consult Clin Psychol. 2020;88(3):179.
21. Allegrini AG, Cheesman R, Rimfeld K, Selzam S, Pingault JB, Eley TC, et al. The p factor: genetic analyses support a general dimension of psychopathology in childhood and adolescence. J Child Psychol Psychiatry. 2020;61(1):30–9.
22. Caspi A, Houts RM, Belsky DW, Goldman-Mellor SJ, Harrington H, Israel S, et al. The p factor: one general psychopathology factor in the structure of psychiatric disorders? Clin Psychol Sci. 2014;2(2):119–37. https://doi.org/10.1177/2167702613497473.
23. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol. 2018;14:91–118.
24. Ostojic D, Lalousis PA, Donohoe G, Morris DW. The challenges of using machine learning models in psychiatric research and clinical practice. Eur Neuropsychopharmacol. 2024;88:53–65.
25. Leung PW, Kwong SL, Tang CP, Ho TP, Hung SF, Lee CC, et al. Test-retest reliability and criterion validity of the Chinese version of CBCL, TRF, and YSR. J Child Psychol Psychiatry. 2006;47(9):970–3. https://doi.org/10.1111/j.1469-7610.2005.01570.x.
26. Alexander P C, Luis Eduardo G, Hudson G. Unique variable analysis: a network psychometrics method to detect local dependence. Multivar Behav Res. 2023. https://doi.org/10.1080/00273171.2023.2194606.
27. Ferketich S. Focus on psychometrics. Aspects of item analysis. Res Nurs Health. 1991. https://doi.org/10.1002/nur.4770140211.
28. Thomas S. Robert F. DeVellis & Carolyn T. Thorpe. (2022) Scale development: theory and applications. California: SAGE Publications Inc, 320 pages, $60.00 softcover. Pers Psychol. 2022;75:243–4. https://doi.org/10.1111/peps.12499.
29. Golino HF, Epskamp S. Exploratory graph analysis: a new approach for estimating the number of dimensions in psychological research. PLoS ONE. 2017;12(6):e0174035.
30. Osborne J, Osborne JW, Costello AB, Kellow JT. Best practices in exploratory factor analysis. Thousand Oaks: SAGE Publications Inc; 2008.
31. Wu H, Estabrook R. Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. Psychometrika. 2016;81(4):1014–45.
32. De Los Reyes A, Kazdin AE. Informant discrepancies in the assessment of childhood psychopathology: a critical review, theoretical framework, and recommendations for further study. Psychol Bull. 2005;131(4):483.
33. Offermans JE, de Bruin EI, Lange AM, Middeldorp CM, Wesseldijk LW, Boomsma DI, et al. The development and validation of a subscale for the school-age child behavior checklist to screen for autism spectrum disorder. J Autism Dev Disord. 2022. https://doi.org/10.1007/s10803-022-05465-7.
34. Conway CC, Forbes MK, South SC. A hierarchical taxonomy of psychopathology (HiTOP) primer for mental health researchers. Clin Psychol Sci. 2022;10(2):236–58. https://doi.org/10.1177/21677026211017834.
35. McLaughlin KA, Sheridan MA, Humphreys KL, Belsky J, Ellis BJ. The value of dimensional models of early experience: thinking clearly about concepts and categories. Perspect Psychol Sci. 2021;16(6):1463–72. https://doi.org/10.1177/1745691621992346.
36. D'Urso ED, De Roover K, Vermunt JK, Tijmstra J. Scale length does matter: recommendations for measurement invariance testing with categorical factor analysis and item response theory approaches. Behav Res Methods. 2022;54(5):2114–45.
37. Petersen IT, Choe DE, LeBeau B. Studying a moving target in development: the challenge and opportunity of heterotypic continuity. Dev Rev. 2020;58:100935.
38. Gershfeld-Litvin A, Hertz-Palmor N, Shtilerman A, Rapaport S, Gothelf D, Weisman H. The development of somatic symptom disorder in children: psychological characteristics and psychiatric comorbidity. J Acad Consult Liaison Psychiatry. 2022;63(4):324–33. https://doi.org/10.1016/j.jaclp.2021.10.008.
39. Huang W, Chang S, Liao S. The psychological features of distinct somatic syndromes: a cluster analysis according to population-based somatic symptom profiles in Taiwan. J Formos Med Assoc. 2022;121(9):1813–22. https://doi.org/10.1016/j.jfma.2022.03.009.
40. Kaiser BN, Weaver JL. Culture-bound syndromes, idioms of distress, and cultural concepts of distress: new directions for an old concept in psychological anthropology. Transcult Psychiatry. 2019;56(4):589–98. https://doi.org/10.1177/1363461519862708.
41. Ibeziako P, Brahmbhatt K, Chapman A, De Souza C, Giles L, Gooden S, et al. Developing a clinical pathway for somatic symptom and related disorders in pediatric hospital settings. Hosp Pediatr. 2019;9(3):147–55. https://doi.org/10.1542/hpeds.2018-0205.
42. Whittle S, Zhang L, Rakesh D. Environmental and neurodevelopmental contributors to youth mental illness. Neuropsychopharmacology. 2025;50(1):201–10. https://doi.org/10.1038/s41386-024-01926-y.
43. Roberts C, Sahakian BJ, Chen S, Sallie SN, Walker C, White SR, et al. Impact and centrality of attention dysregulation on cognition, anxiety, and low mood in adolescents. Sci Rep. 2023;13(1):9106. https://doi.org/10.1038/s41598-023-34399-y.
44. Arslan İB, Lucassen N, van Lier PAC, de Haan AD, Prinzie P. Early childhood internalizing problems, externalizing problems and their co-occurrence and (mal)adaptive functioning in emerging adulthood: a 16-year follow-up study. Soc Psychiatry Psychiatr Epidemiol. 2021;56(2):193–206. https://doi.org/10.1007/s00127-020-01959-w.
45. Shapero BG, Abramson LY, Alloy LB. Emotional reactivity and internalizing symptoms: moderating role of emotion regulation. Cognit Ther Res. 2016;40(3):328–40. https://doi.org/10.1007/s10608-015-9722-4.
46. McHugh L, Egan J. Psychological and somatic manifestations of dissociation: the role of childhood trauma, attachment, and alexithymia. Eur J Trauma Dissociation. 2023;7(1):100316. https://doi.org/10.1016/j.ejtd.2023.100316.
47. Chin M, Robson DA, Woodbridge H, Hawes DJ. Irritability as a transdiagnostic construct across childhood and adolescence: a systematic review and meta-analysis. Clin Child Fam Psychol Rev. 2025;28(1):101–24. https://doi.org/10.1007/s10567-024-00512-4.
48. Forslund T, Brocki KC, Bohlin G, Granqvist P, Eninger L. The heterogeneity of attention-deficit/hyperactivity disorder symptoms and conduct problems: cognitive inhibition, emotion regulation, emotionality, and disorganized attachment. Br J Dev Psychol. 2016;34(3):371–87. https://doi.org/10.1111/bjdp.12136.
49. Löwe B, Toussaint A, Rosmalen JGM, Huang W, Burton C, Weigel A, et al. Persistent physical symptoms: definition, genesis, and management. Lancet. 2024;403(10444):2649–62. https://doi.org/10.1016/S0140-6736(24)00623-8.
50. Williams JC. Editorial: making the case of universal mental health screening in schools. J Am Acad Child Adolesc Psychiatry. 2025. https://doi.org/10.1016/j.jaac.2025.01.035.
51. Watson T, Tindall R, Patrick A, Moylan S. Mental health triage tools: a narrative review. Int J Ment Health Nurs. 2023;32(2):352–64. https://doi.org/10.1111/inm.13073.
52. DeSantis B, Jackson MJ, Duncan BL, Reese RJ. Casting a wider net in behavioral health screening in primary care: a preliminary study of the Outcome Rating Scale. Prim Health Care Res Dev. 2017;18(2):188–93. https://doi.org/10.1017/S1463423616000311.
53. Merten EC, Cwik JC, Margraf J, Schneider S. Overdiagnosis of mental disorders in children and adolescents (in developed countries). Child Adolesc Psychiatry Ment Health. 2017;11(1):5. https://doi.org/10.1186/s13034-016-0140-5.

Publisher's Note