NEW RESEARCH DIMENSIONAL APPROACHES

Check for updates

Studying Development of Psychopathology Using Changing Measures to Account for Heterotypic Continuity

Isaac T. Petersen[a,*], PhD, Zachary Demko[a], MA, Won-Chan Lee[a], PhD, Jacob J. Oleson[a], PhD

Objective: Psychopathology shows changes in behavioral manifestation across development, that is, heterotypic continuity. However, research has paid little attention to how to account for heterotypic continuity when examining the development of psychopathology. This longitudinal study accounted for heterotypic continuity of multiple psychopathology dimensions by using developmental scaling to place multi-informant ratings of children's behavior problems onto the same scale to chart children's trajectories.

Method: The study examined children's (N = 231) development of 3 psychopathology dimensions—externalizing, internalizing, and thought-disordered—using different measures across 7 timepoints from 3 to 7.5 years of age. Psychopathology dimensions were assessed by mother-, father-, and teacher/caregiver-report. We compared 3 assessment approaches: the common items, upward/downward extension, and construct-valid items approaches. We compared 2 scoring approaches: mean scoring and developmental scaling. Developmental scaling aims to place scores from age-differing measures onto the same scale. We compared their accuracy, for externalizing problems, in terms of criterion validity with respect to observations of compliance and attention to task.

Results: Using different measures across ages (ie, construct-valid items approach) was the most accurate assessment approach—modestly more accurate than using the common items or upward/downward extension—in terms of criterion validity with respect to observations of compliance and attention to task (r_{diff} = 0.07-0.13). Developmental scaling was the most accurate scoring approach, modestly more accurate than average scores (r_{diff} = 0.03-0.17).

Conclusion: Using (1) age-differing measures to account for heterotypic continuity and (2) developmental scaling to link scores from the different measures onto the same scale may enable studying development of psychopathology across the lifespan.

Plain language summary: Behavioral health problems manifest differently as children age. This study examined manifestations of 3 dimensions of behavioral health problems from ages 3 to 7.5 years (N = 231). The authors found that using different measures and scaling across ages was the most accurate way to assess behavioral health problems as children age.

Diversity & Inclusion Statement: We worked to ensure sex and gender balance in the recruitment of human participants. We worked to ensure race, ethnic, and/or other types of diversity in the recruitment of human participants. We worked to ensure that the study questionnaires were prepared in an inclusive way.

Study registration information: School readiness study: https://osf.io/jzxb8

Key words: development of psychopathology; heterotypic continuity; changing measures; developmental scaling; longitudinal

JAACAP Open 2026;4(1):111-123.

Psychopathology is thought to show heterotypic continuity—its behavioral manifestations change across development despite persistence in the underlying construct.[1] For instance, externalizing problems—which encompass disinhibition and antagonism—are often expressed as overt acts in early childhood, such as physical aggression; however, externalizing problems are more often expressed later in development as covert and indirect or relational forms of aggression, rule breaking, and substance use.[2] These behaviors are often considered covert or indirect because, unlike overt physical aggression, they are less visible, less confrontational, or occur outside the immediate view of authority figures or of individuals targeted by the aggressive behavior. A consequence of the changing behavioral manifestation of psychopathology is that different measures across development—and methods to link their scores onto the same scale—may be necessary to chart children's development of psychopathology while accounting for heterotypic continuity. Despite considerable evidence[3] and theory[4] demonstrating that psychopathology shows changes in manifestation across development, heterotypic continuity has largely been ignored when studying people's trajectories. Nearly all longitudinal studies that chart people's

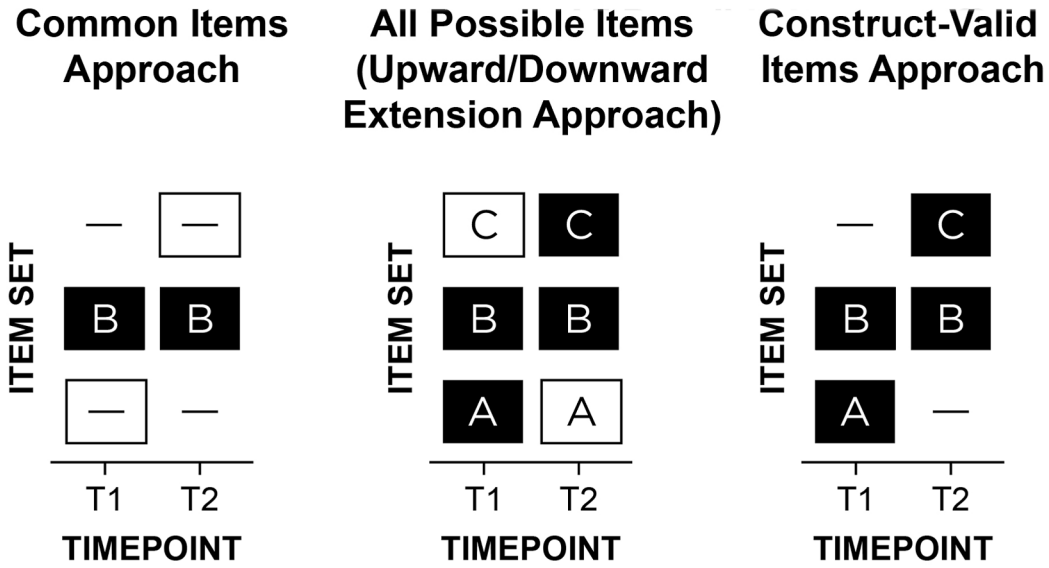
development of psychopathology use the same measure across time,[1] likely leading to inaccurate trajectories.[5–8] Moreover, emerging conceptual frameworks of psychopathology, including the Hierarchical Taxonomy of Psychopathology (HiTOP) and Research Domain Criteria (RDoC), do not account for development.[9–11] Thus, developing approaches to account for heterotypic continuity represents a critical challenge for advancing understanding of how psychopathology develops.

It is important to examine continuity and discontinuity across key developmental transitions, such as the transition from preschool to school entry. Entry into formal schooling often introduces structured demands (eg, sitting still and completing tasks) and additional socialization agents, including teachers and a transition from a reliance on caregivers to closer affiliation with peers. Although heterotypic continuity has since been demonstrated across the lifespan,[3] Kagan argued that heterotypic continuity is particularly likely before age 10—during which time children learn more effective ways of accomplishing their goals[12] and develop skills in inhibiting inappropriate actions[13]—and after major changes in their psychological ecology, including transition to school.[13] Some behaviors likely emerge across this developmental period, such as lying, supported by developing skills in theory of mind and executive functioning,[14] whereas other behaviors may wane with age, such as biting others and having temper tantrums. Determining how to chart individuals' change across the transition from preschool to school entry is a necessary step toward the goal of tracking people's change across the lifespan. The present study examines children's development of multiple dimensions of psychopathology across 3 to 7.5 years of age during this crucial transition from preschool to school entry, and examines the accuracy of various approaches to longitudinal assessment of externalizing behavior.

There are 2 primary ways that prior work has examined people's change in psychopathology across development (Figure 1).[6] The "common items" approach removes items that are age specific. For example, in a study of externalizing problems from early to middle childhood, a study might remove some physical aggression items such as "bites others" that might not be developmentally relevant at all ages. The second approach, the "upward/downward extension" approach, takes items that are valid at a given age and uses the items across ages when the item may not be valid or useful; that is, all items are assessed at all ages. For instance, the upward extension approach might take the item "disobedient to authority figures" and apply it to older individuals for whom such a behavior may no longer validly reflect externalizing behavior and may instead reflect prosocial functions including protesting against societally unjust actions.[6] Both the "common items"[15,16] and "upward/downward extension"[17–19] approaches are widely used, likely because they result in the same items assessed across ages; however, both have key problems. As an example of the common items approach, Odgers *et al.*[15] examined children's development of *DSM-IV* symptoms of conduct disorder from ages 7 to 26 years, but they dropped age-specific symptoms (eg, running away, staying out late) "because [these symptoms] did not cover the study's age span."[p676] As an example of the upward extension approach, Broeren *et al.*[18] examined children's development of anxiety from ages 4 to 11 years; they noted, "Although the questionnaire was originally developed to measure anxiety in preschoolers, the current study also employed the scale with older children to promote uniformity in measures."[p84] Forcing use of the same measure across time naturally limits the ages that a study can span, which prevents charting wider age spans. For example, one could not examine development of depression across ages 11 to 13 years using the Short Mood and Feelings Questionnaire because of changes in its functioning (in particular, some items were less relevant at age 11) that likely reflect developmental changes in manifestation of depression.[20] The common items approach yields low content validity because it does not assess all construct facets, especially age-specific manifestations. The upward/downward extension approach violates construct validity because it assesses developmentally inappropriate items. In sum, despite these approaches being the most widely used for assessing individuals' development, they likely lead to inaccurate scores and thus inaccurate trajectories.

The present study considers a third approach to assess individuals' development of psychopathology: the "construct-valid items" approach (Figure 1). The construct-valid items approach uses different items, as needed, across time to maintain construct validity and to account for the changing behavioral manifestation of psychopathology. Because the construct-valid items approach uses different measures across development when the construct changes in manifestation, scoring approaches are needed to link scores from the age-differing measures onto the same scale in ways that allow tracking individuals' absolute growth. Traditional scoring approaches (eg, mean, sum) do not ensure that scores from the differing measures are comparable, and age-normed scores do not allow observing absolute growth. A potentially useful scoring approach for linking scores from age-differing measures onto the same scale is developmental scaling.[6] Developmental scaling (aka vertical scaling) refers to the process of linking or harmonizing scores from different measures across development onto the same scale while retaining absolute change (unlike age-normed scores).[21] There are various approaches to developmental scaling such as item response theory–based approaches. The present study performs

FIGURE 1 Approaches to Longitudinal Assessment



Note: *Item set A refers to items that are construct-valid at only timepoint 1 (T1); item set B is construct-valid at both T1 and T2; item set C is construct-valid at only T2. A dash indicates that the item set was not assessed at a given timepoint. A white box indicates invalid assessment in terms of either a content gap (ie, important missing items) or intrusion (ie, invalid items at a given timepoint). In a study of externalizing problems from early childhood to adulthood, "biting others" may be in item set A; "noncompliant" in B; "drug use" in C. The 3 approaches are as follows: (1) common items: B at T1 and T2; (2) upward/downward extension: ABC at T1 and T2; or (3) construct-valid items: AB at T1; BC at T2. The "common items" and "upward/downward extension" approaches are by far the most widely used in the literature, even though they likely lead to inaccurate scores and thus inaccurate trajectories. In the present study, the change in measurement occurred between ages 5 and 6 years (ie, the Child Behavior Checklist 1.5–5 and Caregiver–Teacher Report Form were assessed from ages 3 to 5 1/4 years; the Child Behavior Checklist 6–18 and Teacher's Report Form were assessed from ages 6 to 7.5 years).*

developmental scaling as a function of age (ie, age-based scaling). Developmental scaling benefits from having some age-common items from measures at adjacent ages, to serve as an anchor. Developmental scaling can leverage the common items to link the scores from the different measures onto the same scale (ie, the scale of the reference age). By assuming no changes in functioning of the common items across ages (ie, no differential item functioning) or by accounting for any changes in item functioning, this approach ensures that both item and latent parameter estimates are expressed on a common scale across ages. Nevertheless, developmental scaling can use all construct-valid items—both common and unique items—to estimate people's scores on that scale, thus making use of all construct-valid information while estimating people's scores on a comparable scale across development.

Simulation work has demonstrated that, compared to the common items and upward/downward extension approaches, the construct-valid items approach yields more accurate trajectories at the group and person level.[7] Although many studies have examined development of psychopathology, few studies have examined trajectories in ways that account for heterotypic continuity, that is, by using different, age-appropriate measures across time to maintain construct validity,[22,23] which is crucial for accurate inferences. Even fewer have done so in ways that allow identifying absolute change, which is necessary to chart individual development. We are aware of only 3 studies that have used the construct-valid items approach (ie, age-differing measures) to study individuals' absolute change in psychopathology,[8,24,25] in the case of externalizing problems[24,25] and internalizing problems.[8,24] This is surprising, because the use of age-differing measures is common when studying individuals' development in other domains, including cognitive and educational development.[26–28] Moreover, no studies have performed developmental scaling of the 3 primary dimensions of psychopathology simultaneously—including the externalizing, internalizing, and thought-disordered dimensions of psychopathology.[29] Externalizing problems encompass disinhibition and antagonism; internalizing problems encompass mood and anxiety problems; thought-disordered problems encompass psychosis. Collectively, externalizing, internalizing, and thought-disordered problems are thought to parsimoniously describe many forms of psychopathology.[29] Moreover, the 3 dimensions covary, so there is value in modeling them together. In addition, despite theory and simulation work supporting the construct-valid items approach,[7] no studies have compared the common items, upward/downward extension, and construct-valid items approaches empirically.

The Present Study

The present study demonstrates and evaluates the use of the following: (1) different yet construct-valid items across time—ie, the "construct-valid items" approach—to account for the changing behavioral manifestation of psychopathology across development; and (b) developmental scaling methods to place the scores from the different items onto the same scale to allow charting of individuals' development. Consistent with modern conceptualizations of the (multi)dimensionality of psychopathology, we examine the development of externalizing, internalizing, and thought-disordered dimensions of psychopathology during childhood. The present study is the first to use different measures of externalizing, internalizing, and thought-disordered psychopathology over time (and we leverage overlapping items) while using developmental scaling to link scores from the measures onto the same scale for more accurate growth estimates that account for heterotypic continuity. This allows us to chart children's development in externalizing, internalizing, and thought-disordered psychopathology across a lengthy age span despite changes in behavioral manifestation and measurement.

We test 3 research questions (RQs), as follows:

RQ1: What is the developmental course of externalizing, internalizing, and thought-disordered psychopathology for boys and girls across 3 to 7 years of age? To examine this question, we examine growth curves of developmentally scaled scores that leverage the construct-valid items.

RQ2: Which approach to longitudinal assessment is most accurate using mean scoring: the common items, upward/downward extension, or construct-valid items approach? To examine this question, we compare their criterion validity for estimating externalizing problems by examining their strength of association with researcher observations of children's externalizing behavior, including (non)compliance and (in)attention to task.

RQ3: Is developmental scaling of construct-valid items on different measures across time an accurate and useful way to place scores from different measures onto the same scale and account for heterotypic continuity? We compare the accuracy (criterion validity) of developmental scaling using different measures across time (ie, construct-valid items) to traditional approaches (ie, mean scoring of the common items and upward/downward extension approaches) that use the same items across time and ignore heterotypic continuity. We also evaluate whether developmentally scaled scores of externalizing problems show usefulness (incremental validity) in predicting the observations of noncompliance and inattention over and above predictions from traditional approaches. In addition, we evaluate traditional approaches' misclassification of individuals' persistence vs desistance of behavior problems.

We hypothesized that psychopathology undergoes changes in behavioral manifestation across development, consistent with heterotypic continuity; and consequently, that to accurately assess children's development of psychopathology, it is necessary to use age-differing measures with approaches that link the scores from the different measures to be on the same scale. If this hypothesis is true, we predict the following: (1) the construct-valid items approach to assessment will show greater accuracy (ie, stronger criterion validity) than the common items and upward/downward extension approaches; and (2) developmental scaling of the construct-valid items will show incremental validity above and beyond the common items and upward/downward extension approaches. Our hypotheses were preregistered (https://osf.io/jzxb8).

METHOD

Participants

A community sample of children (N = 231) and their caregivers participated in an ongoing accelerated longitudinal study. Participants were recruited at 1 of 4 ages: 36 (n = 62), 45 (n = 54), 54 (n = 53), or 63 (n = 62) months. Spanning all timepoints, children ranged from 3 to 7.5 years of age. The inclusion criterion to be recruited for the study was that the child was one of the target ages (described above). Exclusion criteria were as follows: the child's primary caregiver did not speak English, or the child did not have a permanent guardian, did not have normal or corrected-to-normal vision and hearing, or was not capable of communicating or following basic instructions in English. Participants were recruited in 2018 to 2024 through a biomedical registry of children who had well-child checkups at the University of Iowa Health Care Medical Center, from university listservs, and from advertisements and in-person recruitment activities at local schools, daycare facilities, and preschools, Women, Infants, and Children (WIC) programs, pediatricians' offices, and community events; and by word of mouth. Reasons for participant ineligibility and a flowchart of the final sample are shown in Figure S1, available online. Extent of missingness and tests of systematic missingness are in Supplement 1, available online. The sample consisted of children, their primary caregiver, the primary caregiver's parenting partner (as applicable), and a teacher/secondary caregiver (eg, nanny, babysitter, or someone else who knew the child well). Participant demographics are listed in

Table 1. Compared to the US population, sample participants were somewhat more likely to be White, married, middle or upper class, and to have a college or graduate degree. Participant demographics were broadly reflective of the surrounding area.

Procedures

The child and their primary caregiver (ie, parent) participated in 2 laboratory visits every 9 months for 4 timepoints (Figure S2, available online). The first laboratory visit consisted of parent–child interaction tasks and behavioral tasks; it lasted ∼2.5 hours. During the behavioral tasks, the parent completed questionnaires on their child. The second laboratory visit consisted of the child completing computerized tasks while electroencephalography was recorded; it lasted ∼2 hours. The primary caregiver's parenting partner (as applicable) and a teacher/secondary caregiver were invited to complete electronic questionnaires on the child. Video examples of procedures are available on Databrary (https://nyu.databrary.org/volume/1559).

Measures

The present study is part of a larger study, the School Readiness Study. Measures and hypotheses for the School Readiness Study were preregistered (https://osf.io/jzxb8). Data files, a data dictionary, analysis scripts, and a computational notebook for the present study are published online (https://osf.io/bgxma).

Behavior Ratings of the Child's Psychopathology. Children's psychopathology was rated by mothers, fathers, and teachers/secondary caregivers. Mothers and fathers completed the Child Behavior Checklist (CBCL) 1.5–5[30] or CBCL 6–18,[31] depending on the child's age. Teachers/secondary caregivers completed the Caregiver–Teacher Report Form (C–TRF)[30] or Teacher's Report Form (TRF),[31] depending on the child's age. Items were on a scale of 0 to 2, in which 0 = "Not true (as far as you know)"; 1 = "Somewhat or sometimes true"; 2 = "Very true or often true." Externalizing items were those from the Externalizing scale, which includes the Attention Problems and Aggressive Behavior subscales for the CBCL 1.5–5 and C–TRF; it includes the Rule-Breaking Behavior and Aggressive Behavior subscales for the CBCL 6–18 and TRF. Internalizing items were those from the Internalizing scale, which includes the Emotionally Reactive, Anxious/Depressed, Somatic Complaints, and Withdrawn subscales for the CBCL 1.5–5 and C–TRF; it includes the Anxious/Depressed, Withdrawn/Depressed, and Somatic Complaints subscales for the CBCL 6–18 and TRF. Thought

TABLE 1 Participant Demographics

	n/Mean	%/SD
Child		
Sex		
Male	122	52.8
Female	109	47.2
Age, y	4.96	1.19
Ethnicity		
Hispanic	38	16.5
Not Hispanic	193	83.5
Race		
Asian	9	3.9
Black or African American	16	6.9
White	170	73.6
More than 1 race	24	10.4
Other race	12	5.2
Primary caregiver		
Sex		
Male	21	9.1
Female	210	90.9
Age, y	35.36	5.31
Education		
Some high school (no degree)	6	2.6
High school degree	8	3.5
Some college	30	13.1
Associate's degree	28	12.2
Bachelor's degree	80	34.9
Master's degree	57	24.9
Professional school degree	8	3.5
Doctorate degree	12	5.2
Marital status		
Single/never married	29	12.7
Married	184	80.3
Separated	6	2.6
Divorced	8	3.5
Re-married	2	0.9
Parenting partner		
Sex		
Male	197	86.0
Female	32	14.0
Age, y	37.24	6.48
Education		
No schooling (or <1 y)	1	0.4
Nursery, kindergarten, or elementary (grades 1-8)	1	0.4
Some high school (no degree)	9	4.0
High school degree	25	11.2
Some college	40	17.9
Associate's degree	29	13.0
Bachelor's degree	61	27.4

(continued)

TABLE 1 Continued

	n/Mean	%/SD
Master's degree	35	15.7
Professional school degree	12	5.4
Doctorate degree	10	4.5
Family		
Income-to-needs ratio	3.53 (median = 3.05)	2.84
Hollingshead Four-Factor Index of Socioeconomic Status	47.9 (possible range: 1-66)	12.4
Nam–Powers–Boyd Occupation Status score, averaged across parents	71.2 (possible range: 0-100)	20.6
Study variables		
Externalizing problems (developmentally scaled)[a]	−0.77	2.67
Internalizing problems (developmentally scaled)[a]	−1.63	2.86
Thought-disordered problems (developmentally scaled)[a]	−2.05	2.96
Externalizing problems (construct-valid items)[b]	0.15	0.14
Externalizing problems (common items)[b]	0.14	0.15
Externalizing problems (all possible items; upward/downward extension approach)[b]	0.15	0.13
Externalizing problems (*T* score)	47.11	9.79
Internalizing problems (*T* score)	45.69	9.51
Attention to task (observation)	3.93 (possible range: 1-5)	0.96
Compliance (observation)	4.08 (possible range: 1-5)	0.99

Note: For some variables, the number of primary caregivers and parenting partners is greater than the number of children because the primary caregiver and parenting partner in some cases changed over time. The sample-wide mean of the developmentally scaled scores is negative and the SD is greater than 1.0, indicating that behavior problems (across ages) tended to be lower than the latent level at age 3 years and that scores tended to show greater variability relative to the variability at age 3.

[a]*Scores are on the scale of the factor scores at 3 years of age.*

[b]*Computed with a mean (proportion) score.*

disorder items were those from the Autism Spectrum Problems *DSM*-oriented scale for the CBCL 1.5–5 and C–TRF; thought disorder items were from the Thought Problems subscale for the CBCL 6–18 and TRF.

Externalizing and internalizing problems showed strong internal consistency (Table S1, available online), whereas thought-disordered problems showed weaker internal consistency. Behavior problem ratings showed cross-rater reliability (Table S2, available online) and cross-time rank-order stability (Table S3, available online). Cross-informant associations ranged from modest to large, consistent with prior work.[32,33] Using age and sex norm-referenced *T* scores of 65 as a clinical cutoff, ~12.6% and ~7.8% of children were in the at-risk or clinical range for externalizing problems and internalizing problems, respectively, at one or more timepoints based on ratings from one or more raters.

The Achenbach scales (CBCL 1.5–5, CBCL 6–18, TRF, C–TRF) have some items that are common across all instruments (ie, common items), and some items that are unique to a particular instrument (ie, unique items). The number of common items for each pair of measures is provided in Table 2. To examine effects of upward/downward extension, we assessed both the age-common and age-unique items of each of the CBCL Externalizing scales at all ages. That is, we assessed all items from the parent-reported Externalizing scales of the CBCL 1.5–5 and CBCL 6–18 at all ages (including the non–age-relevant items). Thus, our study is uniquely positioned to determine which approach is most accurate: (1) the "common items" approach, (2) the "upward/downward extension" approach, or (3) the "construct-valid items" approach (Figure 1). To minimize teachers' response burden, teachers did not complete non–age-relevant versions (eg, teachers of 6-year-olds did not complete the C–TRF). Given differing numbers of items and consistent with recommendations,[34] scores for the (non–developmentally scaled) scores were calculated by averaging scores across items and dividing by 2 (the maximum possible score for a given item), thus representing a proportion of the maximum possible score. Developmentally scaled scores were estimated using a Bayesian longitudinal item response model (described later).

Researcher Observations of the Child's Behavior Problems. Trained researchers observed the child's behavior problems live during 2 laboratory visits at each timepoint. Researchers observed the child's global degree of compliance, attention to task, and distress across the duration of each of the 2 laboratory visits. Researchers responded to the following question on a form at the end of each laboratory visit, which was designed specifically for this study: "Rate the target child in the following categories based on the entirety of the laboratory visit (1 = very low, 5 = very high): attention, compliance, distress." The

TABLE 2 Number of Common Items for Each Pair of Measures

Externalizing problems				
Measure	**CBCL 1.5–5**	**CBCL 6–18**	**C–TRF**	**TRF**
CBCL 1.5–5	24			
CBCL 6–18	7	35		
C–TRF	24	10	34	
TRF	9	28	12	32
Internalizing problems				
Measure	**CBCL 1.5–5**	**CBCL 6–18**	**C–TRF**	**TRF**
CBCL 1.5–5	36			
CBCL 6–18	12	32		
C–TRF	30	11	32	
TRF	11	30	11	33
Thought-disordered problems				
Measure	**CBCL 1.5–5**	**CBCL 6–18**	**C–TRF**	**TRF**
CBCL 1.5–5	12			
CBCL 6–18	1	15		
C–TRF	12	1	12	
TRF	1	10	1	10

Note: The upper third of the table presents the number of common items on the Externalizing scale. The middle third of the table presents the number of common items on the Internalizing scale. The lower third of the table presents the number of common thought-disordered problem items. Numbers on the diagonal represent the total number of items in the Externalizing scale (upper), Internalizing scale (middle), or thought-disordered problem items (lower) for that measure (eg, the CBCL 1.5–5 has 24 items on the Externalizing scale, 36 items on the Internalizing scale, and 12 thought-disordered problem items). Numbers below the diagonal represent, for that pair of measures, the number of items that are common to both of the measures. The number of unique items can be calculated by subtracting the number of common items from the total number of items. For instance, the CBCL 6–18 has 7 unique externalizing items when compared with the TRF (ie, 35 total items minus 28 common items). Conversely, the TRF has 3 unique externalizing items when compared with the CBCL 6–18 (ie, 33 total items minus 30 common items). CBCL = Child Behavior Checklist; C–TRF = Caregiver–Teacher Report Form; TRF = Teacher's Report Form.

present study focused on the observations of compliance and attention to laboratory visit tasks, given their relevance for externalizing psychopathology—namely, the inattention and oppositionality aspects of externalizing behavior. As such, each rating reflected observations of child compliance and attention across a range of tasks over the span of 2 to 3 hours, such as executive function and delay-of-gratification tasks during the first laboratory visit, and inhibitory control and attentional orienting tasks while wearing an electroencephalography cap during the second laboratory visit. We did not prompt observers to consider specific behaviors when forming their judgment about their ratings. One example of a behavior that observers considered inattentive was interrupting tasks with unrelated comments; one example of a behavior that observers considered noncompliant was moving around excessively when asked to sit still.

Three researchers provided observation ratings at the end of the first laboratory visit; 2 researchers provided observation ratings at the end of the second laboratory visit. In total, at a given timepoint, 5 observers provided ratings based on ~5 hours of laboratory visits. Each researcher's ratings were made masked to the ratings by the other observers. Such post-visit researcher direct observation ratings are commonly used in developmental research[35] and have been found to be associated with parent ratings and task-based coder ratings of the same behaviors in prior studies.[36] Interobserver reliability was intracorrelation coefficient (ICC)[2,k] = 0 .93 and 0.85 for compliance at the first and second laboratory visit, respectively. Interobserver reliability was ICC[2,k] = 0.91 and 0.84 for attention to task at the first and second laboratory visit, respectively. Correlations of ratings across the first and second laboratory visits were r = 0.60 for compliance (p values <.001) and r = 0.67 for attention to task. Ratings were first averaged across raters within visit, and then were averaged across visits within timepoints.

Statistical Analysis

We used developmental scaling to link scores from the construct-valid items, from different measures, across ages and raters onto the same scale (Supplement 2, available

online). The item response theory (IRT) approach to developmental scaling places scores from different measures onto the same scale by linking the item response theory parameters of the common items (ie, easiness and discrimination) to ensure that the scores are comparable across ages and raters. We used a 2-parameter Bayesian longitudinal item response model in a mixed modeling item response theory framework that simultaneously does the following: (1) performs developmental scaling to link the scores from the differing measures onto the same scale, (2) estimates children's growth curves of psychopathology, and (3) accounts for potential differential item functioning across ages and raters. This allowed charting children's development of psychopathology on a comparable scale across ages 3 to 7 years, despite measurement changes. We included a quadratic trend to allow curvature in children's trajectories over time. Age in years was centered to set intercepts at age 3 years, the youngest age in the sample. Age 3, therefore, serves as the reference scale, with parameter estimates for subsequent time points modeled relative to this baseline. Thus, children's factor scores (theta) were on the scale of the factor scores at 3 years of age.

In this longitudinal item response theory model, item and person parameter estimates across ages and raters were placed on a common scale using common items and simultaneous calibration. We accounted for age-related differences in item functioning of item parameters (easiness and discrimination), which ensured that the person parameter estimates at different age groups were expressed on the same scale. Rather than estimating parameters separately for each age group, all age-specific measures were calibrated simultaneously within a single model, resulting in item and person parameter estimates on the same underlying scale. To prevent arbitrary shifts in scale, the model imposed an additional constraint: the variance of the latent factor at age 3 was fixed at 1, and the mean was ~ 0. The scales for subsequent ages were then determined through the common items, preserving score comparability over time.[37,38]

The 2-parameter item response model estimates 2 parameters of each item: easiness and discrimination. Easiness is the expected score on the item at a given level of the latent factor. Discrimination of an item reflects how strongly the item is associated with the latent factor. The model used a multidimensional item response theory approach to allow items to load onto the latent externalizing, internalizing, externalizing, or thought-disordered factor. This allowed borrowing information from each dimension in estimation of the others for more accurate estimates, given considerable covariation among externalizing, internalizing, and thought-disordered problems. Items on the Externalizing scale of the CBCL and (C–)TRF were allowed to load onto the latent externalizing factor. Items on the Internalizing scale of the CBCL and (C–)TRF were allowed to load onto the latent internalizing factor. Items on the Autism Spectrum Problems (CBCL 1.5–5, C–TRF) and Thought Problems (CBCL 6–18, TRF) subscales were allowed to load onto the latent thought-disordered factor. The item response theory model was a graded response model with a cumulative response distribution and a logit link, which allows ordinal responses.

We also compared criterion validity of developmentally scaled scores and traditional scoring methods including a mean (ie, proportion) score for each of the following: (1) the "common items," (2) items from the "upward/downward extension" approach, and (3) the "construct-valid items" (ie, scoring the items used in the developmental scaling model with traditional methods). To compare their criterion validity, we used the Fisher *r*-to-*z* tests to examine their strength of association with researcher observations of children's behavior problems (ie, low compliance and attention to task). Using Bayesian mixed-effects models, we also evaluated whether developmental scaling of the construct-valid items showed incremental validity above and beyond the traditionally scored common items and upward/downward extension approaches in predicting research observations of behavior. Mixed-effects models included random intercepts for the participant to account for nonindependence of multiple observations from the same child, given the longitudinal and multi-informant study design. Models were fit using the brms package[39] version 2.22.0 in R[40] version 4.3.1. Models were estimated with 4 chains and 4,000 iterations. We kept default brms priors,[39] which use vague but proper priors.

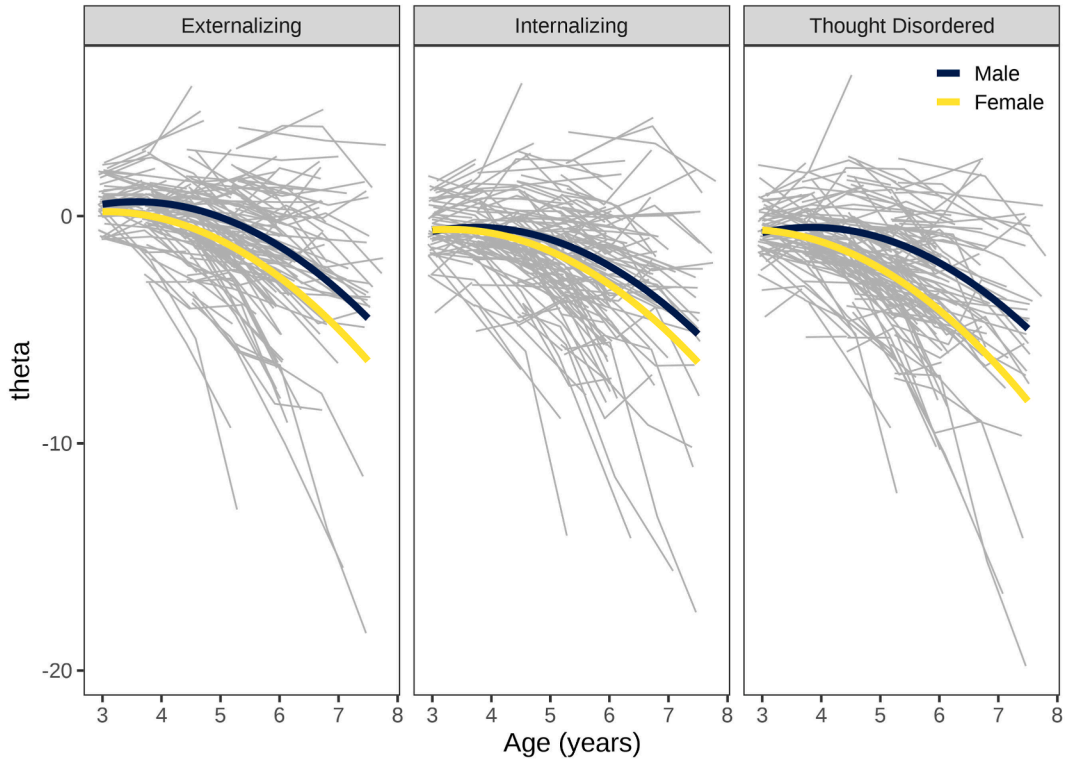
Parents and teachers completed the items for the relevant full scale; however, only parents completed the additional items for the non–age-relevant version that are necessary for evaluating the upward/downward extension approach. Thus, for a fair comparison with the upward/downward extension approach, we used only parents' ratings. For comparisons that did not involve the upward/downward extension approach (common items vs construct-valid items; traditional scoring vs developmentally scaling), we used both parents' and teachers' ratings. Tests were 2-tailed.

To distinguish individuals' growth curves as either persisting or desisting, we conducted latent class growth analyses (Supplement 3, available online).

Sensitivity Analysis

We conducted sensitivity analyses when imposing approximate longitudinal measurement invariance constraints in the developmental scaling model (Supplement 4, available online).

FIGURE 2 Developmentally Scaled Trajectories of Externalizing, Internalizing, and Thought-Disordered Problems



Note: *The figure depicts participants' model-implied values of theta overlaid with the sample's model-implied trajectory by sex. The gray lines represent participants' model implied values of theta by age. The yellow and dark blue lines represent the sample's model-implied trajectory by sex. Theta represents the estimate of the person's level on the latent factor. Theta is on the scale of the factor scores at 3 years of age. A theta of zero represents the average latent level of behavior problems at age 3; a positive theta indicates that the child is above the average latent level of behavior problems at age 3; a negative theta indicates that the child is below the average latent level of behavior problems at age 3.*

RESULTS

Descriptive statistics and bivariate correlations for study variables are provided in Table S4, available online. Numbers of observations at each timepoint are in Table S5, available online. Regression coefficients of the Bayesian item response model are in Table S6, available online. Item parameters by age and rater are in Table S7, available online. Estimates of changes in item functioning with age are in Table S8, available online. A description of item functioning is provided in Supplement 5, available online.

Developmental Course of Externalizing, Internalizing, and Thought-Disordered Problems

Developmentally scaled trajectories of externalizing, internalizing, and thought-disordered problems are depicted in Figure 2. On average, boys and girls showed rapid decreases in levels of psychopathology across 3 to 7 years of age. Boys tended to show higher levels than girls at the older ages, particularly in externalizing and thought-disordered problems. Girls showed steeper declines than boys in externalizing and thought-disordered problems. In latent class growth analyses, the common items and upward/downward extension approaches led to considerable misclassification of persistence vs desistance with respect to developmental scaling.

Criterion Validity

The Fisher r-to-z tests of comparative criterion validity are shown in Table S10, available online. As expected based on theory and simulation work, average scores of externalizing problems using the construct-valid items approach showed stronger criterion validity (compliance: $r = -0.22$; attention to task: $r = -0.24$) compared to the following: (1) the common items approach (compliance: $r = -0.15$; attention to task: $r = -0.16$); (2) the upward/downward extension approach (compliance: $r = -0.09$; attention to task: $r = -0.12$); (3) age and sex norm-referenced T scores from the Achenbach scales; and (4) in-sample z scores normed within age (ie, wave) (Fisher r-to-z tests: $r_{diff} = 0.07$-0.13) (p values $<.001$). Moreover, developmentally scaled scores of the construct-valid items showed the strongest criterion validity (compliance: $r = -0.25$;

TABLE 3 Bayesian Mixed-Effect Models

		Outcome: Attention to task				
Model	**Predictor**	**β**	**SE**	**Lower**	**Upper**	**Significantly greater**
1	Common items	0.06	0.03	0.00	0.11	
1	Developmentally scaled construct-valid items	−0.24	0.04	−0.30	−0.17	X
2	All possible items (upward/downward extension)	0.15	0.04	0.07	0.23	
2	Developmentally scaled construct-valid items	−0.37	0.05	−0.46	−0.28	X
3	*T* scores (age and sex norm-referenced)	0.13	0.03	0.07	0.19	
3	Developmentally scaled construct-valid items	−0.29	0.04	−0.36	−0.22	X
4	*z* Scores (normed within age)	0.13	0.03	0.07	0.18	
4	Developmentally scaled construct-valid items	−0.29	0.04	−0.36	−0.22	X
		Outcome: Compliance				
Model	**Predictor**	**β**	**SE**	**Lower**	**Upper**	**Significantly Greater**
5	Common items	0.04	0.03	−0.02	0.09	
5	Developmentally scaled construct-valid items	−0.18	0.04	−0.25	−0.11	X
6	All possible items (upward/downward extension)	0.15	0.04	0.07	0.23	
6	Developmentally scaled construct-valid items	−0.30	0.05	−0.40	−0.20	X
7	*T* scores (age and sex norm-referenced)	0.12	0.03	0.06	0.18	
7	Developmentally scaled construct-valid items	−0.24	0.04	−0.31	−0.17	X
8	*z* Scores (normed within age)	0.10	0.03	0.04	0.16	
8	Developmentally scaled construct-valid items	−0.22	0.04	−0.30	−0.15	X

Note: *"Lower" and "Upper" represent the bounds of the 95% credible interval. Beta coefficients represent standardized regression coefficients. Developmentally scaled construct-valid items represent significant terms in the expected direction such that externalizing problem ratings for the child are negatively associated with researchers' observation of the child's compliance and attention to task (ie, greater scores on ratings of externalizing problems are associated with poorer compliance and attention to task). "Significantly greater" indicates whether a predictor showed a significantly stronger beta than the other predictor in the model.*

attention to task: $r = -0.28$). Developmentally scaled scores of the construct-valid items showed significantly stronger criterion validity than the common items and upward/downward extension approaches and *T* and *z* scores ($r_{diff} = 0.10$-0.17; *p* values $<.001$). Compared to traditional scoring of the construct-valid items, developmental scaling showed stronger criterion validity in relation to the child's attention to task ($r_{diff} = 0.04$; $z = -2.25$, $p = .025$); however, there was no significant difference in relation to the child's compliance ($r_{diff} = 0.03$; $z = -1.29$, $p = .199$).

Incremental Validity

Results of the Bayesian mixed-effects models for evaluating incremental validity are provided in Table 3. Developmentally scaled scores of externalizing problems predicted observations of the child's compliance and attention to task over and above the following: (1) the common items, (2) items from the upward/downward extension approach, (3) age and sex norm-referenced *T* scores, and (4) *z* scores normed within age. The reverse was not true: the common items, items from the upward/downward extension approach, and *T* and *z* scores did not show incremental validity above the developmentally scaled scores of the construct-valid items.

DISCUSSION

Our approach to developmental scaling placed scores from different measures of psychopathology across ages and raters onto the same scale (ie, the construct-valid items approach). All forms of psychopathology

decreased across 3 to 7 years of age, for both boys and girls; however, girls showed steeper declines than boys, particularly for externalizing and thought-disordered problems. The most accurate approach, in terms of criterion validity of estimates of externalizing problems in relation to observations of compliance and attention to task, was the use of different items across ages—that is, the construct-valid items approach. The traditionally scored construct-valid items approach was modestly more accurate (r_{diff} = 0.07-0.13) than the 2 most widely used traditionally scored approaches to studying people's trajectories—the common items and the upward/downward extension approaches. Developmental scaling of the construct-valid items—to put the different measures onto the same scale—led to the greatest criterion validity, consistent with the idea that it was the most accurate approach. Developmentally scaled scores showed modestly stronger criterion validity than traditional assessment and scoring approaches (r_{diff} = 0.10-0.17). Moreover, developmentally scaled scores showed moderately stronger criterion validity than mean scores of the construct-valid items (r_{diff} = 0.03-0.04). In addition, developmentally scaled scores of externalizing problems showed incremental validity above and beyond traditional scoring approaches in predicting observations of noncompliance and inattention. Furthermore, the common items and upward/downward extension approaches led to considerable misclassification of persistence vs desistance with respect to developmental scaling.

Our findings that developmental scaling using age-differing, construct-valid items was the most accurate approach is consistent with theory and findings from simulation work.[7] The age-related decreases in externalizing problems in early childhood are consistent with prior work.[22–24,41,42] The declines in internalizing problems in early childhood are consistent with some prior work,[16,24,41] whereas other studies have identified that internalizing problems show relative stability in level[24] or even increases[24,42–44] in early childhood, depending on the rater, for example, parent vs teacher. Studies of thought-disordered problems in community samples are less common; nevertheless, in a sample of young children with autism, girls tended to decrease more in autism symptom severity more than boys,[45,46] consistent with our findings.

The study had several limitations. First, the Achenbach scales are potentially less well suited to assess thought-disordered problems than they are to assess externalizing and internalizing problems. Using the *DSM*-oriented Autism Spectrum Problems (CBCL 1.5–5/C–TRF) and the Thought Problems (CBCL 6–18/TRF) subscales, there was only one common item across ages, which may limit the effectiveness of linking thought-disordered scores across development. Moreover, the internal consistency of the thought-disorder items was weak. Second, the sample was a community sample; future work should replicate and extend these findings in a clinical sample. Third, the criterion-related associations with researcher observations were modest, suggesting that researcher observations are an imperfect criterion. Fourth, we observed some systematic missingness, which could lead to altered trajectory estimates such as greater decreases than what might occur normatively. Nevertheless, the developmental courses that we identified are largely consistent with prior work, and we have no reason to expect that this would alter the pattern of associations. Future work should examine additional developmental transitions, such as preadolescence to adolescence.

The study also had key strengths. First, it was innovative; it is the first study to link scores from age- and rater-differing measures of externalizing, internalizing, and thought-disordered problems onto the same scale. Given known rater-role biases[47]—that is, systematic differences in ratings by mothers vs fathers vs teachers, accounting for systematic rater-role biases is a key strength of the study. In addition, it was the first study to empirically compare the construct-valid items approach to traditional measurement approaches. Second, it was longitudinal, allowing one to chart children's change. Third, it leveraged multiple perspectives of the child's behavior for more accurate estimates, including ratings by mothers, fathers, and teachers. Fourth, it evaluated the scoring approaches against researcher observations of the child's behavior, for more objective comparison.

Our study demonstrates the following: (1) that different measures may be necessary to account for heterotypic continuity, and (2) that developmental scaling is an accurate and useful approach to place scores from different measures onto the same scale for charting people's development and to account for heterotypic continuity. Findings suggest that developmental scaling may enable studying the development of psychopathology across the lifespan. Our work has key implications for assessment and charting people's development when studying a "moving target," which applies to many forms of psychopathology. This will lead to more accurate modeling of dimensions of psychopathology and their development. Moreover, researchers may also use this approach to more accurately assess psychological phenomena across cultures.

Our approach also holds promise for better integrating a developmental perspective into the Hierarchical Taxonomy of Psychopathology and the Research Domain Criteria; it allows studying Hierarchical Taxonomy of

Psychopathology and Research Domain Criteria dimensions as they manifest in different ways across development, which researchers had not been able to do effectively with traditional approaches.

CRediT authorship contribution statement

Isaac T. Petersen: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Zachary Demko:** Writing – review & editing, Visualization, Conceptualization. **Won-Chan Lee:** Writing – review & editing, Methodology, Formal analysis. **Jacob J. Oleson:** Writing – review & editing, Methodology, Formal analysis.

This article is part of a special series of review and empirical articles devoted to examining dimensional alternatives to categorical diagnostic approaches for improving research insights and clinical practice in the field of child and adolescent neurodevelopmental and neuropsychiatric conditions. This series is edited by Guest Editors Mirko Uljarević, MD, PhD, Robert Krueger, PhD, Eric Youngstrom, PhD, Andrew Whitehouse, PhD; Associate Editor Robert R. Althoff, MD, PhD; *JAACAP Open* Editor Manpreet K. Singh, MD, MS; and *JAACAP* Editor-in-Chief Douglas K. Novins, MD.

Accepted October 21, 2025.

[a]University of Iowa, Iowa City, Iowa.

This project was funded by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (HD098235). Mr. Demko was funded by the National Institute of General Medical Sciences (T32GM108540; T32GM149386).

The present study was approved by the University of Iowa Institutional Review Board (Study #: 201701837). The authors complied with APA ethical standards in the treatment of participants.

This work has been posted on a preprint server: https://doi.org/10.31234/osf.io/p4kgw_v3.

Data Sharing: The data, a data dictionary, analysis scripts, and a computational notebook for the present study are published online: https://osf.io/bgxma.

Jacob J. Oleson served as the statistical expert for this research.

The authors thank the children and families who participated in the study.

Disclosure: Isaac T. Petersen, Zachary Demko, Won-Chan Lee, and Jacob J. Oleson have reported no biomedical financial interests or potential conflicts of interest.

*Correspondence to Isaac T. Petersen, PhD, University of Iowa, G60 Psychological and Brain Sciences Building, Iowa City, IA 52242; e-mail: isaac-t-petersen@uiowa.edu

https://doi.org/10.1016/j.jaacop.2025.10.008

REFERENCES

1. Petersen IT. Reexamining developmental continuity and discontinuity in the 21st century: better aligning behaviors, functions, and mechanisms. Dev Psychol. 2024;60(11):1992-2007. https://doi.org/10.1037/dev0001657
2. Miller JL, Vaillancourt T, Boyle MH. Examining the heterotypic continuity of aggression using teacher reports: results from a national Canadian study. Soc Dev. 2009;18(1):164-180. https://doi.org/10.1111/j.1467-9507.2008.00480.x
3. Speranza AM, Liotti M, Spoletini I, Fortunato A. Heterotypic and homotypic continuity in psychopathology: a narrative review. Front Psychol. 2023;14:1194249. https://doi.org/10.3389/fpsyg.2023.1194249
4. Patterson GR. Orderly change in a stable world: the antisocial trait as a chimera. J Consult Clin Psychol. 1993;61(6):911-919. https://doi.org/10.1037/0022-006X.61.6.911
5. Chen FR, Jaffee SR. The heterogeneity in the development of homotypic and heterotypic antisocial behavior. J Dev Life-Course Criminol. 2015;1(3):269-288. https://doi.org/10.1007/s40865-015-0012-3
6. Petersen IT, Choe DE, LeBeau B. Studying a moving target in development: the challenge and opportunity of heterotypic continuity. Dev Rev. 2020;58:100935. https://doi.org/10.1016/j.dr.2020.100935
7. Petersen IT, LeBeau B, Choe DE. Creating a developmental scale to account for heterotypic continuity in development: a simulation study. Child Dev. 2021;92(1):e1-e19. https://doi.org/10.1111/cdev.13433
8. Petersen IT, Lindhiem O, LeBeau B, *et al*. Development of internalizing problems from adolescence to emerging adulthood: accounting for heterotypic continuity with vertical scaling. Dev Psychol. 2018;54(3):586-599. https://doi.org/10.1037/dev0000449
9. Conradt E, Crowell SE, Cicchetti D. Using development and psychopathology principles to inform the Research Domain Criteria (RDoC) framework. Dev Psychopathol. 2021;33(5):1521-1525. https://doi.org/10.1017/S0954579421000985
10. Durbin CE, Wilson S, MacDonald AW III. Integrating development into the Research Domain Criteria (RDoC) framework: introduction to the special section. J Psychopathol Clin Sci. 2022;131(6):535-541. https://doi.org/10.1037/abn0000767
11. Tackett JL, Hallquist M. The need to grow: developmental considerations and challenges for modern psychiatric taxonomies. J Psychopathol Clin Sci. 2022;131(6):660-663. https://doi.org/10.1037/abn0000751
12. Kagan J. Change and Continuity in Infancy. Wiley; 1971.
13. Kagan J. The three faces of continuity in human development. In: Goslin DA, ed. Handbook of Socialization Theory and Research. Rand McNally; 1969:983-1002.
14. Sai L, Shang S, Tay C, *et al*. Theory of mind, executive function, and lying in children: a meta-analysis. Dev Sci. 2021;24(5):e13096. https://doi.org/10.1111/desc.13096
15. Odgers CL, Moffitt TE, Broadbent JM, *et al*. Female and male antisocial trajectories: from childhood origins to adult outcomes. Dev Psychopathol. 2008;20(2):673-716. https://doi.org/10.1017/S0954579408000333
16. Sterba SK, Prinstein MJ, Cox MJ. Trajectories of internalizing problems across childhood: heterogeneity, external validity, and gender differences. Dev Psychopathol. 2007;19(2):345-366. https://doi.org/10.1017/S0954579407070174
17. Briggs-Gowan MJ, Godoy L, Heberle A, Carter AS. Assessment of psychopathology in young children. In: Cicchetti D, ed. Developmental Psychopathology. Hoboken, NJ: John Wiley & Sons; 2016:1-45.
18. Broeren S, Muris P, Diamantopoulou S, Baker JR. The course of childhood anxiety symptoms: developmental trajectories and child-related factors in normal children. J Abnorm Child Psychol. 2013;41(1):81-95. https://doi.org/10.1007/s10802-012-9669-9
19. Tong Y, Kolen MJ. Comparisons of methodologies and results in vertical scaling for educational achievement tests. Appl Meas Educ. 2007;20(2):227-253. https://doi.org/10.1080/08957340701301207
20. Schlechter P, Wilkinson PO, Ford TJ, Neufeld SAS. The Short Mood and Feelings Questionnaire from adolescence to emerging adulthood: measurement invariance across time and sex. Psychol Assess. 2023;35(5):405-418. https://doi.org/10.1037/pas0001222
21. Kolen MJ, Brennan RL. Test Equating, Scaling, and Linking: Methods and Practices. (Statistics for Social and Behavioral Sciences.). 3rd ed. Springer; 2014:566.
22. Owens EB, Shaw DS. Predicting growth curves of externalizing behavior across the preschool years. J Abnorm Child Psychol. 2003;31(6):575-590. https://doi.org/10.1023/a:1026254005632
23. Petersen IT, Bates JE, Dodge KA, Lansford JE, Pettit GS. Describing and predicting developmental profiles of externalizing problems from childhood to adulthood. Dev Psychopathol. 2015;27(3):791-818. https://doi.org/10.1017/S0954579414000789
24. Harris JL, LeBeau B, Petersen IT. Reactive and control processes in the development of internalizing and externalizing problems across early childhood to adolescence. Dev Psychopathol. 2025;37(2):836-858. https://doi.org/10.1017/S0954579424000713
25. Petersen IT, LeBeau B. Creating a developmental scale to chart the development of psychopathology with different informants and measures across time. J Psychopathol Clin Sci. 2022;131(6):611-625. https://doi.org/10.1037/abn0000649
26. McArdle JJ, Grimm KJ, Hamagami F, Bowles RP, Meredith W. Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. Psychol Methods. 2009;14(2):126-149. https://doi.org/10.1037/a0015857
27. McArdle JJ, Grimm KJ. An empirical example of change analysis by linking longitudinal item response data from multiple tests. In: von Davier AA, ed. Statistical Models

for Test Equating, Scaling, and Linking. Springer Science & Business Media; 2011:71-88.
28. Weeks JP. An application of multidimensional vertical scaling. Meas Interdiscip Res Perspect. 2018;16(3):139-154. https://doi.org/10.1080/15366367.2018.1502005
29. Caspi A, Houts RM, Belsky DW, *et al*. The p factor: one general psychopathology factor in the structure of psychiatric disorders? Clin Psychol Sci. 2014;2(2):119-137. https://doi.org/10.1177/2167702613497473
30. Achenbach TM, Rescorla LA. Manual for the ASEBA Preschool Forms and Profiles: An Integrated System of Multi-Informant Assessment. University of Vermont, Department of Psychiatry; 2000.
31. Achenbach TM, Rescorla LA. Manual for the ASEBA School-Age Forms and Profiles. University of Vermont, Research Center for Children, Youth, and Families; 2001.
32. Achenbach TM, McConaughy SH, Howell CT. Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. Psychol Bull. 1987;101(2):213-232. https://doi.org/10.1037/0033-2909.101.2.213
33. De Los Reyes A, Augenstein TM, Wang M, *et al*. The validity of the multi-informant approach to assessing child and adolescent mental health. Psychol Bull. 2015;141(4):858-900. https://doi.org/10.1037/a0038498
34. Little TD. Longitudinal Structural Equation Modeling. Methodology in the Social Sciences. Guilford Press; 2013.
35. Al-Hendawi M, Hussein E, Darwish S. Direct observation systems for child behavior assessment in early childhood education: a systematic literature review. Disc Ment Health. 2025;5(1):21. https://doi.org/10.1007/s44192-025-00139-z
36. Augustine ME, Moding KJ, Stifter CA. Person-centered profiles of child temperament: a comparison of coder, mother, and experimenter ratings. Infant Behav Dev. 2022;68:101725. https://doi.org/10.1016/j.infbeh.2022.101725
37. Wang C, Nydick SW. On longitudinal item response theory models: a didactic. J Educ Behav Stat. 2020;45(3):339-368. https://doi.org/10.3102/1076998619882026
38. Huang H-Y. A multilevel higher order item response theory model for measuring latent growth in longitudinal data. Appl Psychol Meas. 2015;39(5):362-372. https://doi.org/10.1177/0146621614568112
39. Bürkner P-C. brms: an R package for Bayesian multilevel models using Stan. J Stat Softw. 2017;80(1):28. https://doi.org/10.18637/jss.v080.i01
40. R: a language and environment for statistical computing. R Foundation for Statistical Computing; 2023; http://www.R-project.org
41. Shi Q, Ettekal I, Deutz MHF, Woltering S. Trajectories of pure and co-occurring internalizing and externalizing problems from early childhood to adolescence: associations with early childhood individual and contextual antecedents. Dev Psychol. 2020;56(10):1906-1918. https://doi.org/10.1037/dev0001095
42. Bongers IL, Koot HM, van der Ende J, Verhulst FC. The normative development of child and adolescent problem behavior. J Abnorm Psychol. 2003;112(2):179-192. https://doi.org/10.1037/0021-843x.112.2.179
43. Colder CR, Mott JA, Berman AS. The interactive effects of infant activity level and fear on growth trajectories of early childhood behavior problems. Dev Psychopathol. 2002;14(1):1-23.
44. Gilliom M, Shaw DS. Codevelopment of externalizing and internalizing problems in early childhood. Dev Psychopathol. 2004;16(2):313-333. https://doi.org/10.1017/S0954579404044530
45. Waizbard-Bartov E, Ferrer E, Heath B, *et al*. Identifying autism symptom severity trajectories across childhood. Autism Res. 2022;15(4):687-701. https://doi.org/10.1002/aur.2674
46. Waizbard-Bartov E, Ferrer E, Young GS, *et al*. Trajectories of autism symptom severity change during early childhood. J Autism Dev Disord. 2021;51(1):227-242. https://doi.org/10.1007/s10803-020-04526-z
47. Bauer DJ, Howard AL, Baldasaro RE, *et al*. A trifactor model for integrating ratings across multiple informants. Psychol Methods. 2013;18(4):475-493. https://doi.org/10.1037/a0032475

Supplement 1. Description of Missing Data.

A total of 230 children had behavior problem ratings from a mother, father, and/or teacher/secondary caregiver; 225 children had scores from researchers' observations of behavior. The number of participants who had data, disaggregated by the number of time points, is in Table S5. Due to the ongoing nature of the longitudinal study, some data are missing because they are not yet available: 14% of participant-by-timepoint instances (i.e., lab visits) are not yet eligible, 4% are eligible and to-be-scheduled, and 2% are eligible and scheduled but not yet conducted. Among eligible participant-by-timepoint instances, 77% of time points had one or more raters provide ratings, and 66% of time points had researchers' observations of behavior. Among missing lab visits at a given time point for which the child reached eligibility, reasons for missingness included: not interested (15%), too busy (15%), moved/relocated (6%), unable to contact (29%), coronavirus (COVID-19) pandemic (28%), and other (6%). Thus, much of the missing instances were due to the COVID-19 pandemic or to not yet being eligible. We suspended lab visits for 14 months during the COVID-19 pandemic (March 2020 – April 2021). We continued to collect online questionnaires from families during the pandemic but were unable to perform lab visits and behavior observations during this period.

We examined whether missingness was systematic in the behavior problem ratings or researchers' behavior observations. In general, older children were more likely to be missing behavior problem ratings (at a trend level; $t[10.48] = -2.14$, $p = .057$) and behavior observations ($t[138.81] = -6.81$, $p < .001$) compared to younger children, presumably due to attrition, which is common in longitudinal studies. Behavior problem ratings ($t[234.42] = 4.97$, $p < .001$) and behavior observations ($t[465.59] = 3.81$, $p < .001$) were also more likely to be missing for those from lower socioeconomic status households. Boys were more likely than girls to be missing

behavior problem ratings ($\chi^2[1] = 7.67$, $p = .005$) and behavior observations ($\chi^2[1] = 3.89$, $p = .049$). For behavior problem ratings, compared to Non-Hispanic White participants, ratings were more likely to be missing for Asian ($\chi^2[1] = 11.47$, $p < .001$), Black ($\chi^2[1] = 13.85$, $p < .001$), and Hispanic ($\chi^2[1] = 7.40$, $p = .007$) participants, and participants with some other race ($\chi^2[1] = 5.90$, $p = .015$), but not for multiracial participants ($\chi^2[1] = 2.57$, $p = .109$). For behavior observations, compared to Non-Hispanic White participants, ratings were more likely to be missing for Black ($\chi^2[1] = 5.06$, $p = .024$) and multiracial ($\chi^2[1] = 5.77$, $p = .016$) participants, but not for Asian ($\chi^2[1] = 2.77$, $p = .096$), or Hispanic ($\chi^2[1] = 2.01$, $p = .156$) participants, or for participants with some other race ($\chi^2[1] = 0.66$, $p = .416$). Compared to participants with behavior problem ratings, participants who were missing behavior problem ratings tended to be higher in observed compliance ($t[4.42] = 3.38$, $p = .024$) but did not differ in observed attention to task ($t[1.13] = 3.97$, $p = .134$). Compared to participants with behavior observation scores, participants who were missing behavior observations actually tended to be *lower* in externalizing ($t[194.26] = 4.50$, $p < .001$), internalizing ($t[193.68] = 4.46$, $p < .001$), and thought-disordered ($t[192.44] = 5.17$, $p < .001$) problems.

Supplement 2. Developmental Scaling Approach.

We used developmental scaling to link scores from different age-specific measures onto a common latent scale. This allowed meaningful longitudinal comparisons and accurate estimation of psychopathology trajectories. Developmental scaling benefits from having some age-common items from measures at adjacent ages, to serve as an anchor. Developmental scaling can leverage the common items to link the scores from the different measures onto the same scale (i.e., the scale of the reference age). By assuming no changes in functioning of the common items across ages (i.e., no differential item functioning) or by accounting for any changes in item functioning, this approach ensures that both item and latent parameter estimates are expressed on a common scale across ages. Nevertheless, developmental scaling can use all construct-valid items—common and unique items—to estimate people's scores on that scale, thus making use of all construct-valid information while estimating people's scores on a comparable scale across development.

To perform developmental scaling, we used a two-parameter Bayesian longitudinal multidimensional item response model in a mixed modeling item response theory (IRT) framework. Such a model allows us to simultaneously account for heterotypic continuity using different measures across time and to model children's trajectories. The model linked scores from measures across all ages in the same model through simultaneous calibration with a specified reference age. This approach is similar to multiple-group concurrent calibration in test equating and linking[1,2], but it differs in that our longitudinal modeling approach focuses on growth trajectories over time by examining repeated measures of individuals. By incorporating the longitudinal IRT framework, the simultaneous calibration accounts for the within-person dependence of scores over time and is thus more suitable than the group-based concurrent

calibration approach. Moreover, simultaneous calibration results in more precise and stable estimates than separate (two-stage) calibration in which separate models across age are fit.[1,3]

In the longitudinal multidimensional IRT model of the present study, item and person parameter estimates across ages and raters were placed on a common scale using common items and simultaneous calibration. We accounted for age-related differences in item functioning of item parameters (easiness and discrimination), which ensured that the person parameter estimates at different age groups were expressed on the same scale. Rather than estimating parameters separately for each age group, all age-specific measures were calibrated simultaneously within a single model, resulting in item and person parameter estimates on the same underlying scale. To prevent arbitrary shifts in scale, the model imposed an additional constraint: The variance of the latent factor at age 3 was fixed at 1 and the mean was ~0. The scales for subsequent ages were then determined through the common items, preserving score comparability over time.[4,5]

The two-parameter item response model applied intercept–slope parameterization:

$$y_{ij} = \alpha_i \cdot \theta_j + \xi_i \quad (1)$$

where y_{ij} is the score for person j on item i. The model estimates two parameters for each item: easiness (ξ; equivalent to −1 × difficulty, severity, or threshold) and discrimination (α). The item's easiness parameter is the expected score on an item at a given level of the construct, and is similar to the intercept parameter in factor analysis.[6] The item's discrimination parameter is how strongly the item is associated with the construct, and is similar to a factor loading. Easiness and discrimination provide information about the functioning and usefulness of each item—and the whole measurement scheme—at a given age. In addition, the model estimates a person parameter for each person (i.e., person j): theta (θ). The person parameter, theta, represents a person's level on the latent construct and is similar to a factor score.

A two-parameter logistic IRT model takes the following form:

$$P(y_{ij} = 1|\theta_j, \alpha_i, \xi_i) = \frac{e^{\alpha_i(\theta_j+\xi_i)}}{1+e^{\alpha_i(\theta_j+\xi_i)}} \quad (2)$$

where y_{ij} is the score for person j on item i, theta (θ_j) is the level on the construct for person j, xi (ξ_i) is the easiness parameter for item i, and alpha (α_i) is the discrimination parameter for item i.

For robust estimates of the child's level on each psychopathology dimension, we fit multidimensional item response models that included items assessing the three primary dimensions of psychopathology: externalizing problems, internalizing problems, and thought-disordered problems. This allowed borrowing information from each dimension in the estimation of the other, for more accurate estimates given considerable covariation between externalizing, internalizing, and thought-disordered problems.[7]

In the present study, behavior problem items were rated on a three-point scale that ranged from $y_{ij} = 0$–2. There were three possible response options (0, 1, 2), so there were two category boundaries: one boundary between 0 and 1, and one boundary between 1 and 2. Because the response options were ordinal, we fit a graded response model, which allows ordinal responses. We used a cumulative response distribution with a logit link. A two-parameter graded response model takes the following general form of Equation (3):

$$P(Y_{ij} = y_{ij}|\theta_j) = P^*_{y_{ij}}(\theta_j) - P^*_{y_{ij}+1}(\theta_j) \quad (3)$$

where:

$$P^*_{y_{ij}}(\theta_j) = P(Y_{ij} \geq y_{ij}|\theta_j, \alpha_i, \xi_{ic}) = \frac{e^{\alpha_i(\theta_j+\xi_{ic})}}{1+e^{\alpha_i(\theta_j+\xi_{ic})}} \quad (4)$$

where y_{ij} is the score for person j on item i, theta (θ_j) is the level on the construct for person j, xi (ξ_{ic}) is the easiness parameter for item i for category c, and alpha (α_i) is the discrimination

parameter for item i.

A cumulative ordinal logistic model involves two separate logistic regression models. The two logistic regressions have different intercepts (easiness) but share the same slopes (discrimination, theta). The first logistic model is outcome 0 versus 1 and 2. The second logistic model is outcome 0 and 1 versus 2. The regression coefficients are on the log odds scale and can be exponentiated for the odds ratio. As an item's easiness increases, the likelihood of endorsing the item increases. As a person's theta increases, the person's level on the construct increases, and thus, their likelihood of endorsing the item increases. In a cumulative ordinal model with a logit link, an item's discrimination parameter represents the increase in log-odds of endorsing a higher response category (e.g., from 0 to 1 or from 1 to 2) for a one-unit increase in a person's level on the latent construct (theta). For instance, if an item's discrimination parameter (α_i) is 1, then a one-unit increase in theta multiplies the odds of endorsing a higher category by exp(1) ≈ 2.72. As an item's discrimination increases, the item becomes more sensitive to differences in theta, meaning that small differences in the person's level on the construct lead to larger differences in the probability of endorsing the item. For instance, if an item's easiness parameter (ξ_{ic}) is -1, the item's discrimination parameter (α_i) is 2, and the person's level on the construct (θ_j) is 0.5, the linear predictor that gets passed to the logit link function is (based on Equation 1):

$$0 = (2 \times 0.5) + (-1)$$

In this case, when passed through the logit link, the probability of endorsing the item (i.e., endorsing 1 or 2 instead of 0, or endorsing 2 instead of 0 or 1) would be 0.5:

$$0.5 = \frac{1}{1+e^{-x}} = \frac{1}{1+e^{-0}} = \frac{1}{1+1}$$

We accounted for potential differential item functioning across ages and raters to ensure we were measuring the same construct across development in a comparable way. We estimated

the item's easiness parameter (ξ_{ic}) with fixed effects for the role of the rater (mother, father, or secondary caregiver), linear and quadratic terms for the child's age, and an age × role interaction. The rater role was dummy coded so that the mother rater was the reference group. The model included a random intercept and random slope for each item. The random slopes for each item were age, role, and an age × role interaction. This allowed each item to differ in its change in easiness over time for each rater type.

We estimated the item's discrimination parameter (α_i) with fixed effects for the psychopathology dimension assessed (externalizing, internalizing, or thought-disordered problems), the role of the rater, linear and quadratic terms for the child's age, and an age × role interaction. The model included a random intercept and random slope for each item. The random slopes for each item were age, rater role, and an age × role interaction. This allowed each item to differ in its change in discrimination over time for each rater type.

We performed the developmental scaling and estimation of growth curves in the same model. Growth curves were estimated for the person parameter (theta; θ_j), representing the child's level on the latent factor for a given dimension of psychopathology (externalizing, internalizing, or thought-disordered problems). A given child had up to four time points. Thus, a quadratic was the most complex polynomial of nonlinear growth we could estimate for children's trajectories that still allow measurement error. Because of prior work demonstrating that developmental trajectories of psychopathology are nonlinear[8], we modeled children's growth with a quadratic term. We modeled random intercepts and random linear and quadratic slopes to allow each child to differ in their starting point, form of growth, and curvature. Age in years was centered to set the intercepts (i.e., reference age) at age 3, the youngest age in the sample. Age 3, therefore, serves as the reference scale, with parameter estimates for subsequent

time points modeled relative to this baseline. Thus, the person and item parameters were on the same scale and were scaled relative to the parameters at 3 years of age. A theta of zero thus represents the average latent level of behavior problems at age 3; a positive theta indicates that the child is above the average latent level of behavior problems at age 3; a negative theta indicates that the child is below the average latent level of behavior problems at age 3.

We included the child's sex (female = 1, male = 0) and the rater role as a predictor of the intercepts and slopes. The fixed effect predictors of the intercepts were the child's age (linear and quadratic terms), rater role, dimension, dimension × role interaction, sex, and a sex × dimension interaction. The fixed effect predictors of the linear slopes were the rater role, dimension, dimension × role interaction, sex, and a sex × dimension interaction. This allowed evaluating sex-related differences in trajectories as a function of the psychopathology dimension. The random slopes for each person were age, quadratic age, role, age × role interaction, dimension, age × dimension interaction, dimension × role interaction, and an age × dimension × role interaction. This allowed each person to have a unique trajectory for each psychopathology dimension and rater type.

In a Bayesian model, the final step is to specify prior distributions for all remaining parameters in the model. The person parameters' (theta) variance was fixed to 1 for model identification and the lower bound of the discrimination parameter to zero, following recommendations for identifiability.[6,9] With two exceptions (described in the previous sentence), we kept the default priors used in the brms package[10], which uses vague but proper priors. The default priors for regression parameters were multivariate normal with mean zero and unknown covariance matrix Σ which follows a LKJ-correlation prior.[11] All standard deviation parameters were given a half Student-*t*-distribution prior with 3 degrees of freedom, mean 0, and scale

parameter 2.5. The prior for the intercept of item discrimination, item easiness, and theta was a flat prior.

Our model had no missing data in the predictors (age, sex, and rater); missingness was only in the outcome (scores on psychopathology items). Mixed models handle missing data in the outcomes. Mixed models provide valid inferences if the data are missing at random or completely at random.[12] Furthermore, our Bayesian hierarchical mixed model also provides valid inference when data are missing at random or completely at random. Because much of our missingness was due to the coronavirus 2019 (COVID-19) pandemic, we felt this modeling approach was appropriate. Moreover, researchers have argued against using multiple imputation in longitudinal designs that use mixed models because multiple imputation can lead to unstable estimates.[13]

Developmentally scaled factor scores were estimated from the posterior distribution by averaging model-predicted posterior samples across chains and iterations, using the posterior_epred() function from the brms package. Model-predicted posterior samples were averaged within combinations of child-by-measurement occasion-by-rater. This allowed each child to have a different factor score of externalizing, internalizing, and thought-disordered problems for each rater at each of their measurement occasions.

We fit the Bayesian longitudinal mixed models using the brm() function of the brms package 2.22[10] in R, which uses the RStan 2.32.6[14] interface to Stan 2.32.2[15] for Bayesian modeling. The models included four chains and 4,000 iterations.

Supplement 3. Latent Class Growth Analyses.

Method

A key goal for developmental psychopathology is to understand the factors that lead to persistence versus desistance of behavior problems. However, to identify processes that influence such developmental courses, it is first necessary to accurately establish which children persist versus desist. As described, there are many problems of using the common items and upward/downward extension approaches to assess children's development of behavior problems. Thus, it is important to determine the extent of misclassification of persistence and desistance that the common items and upward/downward extension approaches would yield with respect to more developmentally sensitive approaches. To determine the extent of misclassification of persistence and desistance that the common items and upward/downward extension approaches yield with respect to developmental scaling, we applied latent class growth analysis using each scoring approach.

Latent class growth analysis models were estimated in Mplus version 8.6[16]. All models used maximum likelihood estimation with robust standard errors (MLR) to account for the nonnormally distributed data. We fit separate latent class growth models for each behavior problem: externalizing, internalizing, and thought-disordered problems.

To estimate latent class growth trajectories given the multiple informants, we estimated a multi-informant composite for that behavior problem at each age based on mothers', fathers', and teachers'/secondary caregivers' reports as indicators. We constrained each informant to have a loading of 1.0 on the multi-informant composite for that behavior problem to (a) keep the composite on an interpretable metric across ages, (b) so as not to give undue weight to some informants than others (given the shared context of mothers and fathers), and (c) given findings

that simpler examination of information from multiple informants tends to do just as well if not better than more complex latent variables.[17] Then, using children's trajectories on the multi-informant composite for that behavior problem, we estimated the intercept, linear slope, and quadratic slope for each latent class. We estimated quadratic growth curves to allow for curvature. Quadratic was the most complex function form we could fit to each child's trajectory (while still accounting for measurement error) because each child had up to four timepoints. We compared the fit criteria and interpretability for one, two, three, and four-class models. Then, for fairness of comparison to the developmentally scaled trajectories, we estimated latent class growth analysis models with the same number of classes for the other scoring approaches: the common items approach and the upward/downward extension approach (which were available, by design, for externalizing problems only). For each model, we determined which class each child belonged to based on the class with which they had the highest assigned probability of membership.

Results

Fit criteria for the one, two, three, and four-class models are in Table S9. In general, as is common in studies using latent class growth analysis, model fit tended to improve as the number of classes increased. However, there appeared to be diminishing returns in fit as the number of classes increased. With three classes, the smallest class for all behavior problem outcomes was small (< 15%). Moreover, the classes tended to represent high, medium, and low trajectories, suggesting dimensional rather than categorical differences. Thus, models with three (or four) trajectories were not particularly interpretable given our goals. Given the modest sample size and our goals of evaluating persistence versus desistance, we selected the two-class models for interpretability. The two-class latent class growth trajectories for externalizing, internalizing, and

thought-disordered problems are in Figure S3. For fairness of comparison to the developmentally scaled trajectories, we estimated two-class models for the other scoring approaches: the common items approach and the upward/downward extension approach (for externalizing problems only).

For externalizing problems, the two-class model of developmentally scaled scores showed a large class that was high and stable/increasing (79.2%; "persisting") and a second class that was lower and decreasing (20.8%; "desisting"). Using the common items approach, there was a large class that was high (78.7%; "persisting") and a second class that was lower (21.3%; "low"). With respect to the classifications from the developmental scaling approach, the common items approach misclassified nearly 40% of children: 18.7% were misclassified as persisting and 20.9% were misclassified as low. Using the upward/downward extension approach, both classes were indistinguishable and increased. Thus, the upward/downward extension approach was particularly problematic in terms of accuracy for characterizing children's trajectories.

For internalizing problems, the two-class model of developmentally scaled scores showed a large class that was high and decreasing (63.8%; "high") and a second class that was lower and decreasing (36.2%; "low"). Using the common items approach, both classes were indistinguishable and decreased.

For thought-disordered problems, the two-class model of developmentally scaled scores showed a large class that was high and stable/decreasing (53.9%; "high") and a second class that was lower and decreasing (46.1%; "desisting"). There were not enough age-common items of thought-disordered problems to examine their trajectories in a latent class growth analysis.

Some of the classification discrepancies between the scoring approaches could reflect different percentages of children assigned to each class (i.e., different cutoffs). Nevertheless, the latent class growth analysis demonstrated that the traditional scoring approaches were unable to

characterize the children's trajectories in ways that were consistent with those obtained from the developmental scaling approach. In sum, the common items and upward/downward extension approaches led to considerable misclassification of persistence versus desistance with respect to developmental scaling.

Supplement 4. Sensitivity Analyses.

As a sensitivity analysis, we fit an additional developmental scaling model with longitudinal measurement invariance constraints. In frequentist approaches, measurement invariance is often tested in a four-step approach (configural, scalar, metric, and residual invariance). In a Bayesian approach, however, a conditional logic structure is preferred to sequential model testing because the models provide the information necessary to evaluate the extent of any non-invariance. The advantage of the conditional formulation of Bayesian mixed models is well described in many excellent textbooks[18-20] and the documentation for the R package brms[10]. Wikle et al.[21] (p. 10) wrote: "If most of the complex dependencies in the data are due to the underlying process of interest, then one should model the distribution of the data conditioned on that process (data model), followed by a model of the process' behavior and its uncertainties (process model)." Therefore, within a Bayesian mixed modeling framework, approximate measurement invariance is a process of interest which can be used to account for small instances of non-invariance.[22,23] Approximate measurement invariance involves setting narrow priors on the invariance parameters rather than fixing invariance parameters to zero.[22] Approximate measurement invariance is more accurate than full or partial measurement invariance for estimating true latent mean differences when there are many small differences in the intercepts and factor loadings across groups.[22,24] Thus, as a sensitivity analysis in the present study, we also fit a model that imposes approximate longitudinal measurement invariance.

In this model, we set the slopes of the discrimination parameters to be close to zero, by setting the prior of the discrimination parameter and easiness parameter to have a normal distribution with a mean of zero and a small standard deviation of 0.05. We also set the prior of the standard deviation of the random effect of item on the association of age with discrimination

and easiness to be small (normal distribution with mean = 0, *SD* = 0.01) so that items were restricted to be similar in their change of discrimination and easiness (i.e., near zero). This approximate measurement invariance approach successfully constrained the slopes of the discrimination and easiness parameters to be near-zero. None of the items showed significant differences in easiness (i.e., intercepts) across ages in this model. Several items showed reliable differences in discrimination (i.e., factor loadings) across ages, but differences were small.

Supplement 5. Item Functioning.

Item Severity and Discrimination

Item parameters by age and rater are in Table S7. For externalizing problems, items that were most strongly associated with the construct (i.e., had the highest discrimination) included items assessing the extent to which the child destroys others' things, has temper tantrums, and is physically aggressive toward others. Items that were more severe (i.e., had a lower easiness) included items assessing the extent to which the child is cruel to animals, is physically aggressive towards others, runs away from home, sets fires, and uses substances. Items that were less severe (i.e., had a higher easiness) included items assessing the extent to which the child argues, wants or demands attention, does not comply at home, and is stubborn.

For internalizing problems, items that were most strongly associated with the construct included items assessing the extent to which the child worries, has rapid changes in moods, feels guilt, and is withdrawn from others. Items that were more severe included items assessing the extent to which the child has somatic issues (e.g., headaches, nausea, vomiting, eye-related problems without a known medical cause) and talks about killing themself. Items that were less severe included items assessing the extent to which the child whines, is clingy to adults, has nightmares, is shy, and has their feelings hurt easily.

For thought-disordered problems, items that were most strongly associated with the construct included items assessing the extent to which the child shows strange behavior, little affection, repetitive behavior, and has hallucinations. Items that were more severe included items assessing the extent to which the child repeatedly rocks their head or body, has hallucinations, and plays with their own genitals in public. Items that were less severe included items assessing the extent to which the child avoids eye contact, has difficulty when things are out of place, and

does not answer when others talk to them.

Changes in Item Functioning With Age

Estimates of changes in item functioning with age are in Table S8. There were several general patterns. First, items tended to increase in easiness with age, suggesting that as children get older, for many of the items, a lower level of the construct is required for the item to be endorsed. For instance, an item assessing the extent to which the child has difficulty sitting still showed increases in easiness with age. Perhaps the increases in item easiness reflect that such behaviors become less developmentally appropriate with age. Consequently, caregivers and teachers may be more concerned or attentive to these behaviors as children age, and when the behavior occurs, it may be more likely to be perceived as a problem. In addition, situational demands increase across the transition from preschool to school entry; for instance, with greater classroom structure, the problematic behaviors may become more visible.

Second, items tended to decrease in discrimination with age, suggesting that some of the behaviors wane in construct relevance. For instance, an item assessing the extent to which the child screams a great deal showed decreases in discrimination with age. It is possible that screaming is a canonical form of dysregulated behavior in younger children, whereas as children get older, they may develop more advanced capacities for expressing their frustration such as arguing or sarcasm that replace the need for screaming.

Third, most of the behaviors showed changes in item functioning, consistent with the notion that psychopathology demonstrates heterotypic continuity. Items that do not show changes in item functioning (e.g., stealing, deviant peers, noncompliance at school) could be valuable anchor items for future longitudinal studies and models. The present study accounted for differential item functioning in the Bayesian item response model.

References

1. Kolen MJ, Brennan RL. *Test equating, scaling, and linking: Methods and practices*. 3rd ed. Statistics for Social and Behavioral Sciences. Springer; 2014:566.
2. Lee W-C, Lee G. IRT linking and equating. *The Wiley Handbook of Psychometric Testing*. 2018:639–673.
3. McArdle JJ, Grimm KJ, Hamagami F, Bowles RP, Meredith W. Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychol Methods*. 2009;14(2):126–149. doi: 10.1037/a0015857
4. Wang C, Nydick SW. On longitudinal item response theory models: A didactic. *J Educ Behav Stat*. 2020;45(3):339–368. doi: 10.3102/1076998619882026
5. Huang H-Y. A multilevel higher order item response theory model for measuring latent growth in longitudinal data. *Appl Psychol Meas*. 2015;39(5):362–372. doi: 10.1177/0146621614568112
6. Bürkner P-C. Bayesian item response modeling in R with brms and Stan. *J Stat Softw*. 2021;100(5):1–54. doi: 10.18637/jss.v100.i05
7. Caspi A, Houts RM, Belsky DW, et al. The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clin Psychol Sci*. 2014;2(2):119–137. doi: 10.1177/2167702613497473
8. Petersen IT, Bates JE, Dodge KA, Lansford JE, Pettit GS. Describing and predicting developmental profiles of externalizing problems from childhood to adulthood. *Dev Psychopathol*. 2015;27(3):791–818. doi: 10.1017/S0954579414000789
9. Bürkner P-C. Analysing standard progressive matrices (SPM-LS) with Bayesian item response models. *J Intell*. 2020;8(1):5. doi: 10.3390/jintelligence8010005
10. Bürkner P-C. brms: An R package for Bayesian multilevel models using Stan. *J Stat Softw*. 2017;80(1):28. doi: 10.18637/jss.v080.i01
11. Lewandowski D, Kurowicka D, Joe H. Generating random correlation matrices based on vines and extended onion method. *J Multivar Anal*. 2009;100(9):1989–2001. doi: 10.1016/j.jmva.2009.04.008
12. Detry MA, Ma Y. Analyzing repeated measurements using mixed models. *JAMA*. 2016;315(4):407–408. doi: 10.1001/jama.2015.19394
13. Twisk J, de Boer M, de Vente W, Heymans M. Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *J Clin Epidemiol*. 2013;66(9):1022–1028. doi: 10.1016/j.jclinepi.2013.03.017
14. *RStan: the R interface to Stan*. 2020. http://mc-stan.org
15. *Stan modeling language users guide and reference manual*. 2020. https://mc-stan.org
16. *Mplus version 8.6*. Muthén & Muthén; 2021. https://www.statmodel.com/programs.shtml
17. Aitken M, Plamondon A, Krzeczkowski J, Kil H, Andrade BF. Systematic integration of multi-informant externalizing ratings in clinical settings. *Res Child Adolesc Psychopathol*. 2023;doi: 10.1007/s10802-023-01119-z
18. Cowles MK. *Applied Bayesian statistics: With R and OpenBUGS examples*. Springer; 2013.
19. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian data analysis*. 3rd ed. Taylor & Francis; 2013.
20. Kruschke J. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Elsevier Science; 2014.

21. Wikle CK, Zammit-Mangion A, Cressie N. *Spatio-temporal statistics with R*. CRC Press; 2019.
22. Van De Schoot R, Kluytmans A, Tummers L, Lugtig P, Hox J, Muthen B. Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front Psychol*. 2013;4(770)doi: 10.3389/fpsyg.2013.00770
23. Van De Schoot R, Schmidt P, De Beuckelaer A, Lek K, Zondervan-Zwijnenburg M. Editorial: Measurement invariance. *Front Psychol*. 2015;6(1064)doi: 10.3389/fpsyg.2015.01064
24. Cieciuch J, Davidov E, Schmidt P, Algesheimer R, Schwartz SH. Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values. *Front Psychol*. 2014;5(982)doi: 10.3389/fpsyg.2014.00982

Table S1: Internal Consistency of Behavior Problem Ratings

	Measure							
	CBCL 1.5–5		CBCL 6–18		C–TRF		TRF	
Construct	α	$\omega_{hierarchical}$	α	$\omega_{hierarchical}$	α	$\omega_{hierarchical}$	α	$\omega_{hierarchical}$
Externalizing Problems	.91	.91	.87	.89	.94	.93	.93	–
Internalizing Problems	.82	.80	.80	.80	.84	.81	.78	.76
Thought-Disordered Problems	.62	.63	.50	.51	.70	.38	.47	.47

Note: All α = Cronbach's alpha; $\omega_{hierarchical}$ = omega hierarchical; "–" indicates that the omega coefficient was unable to be estimated due to a model convergence error.

Table S2: Cross-Rater Reliability (Pearson Correlation Coefficients)

		Rater		
	Externalizing			
		Mother	Father	Teacher
Rater	Mother	–		
	Father	.52	–	
	Teacher	.42	.40	–
	Internalizing			
		Mother	Father	Teacher
Rater	Mother	–		
	Father	.40	–	
	Teacher	.25	.18	–
	Thought-Disordered			
		Mother	Father	Teacher
Rater	Mother	–		
	Father	.32	–	
	Teacher	.39	.25	–

Note: All *p*s < .01.

Table S3: 9-Month Lag Cross-Time Rank-Order Stability (Pearson Correlation Coefficients)

Construct	Rater	Cross-Time Stability
Externalizing	Mother	.64
Externalizing	Father	.64
Externalizing	Teacher	.53
Internalizing	Mother	.67
Internalizing	Father	.67
Internalizing	Teacher	.36
Thought-Disordered	Mother	.53
Thought-Disordered	Father	.42
Thought-Disordered	Teacher	.40

Note: All *p*s < .001.

Table S4: Descriptive Statistics and Correlation Matrix of Study Variables

Variable	Age	Sex	SES	EXT (DS)	INT (DS)	TD (DS)	EXT (CV)	EXT (Common)	EXT (All)	Attention (obs)	Compliance (obs)
Age	–										
Sex	.06*	–									
SES	.14***	.06***	–								
EXT (DS)	−.35***	−.09***	−.17***	–							
INT (DS)	−.30***	.03	−.13***	.83***	–						
TD (DS)	−.36***	−.10***	−.16***	.82***	.94***	–					
EXT (CV)	−.26***	−.12***	−.15***	.72***	.56***	.54***	–				
EXT (Common)	−.08***	−.05†	−.12***	.66***	.50***	.46***	.89***	–			
EXT (All)	.02	−.08*	−.13***	.74***	.64***	.59***	.90***	.92***	–		
Attention (obs)	.50***	.14***	.24***	−.28***	−.19***	−.23***	−.24***	−.16***	−.11***	–	
Compliance (obs)	.49***	.12***	.21***	−.24***	−.16***	−.18***	−.22***	−.15***	−.09*	.91***	–
observations	1,857	2,772	2,748	1,328	1,322	1,290	1,328	1,327	981	1,557	1,557
M	4.96	0.47	−0.16	−0.77	−1.63	−2.05	0.15	0.14	0.15	3.93	4.08
SD	1.19	0.50	0.90	2.67	2.86	2.96	0.14	0.15	0.13	0.96	0.99
minimum	2.92	0.00	−3.18	−19.70	−21.71	−24.01	0.00	0.00	0.00	1.00	1.00
maximum	7.80	1.00	2.20	7.49	6.92	6.21	0.77	0.89	0.76	5.00	5.00

Note: "SES" = socioeconomic status (derived from averaging *z*-scores of the parents' educational attainment, the parent's occupational prestige, and of the log transform of the family's income-to-needs ratio); "EXT" = externalizing problems; "INT" = internalizing problems; "TD" = thought-disordered problems; "DS" = developmentally scaled; "CV" = construct-valid items (proportion score); "Common" = common items (proportion score); "All" = all possible items (from the upward/downward extension approach; proportion score); "obs" = observation by researchers. Data are in long form such that each row represents a unique combination of child, wave, and rater. Each participant could have up to 12 rows: 4 timepoints × 3 raters (mother, father, and teacher). Data for the child's sex and socioeconomic status at later timepoints are determined using the last observation carried forward, thus accounting for the larger number of observations. Only parents reported on all possible items (i.e., the upward/downward extension approach), whereas teachers (and parents) reported on the common items and construct-valid items, thus accounting for the relatively smaller number of observations for the externalizing score that leverages all possible items.

Table S5: Number of Participants who had Data by Number of Time Points

	Number of Time Points With Ratings				
Variable	0	1	2	3	4
Behavior Problems: Any Rater	1	49	62	42	77
Behavior Problems: Mother-Report	6	47	62	44	72
Behavior Problems: Father-Report	56	57	61	28	29
Behavior Problems: Teacher/Secondary Caregiver-Report	55	65	59	31	21
Behavior Observation	6	66	58	67	34

Table S6: Regression Coefficients from Bayesian Item Response Model

Parameter	Estimate	*SE*	Lower	Upper
intercept[1]	−1.58	2.26	−6.06	2.83
intercept[2]	1.59	2.26	−2.91	6.01
Predicting Theta				
age (centered)	0.23	0.22	−0.20	0.67
age (centered) quadratic	−0.33	0.09	−0.51	−0.16
mother	0.70	0.46	−0.20	1.62
father	0.43	0.52	−0.56	1.47
secondary caregiver	0.43	0.38	−0.26	1.24
internalizing	−1.34	0.51	−2.29	−0.30
thought disorder	−1.63	0.75	−3.08	−0.12
female	−0.33	0.21	−0.73	0.08
age (centered) × father	0.25	0.15	−0.04	0.55
age (centered) × secondary caregiver	0.19	0.19	−0.21	0.57
age (centered) × internalizing	0.05	0.20	−0.35	0.44
age (centered) × thought disorder	0.11	0.27	−0.44	0.62
father × internalizing	−0.27	0.29	−0.84	0.28
secondary caregiver × internalizing	0.71	0.34	0.02	1.37
father × thought disorder	−0.16	0.37	−0.90	0.56
secondary caregiver × thought disorder	1.30	0.48	0.36	2.25
female × internalizing	0.39	0.21	−0.02	0.79
female × thought disorder	0.44	0.27	−0.10	0.98
age (centered) × female	−0.34	0.17	−0.67	−0.01
age (centered) × father × internalizing	−0.07	0.16	−0.39	0.25
age (centered) × secondary caregiver × internalizing	0.25	0.19	−0.12	0.63
age (centered) × father × thought disorder	0.09	0.21	−0.33	0.51
age (centered) × secondary caregiver × thought disorder	0.12	0.25	−0.38	0.62
age (centered) × female × internalizing	0.05	0.13	−0.21	0.31
age (centered) × female × thought disorder	−0.39	0.17	−0.73	−0.05
Predicting Item Easiness				
intercept	−4.39	2.30	−9.03	0.20
age (centered)	0.41	0.17	0.09	0.75
age (centered) quadratic	−0.02	0.03	−0.07	0.03
father	0.34	0.24	−0.12	0.83
secondary caregiver	−0.58	0.41	−1.40	0.18
age (centered) × father	−0.13	0.08	−0.29	0.02
age (centered) × secondary caregiver	−0.08	0.18	−0.43	0.28

Predicting Item Discrimination				
intercept	0.35	0.04	0.28	0.43
age (centered)	0.00	0.00	0.00	0.01
age (centered) quadratic	0.00	0.00	0.00	0.00
father	0.01	0.01	0.00	0.03
secondary caregiver	0.43	0.10	0.23	0.63
internalizing	0.01	0.01	0.00	0.05
thought disorder	0.03	0.03	0.00	0.10
age (centered) × father	0.00	0.00	0.00	0.01
age (centered) × secondary caregiver	0.05	0.04	0.00	0.15

Note: Regression coefficients are unstandardized. “Lower” and “Upper” represent the bounds of the 95% credible interval. "intercept[1]" represents the intercept for the threshold from 0 to 1; "intercept[2]" represents the intercept for the threshold from 1 to 2. Mother-report is the reference group for the rater role (mother, father, teacher/secondary caregiver). Male is the reference group for the child’s sex. Externalizing is the reference group for the psychopathology dimension (externalizing, internalizing, thought disorder).

Table S7: Item Easiness and Discrimination by Age and Rater

						Easiness						Discrimination					
						Age 3			Age 7.5			Age 3			Age 7.5		
Scale	Construct	CBCL 1.5–5	C–TRF	CBCL 6–18	TRF	M	F	T	M	F	T	M	F	T	M	F	T
EXT	EXT	5	5			−2.87	−2.51	−3.21				1.00	0.94	1.44			
EXT	EXT	6	6			−2.44	−1.90	−2.99				1.04	1.01	1.53			
EXT	EXT	56	56			−4.34	−4.08	−4.69				0.46	0.45	0.75			
EXT	EXT	59	59			−1.50	−1.19	−2.05				0.76	0.73	1.11			
EXT	EXT	95	95			−4.48	−4.21	−4.84				0.88	0.87	1.35			
EXT	EXT	8	8			−1.47	−0.97	−2.61				1.04	1.04	1.65			
EXT	EXT	15	15		6	−2.68	−2.11	−3.56			−2.64	1.17	1.12	1.82			1.54
EXT	EXT	16	16		77	−2.04	−1.62	−3.14			−3.54	1.09	1.07	1.71			1.38
EXT	EXT	18	18	21	21	−4.58	−4.24	−5.25	−2.89	−3.17	−4.15	1.23	1.21	1.91	0.33	0.32	1.60
EXT	EXT	20	20			−2.33	−1.95	−3.17				0.97	0.94	1.63			
EXT	EXT	27	27	26	26	−3.35	−2.82	−3.69	−3.07	−3.09	−3.12	1.04	1.01	1.75	0.21	0.17	1.33
EXT	EXT	29	29			−2.47	−2.14	−3.16				1.12	1.07	1.69			
EXT	EXT	35	35	37	37	−5.64	−5.28	−5.98	−4.70	−4.94	−4.97	1.11	1.08	1.92	0.36	0.33	1.70
EXT	EXT	40	40			−2.94	−2.67	−3.67				1.01	0.96	1.57			
EXT	EXT	42	42			−4.86	−4.55	−5.52				1.05	0.99	1.64			
EXT	EXT	44	44			−3.56	−3.22	−4.20				1.14	1.11	1.80			
EXT	EXT	53	53	57	57	−5.47	−5.19	−6.03	−4.04	−4.36	−5.16	1.15	1.10	1.86	0.39	0.35	1.58
EXT	EXT	58	58			−3.52	−3.14	−4.11				1.23	1.22	1.92			
EXT	EXT	66	66	68	68	−3.73	−3.21	−4.52	−2.85	−2.87	−4.57	1.23	1.19	1.76	0.47	0.46	1.43
EXT	EXT	69	69			−2.74	−2.31	−3.24				0.77	0.77	1.45			
EXT	EXT	81	81	86	86	−3.02	−2.68	−3.55	−0.61	−0.89	−2.19	1.16	1.14	1.72	0.21	0.19	1.30
EXT	EXT	85	85	95	95	−2.65	−2.32	−3.84	−0.87	−1.14	−3.58	1.25	1.22	1.92	0.38	0.36	1.64
EXT	EXT	88	88			−3.26	−3.02	−4.02				1.13	1.09	1.67			
EXT	EXT	96	96			−1.98	−1.66	−2.81				0.93	0.92	1.21			

EXT	EXT			2				−6.02	−6.28				0.20	0.18	
EXT	EXT			28	28			−1.11	−1.36	−2.66			0.41	0.39	1.17
EXT	EXT			39	39			−3.99	−4.12	−4.35			0.29	0.29	0.87
EXT	EXT			43	43			−1.65	−1.84	−2.81			0.36	0.37	0.90
EXT	EXT			63	63			−1.80	−1.93	−3.74			0.20	0.23	0.40
EXT	EXT			67				−7.44	−7.75				0.34	0.36	
EXT	EXT			72				−7.41	−7.72				0.34	0.35	
EXT	EXT			73				−7.44	−7.75				0.34	0.35	
EXT	EXT			81				−4.33	−4.69				0.42	0.41	
EXT	EXT			82	82			−4.99	−5.24	−5.41			0.36	0.35	1.40
EXT	EXT			90	90			−3.87	−4.07	−4.83			0.46	0.48	1.18
EXT	EXT			96	96			−6.77	−7.04	−7.09			0.26	0.26	0.80
EXT	EXT			99	99			−7.68	−7.99	−7.70			0.34	0.36	0.90
EXT	EXT			101	101			−6.27	−6.59	−6.77			0.30	0.32	0.81
EXT	EXT			105	105			−7.66	−7.96	−7.67			0.34	0.36	0.90
EXT	EXT			106				−5.42	−5.72				0.35	0.39	
EXT	EXT			3	3			0.54	0.30	−1.60			0.32	0.32	1.14
EXT	EXT		22	16	16		−5.25	−3.09	−3.47	−3.25		1.80	0.31	0.30	1.53
EXT	EXT			19	19			0.15	−0.05	−1.65			0.45	0.45	1.14
EXT	EXT		17	20	20		−5.24	−2.71	−2.99	−4.27		1.27	0.35	0.35	1.37
EXT	EXT			22				−0.46	−0.70				0.36	0.38	
EXT	EXT			23	23			−2.75	−2.96	−2.95			0.24	0.21	1.10
EXT	EXT			87	87			−1.45	−1.79	−3.07			0.44	0.43	0.94
EXT	EXT			88	88			−2.98	−3.26	−3.07			0.35	0.35	0.94
EXT	EXT			89	89			−5.91	−6.22	−5.84			0.49	0.51	1.48
EXT	EXT		84	94	94		−6.02	−2.62	−3.07	−3.41		1.16	0.34	0.35	1.27
EXT	EXT			97	97			−4.39	−4.66	−5.77			0.39	0.41	0.90
EXT	EXT			104	104			−2.57	−2.75	−4.23			0.55	0.59	1.48
EXT	EXT		24				−2.90					1.38			

EXT	EXT			48				−3.41					1.05				
EXT	EXT			51				−3.77					0.83				
EXT	EXT			64				−3.20					1.16				
EXT	EXT			14				−6.62					0.75				
EXT	EXT			28				−4.02					1.41				
EXT	EXT			74				−6.31					1.05				
EXT	EXT					98					−4.98						0.83
EXT	EXT					76					−4.22						1.44
INT	INT	21	21				−2.69	−2.20	−3.62				0.56	0.60	0.98		
INT	INT	46	46				−6.37	−6.10	−6.57				0.59	0.61	1.03		
INT	INT	51					−5.95	−5.67					1.26	1.21			
INT	INT	79					−3.97	−3.71					0.99	0.97			
INT	INT	82	82				−3.13	−2.83	−3.87				1.00	0.98	1.55		
INT	INT	83	83				−4.50	−4.26	−4.76				0.64	0.65	1.14		
INT	INT	92	92				−4.49	−4.12	−5.26				0.86	0.83	1.27		
INT	INT	97	97				−0.97	−0.46	−2.39				0.50	0.49	1.04		
INT	INT	99	99	112	112		−3.96	−3.81	−4.59	−1.01	−1.53	−2.34	0.53	0.53	0.94	0.35	0.37	1.04
INT	INT	10	10				−1.73	−1.07	−2.90				0.64	0.64	1.04		
INT	INT	33	33				−1.94	−1.68	−2.38				0.51	0.52	0.90		
INT	INT	37	37				−2.67	−2.20	−3.42				0.70	0.71	1.14		
INT	INT	43	43				−5.50	−5.30	−5.69				0.61	0.60	1.14		
INT	INT	47	47	45	45		−4.75	−4.56	−5.27	−1.67	−2.12	−2.88	0.96	0.91	1.32	0.42	0.38	1.16
INT	INT	68	68	71	71		−3.28	−3.02	−3.80	−0.64	−1.03	−1.42	0.45	0.45	0.82	0.24	0.25	0.79
INT	INT	87	87	50	50		−4.64	−4.27	−5.42	−1.95	−2.17	−3.48	0.75	0.75	1.07	0.48	0.52	1.16
INT	INT	90	90	103	103		−5.99	−5.78	−6.25	−3.68	−4.10	−3.68	0.67	0.68	1.26	0.32	0.33	1.21
INT	INT	1	1	56a	56a		−4.05	−3.68	−4.88	−3.61	−3.82	−5.05	0.24	0.26	0.71	0.30	0.31	0.99
INT	INT	7	7				−2.59	−2.11	−3.45				0.44	0.46	0.73		
INT	INT	12		49			−3.00	−2.69		−2.45	−2.74		0.28	0.29		0.33	0.34	
INT	INT	19					−4.60	−4.22					0.38	0.37				

INT	INT		24					−2.52	−2.14					0.57	0.56		
INT	INT		39	39	56b	56b		−6.48	−6.30	−6.88	−2.93	−3.43	−3.60	0.43	0.43	0.83	0.13	0.13	0.66
INT	INT		45	45	56c	56c		−6.41	−6.12	−6.67	−3.73	−4.08	−3.84	0.44	0.46	0.95	0.48	0.51	1.26
INT	INT		52					−4.51	−4.19					0.18	0.22				
INT	INT		78	78	56f	56f		−5.25	−4.96	−5.81	−3.31	−3.66	−3.93	0.32	0.34	0.76	0.29	0.32	0.88
INT	INT		86	86				−3.79	−3.47	−4.59				0.48	0.45	0.76			
INT	INT		93	93	56g	56g		−6.17	−5.80	−6.89	−4.96	−5.18	−6.48	0.28	0.29	0.63	0.45	0.48	0.96
INT	INT		2	2				−3.57	−3.10	−3.91				0.52	0.50	1.07			
INT	INT		4	4				−3.16	−2.69	−3.64				0.53	0.52	1.11			
INT	INT		23	23				−1.88	−1.42	−2.43				0.67	0.62	1.30			
INT	INT		62	62				−4.59	−4.16	−5.55				0.75	0.72	1.28			
INT	INT		67	67				−6.74	−6.47	−7.00				1.17	1.13	1.88			
INT	INT		70	70				−5.42	−5.05	−5.52				0.71	0.71	1.41			
INT	INT		71	71				−5.53	−5.17	−5.85				0.64	0.63	1.28			
INT	INT		98	98	111	111		−5.31	−4.94	−5.67	−4.36	−4.58	−4.78	0.93	0.92	1.58	0.37	0.36	1.42
INT	INT				14	14					−1.59	−1.71	−2.78				0.33	0.36	0.94
INT	INT				29	29					−2.06	−2.21	−4.62				0.25	0.26	0.70
INT	INT				30	30					−3.66	−3.92	−4.67				0.24	0.23	0.91
INT	INT				31	31					−3.15	−3.26	−4.21				0.39	0.41	1.19
INT	INT				32	32					−1.19	−1.50	−1.80				0.36	0.36	0.69
INT	INT				33	33					−2.48	−2.68	−4.29				0.10	0.05	1.03
INT	INT				35	35					−3.64	−3.87	−5.62				0.38	0.38	1.08
INT	INT				52	52					−3.97	−4.13	−5.95				0.45	0.48	1.09
INT	INT				91	91					−5.15	−5.40	−6.35				0.42	0.43	1.14
INT	INT				5	5					−4.33	−4.73	−3.38				0.21	0.18	1.17
INT	INT				42	42					−2.69	−2.85	−3.22				0.26	0.25	0.76
INT	INT				65	65					−3.73	−3.87	−4.90				0.26	0.27	0.82
INT	INT				69	69					−3.06	−3.37	−3.71				0.29	0.30	0.85
INT	INT				75	75					−1.32	−1.52	−2.57				0.19	0.23	0.55

INT	INT			102	102				−4.32	−4.68	−3.83			0.31	0.33	0.58
INT	INT			47					−1.36	−1.67				0.22	0.25	
INT	INT			51	51				−5.16	−5.42	−6.17			0.39	0.43	0.78
INT	INT			54	54				−4.28	−4.64	−4.68			0.35	0.37	1.01
INT	INT			56d	56d				−6.13	−6.34	−6.77			0.31	0.32	0.80
INT	INT			56e	56e				−2.67	−2.85	−4.48			0.24	0.20	0.41
INT	INT		12				−5.24					1.02				
INT	INT		19				−3.32					0.70				
INT	INT				81					−2.03						0.63
INT	INT				106					−3.86						0.57
INT	INT				108					−2.20						0.63
DSM	TD	4	4			−3.07	−2.63	−3.68				0.61	0.60	1.13		
DSM	TD	7	7			−2.48	−2.04	−3.42				0.48	0.50	0.64		
DSM	TD	21	21			−2.56	−2.13	−3.56				0.57	0.59	0.83		
DSM	TD	23	23			−1.74	−1.35	−2.42				0.71	0.66	1.21		
DSM	TD	25	25			−4.08	−3.81	−4.48				0.58	0.56	1.06		
DSM	TD	63	63			−6.44	−6.14	−6.81				0.66	0.67	1.23		
DSM	TD	67	67			−6.78	−6.50	−7.21				1.23	1.19	1.90		
DSM	TD	70	70			−5.45	−5.08	−5.69				0.85	0.84	1.49		
DSM	TD	76	76			−3.36	−3.04	−3.73				0.39	0.41	0.96		
DSM	TD	80	80	84	84	−6.63	−6.40	−6.75	−5.09	−5.48	−4.31	0.87	0.85	1.24	0.39	0.39	1.05
DSM	TD	92	92			−4.28	−3.93	−5.13				0.78	0.76	1.05		
DSM	TD	98	98			−5.36	−5.02	−5.83				0.95	0.94	1.47		
TP	TD			9	9				−0.68	−0.97	−1.74			0.54	0.56	1.12
TP	TD			18	18				−6.33	−6.64	−6.81			0.37	0.39	0.93
TP	TD			40	40				−7.29	−7.58	−7.46			0.39	0.41	0.96
TP	TD			46	46				−4.23	−4.47	−5.22			0.40	0.42	0.78
TP	TD			58	58				−1.51	−1.62	−3.40			0.18	0.20	0.54
TP	TD			59					−5.33	−5.53				0.39	0.40	

TP	TD	60					−4.32	−4.56				0.17	0.20	
TP	TD	66	66				−3.75	−4.00	−4.21			0.43	0.44	1.16
TP	TD	70	70				−5.81	−6.11	−6.54			0.40	0.42	0.97
TP	TD	76					−2.01	−2.36				0.28	0.32	
TP	TD	83	83				−2.58	−2.79	−4.96			0.34	0.34	0.86
TP	TD	85	85				−5.74	−5.99	−6.21			0.38	0.39	1.11
TP	TD	92					−3.56	−3.95				0.15	0.17	
TP	TD	100					−2.85	−3.29				0.33	0.36	

Note: “EXT” = Externalizing; “INT” = Internalizing; “TD” = Thought Disorder; “DSM” = DSM Autism Spectrum; “TP” = Thought Problems. Yellow highlighted cells represent items that span across the age-differing measures. Gray highlighted cells indicates items that were not included in the Bayesian item response model for that rater at a given age (based on the items that are intended to be assessed by each rater on the Achenbach System of Empirically Based Assessment), consistent with the construct-valid items approach. Some of the items are included in multiple scales; for instance, item 46 of the CBCL 6–18 is part of the Thought Problems (Thought Disorder) subscale, whereas it is part of the Internalizing scale of the CBCL 1.5–5 and C–TRF. In addition, items 4, 7, 21, 23, 67, 70, 92, 98 of the of the CBCL 1.5–5 and C–TRF are in the DSM Autism Spectrum (Thought Disorder) subscale, whereas they are also part of the Internalizing scale of the CBCL 1.5–5 and C–TRF. For these items, they were allowed to load onto each of the relevant factors at the relevant ages as part of the Bayesian item response model. For example, item 4 was allowed to load onto both the Thought Disorder and Internalizing factors at ages 3–5.

Table S8: Estimates of Differential Item Functioning by Age

		Slope of Easiness				Slope of Discrimination			
Construct	Item Number	Estimate	Error (*SD*)	Lower	Upper	Estimate	Error (*SD*)	Lower	Upper
Externalizing	CBCL 1.5–5: 15	**0.58**	0.20	0.19	0.97	**−0.18**	0.05	−0.28	−0.09
Externalizing	CBCL 1.5–5: 16	**0.47**	0.19	0.10	0.85	**−0.17**	0.05	−0.26	−0.07
Externalizing	CBCL 1.5–5: 18	**0.47**	0.20	0.08	0.85	**−0.20**	0.04	−0.29	−0.12
Externalizing	CBCL 1.5–5: 20	**0.49**	0.19	0.13	0.87	**−0.15**	0.05	−0.24	−0.07
Externalizing	CBCL 1.5–5: 27	0.15	0.18	−0.20	0.52	**−0.19**	0.04	−0.26	−0.11
Externalizing	CBCL 1.5–5: 29	**0.75**	0.20	0.37	1.16	**−0.18**	0.05	−0.28	−0.09
Externalizing	CBCL 1.5–5: 35	0.30	0.21	−0.10	0.72	**−0.17**	0.05	−0.27	−0.06
Externalizing	CBCL 1.5–5: 40	0.25	0.20	−0.12	0.65	**−0.16**	0.05	−0.26	−0.07
Externalizing	CBCL 1.5–5: 42	0.29	0.22	−0.13	0.73	**−0.17**	0.06	−0.28	−0.06
Externalizing	CBCL 1.5–5: 44	**0.58**	0.21	0.18	1.00	**−0.19**	0.05	−0.29	−0.09
Externalizing	CBCL 1.5–5: 5	**0.56**	0.19	0.19	0.97	**−0.17**	0.05	−0.27	−0.08
Externalizing	CBCL 1.5–5: 53	**0.41**	0.21	0.01	0.82	**−0.17**	0.05	−0.27	−0.07
Externalizing	CBCL 1.5–5: 56	0.36	0.19	0.00	0.73	−0.04	0.05	−0.15	0.06
Externalizing	CBCL 1.5–5: 58	**0.41**	0.21	0.00	0.84	**−0.20**	0.05	−0.31	−0.10
Externalizing	CBCL 1.5–5: 59	0.24	0.17	−0.09	0.58	**−0.11**	0.04	−0.20	−0.03
Externalizing	CBCL 1.5–5: 6	**0.71**	0.20	0.33	1.11	**−0.19**	0.05	−0.29	−0.10
Externalizing	CBCL 1.5–5: 66	0.29	0.20	−0.09	0.69	**−0.17**	0.04	−0.25	−0.09
Externalizing	CBCL 1.5–5: 69	0.30	0.18	−0.04	0.65	**−0.09**	0.05	−0.19	0.00
Externalizing	CBCL 1.5–5: 8	**0.68**	0.19	0.31	1.06	**−0.17**	0.05	−0.26	−0.08
Externalizing	CBCL 1.5–5: 81	**0.63**	0.19	0.27	1.00	**−0.21**	0.03	−0.27	−0.15
Externalizing	CBCL 1.5–5: 85	**0.49**	0.19	0.12	0.86	**−0.20**	0.03	−0.26	−0.13
Externalizing	CBCL 1.5–5: 88	**0.54**	0.21	0.15	0.95	**−0.17**	0.05	−0.27	−0.07
Externalizing	CBCL 1.5–5: 95	0.13	0.21	−0.28	0.57	**−0.11**	0.06	−0.22	−0.01
Externalizing	CBCL 1.5–5: 96	**0.53**	0.18	0.18	0.89	**−0.15**	0.05	−0.24	−0.05
Externalizing	CBCL 6–18: 101	0.23	0.29	−0.35	0.82	0.02	0.11	−0.20	0.22

Externalizing	CBCL 6–18: 104	0.47	0.26	−0.03	0.98	0.00	0.09	−0.17	0.18
Externalizing	CBCL 6–18: 105	0.12	0.31	−0.50	0.72	0.01	0.11	−0.20	0.23
Externalizing	CBCL 6–18: 16	**0.51**	0.23	0.08	0.96	**−0.16**	0.07	−0.29	−0.03
Externalizing	CBCL 6–18: 19	**0.72**	0.24	0.26	1.22	−0.07	0.07	−0.20	0.06
Externalizing	CBCL 6–18: 20	**0.48**	0.22	0.06	0.91	−0.07	0.06	−0.20	0.05
Externalizing	CBCL 6–18: 23	0.39	0.25	−0.10	0.88	−0.08	0.07	−0.22	0.07
Externalizing	CBCL 6–18: 28	**0.54**	0.24	0.07	1.01	−0.03	0.07	−0.16	0.10
Externalizing	CBCL 6–18: 3	**0.63**	0.23	0.18	1.10	−0.09	0.06	−0.20	0.02
Externalizing	CBCL 6–18: 39	0.34	0.26	−0.16	0.87	−0.03	0.09	−0.21	0.14
Externalizing	CBCL 6–18: 43	**0.60**	0.24	0.13	1.09	0.03	0.07	−0.10	0.17
Externalizing	CBCL 6–18: 63	0.33	0.24	−0.14	0.80	−0.03	0.06	−0.15	0.10
Externalizing	CBCL 6–18: 82	0.26	0.27	−0.27	0.80	−0.11	0.10	−0.31	0.08
Externalizing	CBCL 6–18: 87	**0.65**	0.24	0.18	1.13	−0.01	0.07	−0.15	0.14
Externalizing	CBCL 6–18: 88	0.42	0.25	−0.05	0.92	−0.01	0.08	−0.17	0.15
Externalizing	CBCL 6–18: 89	0.30	0.29	−0.27	0.86	−0.02	0.11	−0.23	0.19
Externalizing	CBCL 6–18: 90	0.27	0.26	−0.24	0.79	0.00	0.09	−0.17	0.18
Externalizing	CBCL 6–18: 94	**0.74**	0.23	0.30	1.20	−0.05	0.07	−0.18	0.08
Externalizing	CBCL 6–18: 96	0.18	0.30	−0.42	0.77	−0.01	0.11	−0.23	0.21
Externalizing	CBCL 6–18: 97	0.40	0.28	−0.14	0.96	0.02	0.10	−0.17	0.21
Externalizing	CBCL 6–18: 99	0.12	0.31	−0.49	0.74	0.01	0.11	−0.21	0.24
Externalizing	CBCL 6–18: 106	0.33	0.29	−0.22	0.90	0.08	0.11	−0.13	0.29
Externalizing	CBCL 6–18: 2	0.27	0.30	−0.31	0.86	−0.05	0.11	−0.27	0.17
Externalizing	CBCL 6–18: 22	**0.55**	0.24	0.09	1.03	−0.03	0.06	−0.16	0.09
Externalizing	CBCL 6–18: 67	0.14	0.31	−0.47	0.77	0.00	0.12	−0.24	0.23
Externalizing	CBCL 6–18: 72	0.15	0.31	−0.45	0.75	0.00	0.11	−0.23	0.22
Externalizing	CBCL 6–18: 73	0.15	0.30	−0.44	0.74	0.00	0.11	−0.22	0.22
Externalizing	CBCL 6–18: 81	0.43	0.28	−0.10	1.00	0.01	0.10	−0.19	0.21
Externalizing	C–TRF: 14	0.22	0.31	−0.38	0.83	0.00	0.09	−0.18	0.18
Externalizing	C–TRF: 24	**0.59**	0.26	0.10	1.09	−0.09	0.07	−0.23	0.06

Externalizing	C–TRF: 28	**0.62**	0.26	0.12	1.14	−0.09	0.08	−0.23	0.07
Externalizing	C–TRF: 48	**0.62**	0.25	0.14	1.14	−0.04	0.07	−0.17	0.11
Externalizing	C–TRF: 51	**0.80**	0.26	0.29	1.33	−0.01	0.07	−0.15	0.13
Externalizing	C–TRF: 64	**0.71**	0.25	0.21	1.20	−0.08	0.07	−0.22	0.06
Externalizing	C–TRF: 74	0.42	0.29	−0.14	1.00	−0.03	0.09	−0.20	0.14
Externalizing	TRF: 76	0.38	0.28	−0.16	0.94	−0.09	0.11	−0.30	0.13
Externalizing	TRF: 98	0.34	0.29	−0.24	0.91	0.03	0.11	−0.19	0.24
Internalizing	CBCL 1.5–5: 1	0.19	0.17	−0.15	0.53	0.01	0.04	−0.06	0.09
Internalizing	CBCL 1.5–5: 10	0.28	0.20	−0.09	0.67	−0.05	0.04	−0.14	0.03
Internalizing	CBCL 1.5–5: 12	0.22	0.17	−0.11	0.55	0.01	0.03	−0.05	0.07
Internalizing	CBCL 1.5–5: 2	**0.52**	0.19	0.15	0.91	−0.05	0.05	−0.14	0.03
Internalizing	CBCL 1.5–5: 21	**0.43**	0.20	0.06	0.83	−0.01	0.05	−0.10	0.08
Internalizing	CBCL 1.5–5: 23	0.22	0.20	−0.15	0.61	−0.08	0.04	−0.17	0.00
Internalizing	CBCL 1.5–5: 33	**0.71**	0.18	0.36	1.08	−0.03	0.04	−0.11	0.04
Internalizing	CBCL 1.5–5: 37	0.00	0.21	−0.41	0.41	**−0.10**	0.05	−0.20	−0.02
Internalizing	CBCL 1.5–5: 39	**0.88**	0.21	0.47	1.31	−0.07	0.05	−0.16	0.03
Internalizing	CBCL 1.5–5: 4	**0.49**	0.19	0.10	0.87	−0.03	0.05	−0.12	0.06
Internalizing	CBCL 1.5–5: 43	**0.59**	0.25	0.13	1.10	−0.05	0.06	−0.17	0.06
Internalizing	CBCL 1.5–5: 45	**0.69**	0.23	0.26	1.14	0.01	0.06	−0.10	0.12
Internalizing	CBCL 1.5–5: 46	**0.65**	0.27	0.14	1.19	−0.03	0.07	−0.16	0.10
Internalizing	CBCL 1.5–5: 47	**0.78**	0.22	0.37	1.21	**−0.12**	0.04	−0.20	−0.04
Internalizing	CBCL 1.5–5: 62	0.41	0.23	−0.03	0.87	**−0.11**	0.05	−0.23	−0.01
Internalizing	CBCL 1.5–5: 67	0.12	0.30	−0.46	0.70	**−0.20**	0.07	−0.34	−0.06
Internalizing	CBCL 1.5–5: 68	**0.68**	0.17	0.35	1.01	−0.05	0.03	−0.10	0.00
Internalizing	CBCL 1.5–5: 7	0.20	0.18	−0.15	0.57	0.00	0.04	−0.08	0.09
Internalizing	CBCL 1.5–5: 70	**0.54**	0.25	0.06	1.05	−0.09	0.06	−0.21	0.03
Internalizing	CBCL 1.5–5: 71	0.33	0.25	−0.15	0.83	−0.07	0.06	−0.19	0.05
Internalizing	CBCL 1.5–5: 78	**0.52**	0.19	0.15	0.91	−0.01	0.04	−0.10	0.08
Internalizing	CBCL 1.5–5: 82	0.44	0.24	−0.01	0.92	**−0.15**	0.05	−0.25	−0.05

Internalizing	CBCL 1.5–5: 83	**0.45**	0.22	0.02	0.88	−0.09	0.05	−0.20	0.01
Internalizing	CBCL 1.5–5: 86	0.10	0.20	−0.28	0.49	−0.05	0.05	−0.15	0.05
Internalizing	CBCL 1.5–5: 87	**0.69**	0.21	0.29	1.11	−0.06	0.04	−0.14	0.02
Internalizing	CBCL 1.5–5: 90	**0.61**	0.22	0.19	1.06	−0.08	0.05	−0.19	0.02
Internalizing	CBCL 1.5–5: 92	0.13	0.25	−0.33	0.63	**−0.12**	0.06	−0.23	−0.01
Internalizing	CBCL 1.5–5: 93	0.36	0.24	−0.08	0.84	0.04	0.06	−0.08	0.15
Internalizing	CBCL 1.5–5: 97	0.22	0.18	−0.12	0.56	−0.03	0.04	−0.10	0.04
Internalizing	CBCL 1.5–5: 98	0.30	0.22	−0.14	0.75	**−0.13**	0.05	−0.23	−0.02
Internalizing	CBCL 1.5–5: 99	**0.75**	0.18	0.39	1.11	−0.04	0.03	−0.10	0.02
Internalizing	CBCL 6–18: 102	0.26	0.27	−0.28	0.78	0.05	0.09	−0.14	0.24
Internalizing	CBCL 6–18: 14	**0.50**	0.25	0.03	1.01	0.01	0.07	−0.12	0.14
Internalizing	CBCL 6–18: 29	0.32	0.25	−0.16	0.82	−0.01	0.07	−0.15	0.12
Internalizing	CBCL 6–18: 30	0.36	0.26	−0.15	0.90	−0.01	0.08	−0.17	0.16
Internalizing	CBCL 6–18: 31	0.27	0.26	−0.22	0.79	−0.05	0.09	−0.22	0.12
Internalizing	CBCL 6–18: 32	**0.54**	0.25	0.07	1.03	−0.06	0.08	−0.21	0.08
Internalizing	CBCL 6–18: 33	**0.52**	0.26	0.03	1.03	**−0.20**	0.07	−0.33	−0.06
Internalizing	CBCL 6–18: 35	0.36	0.27	−0.16	0.90	−0.07	0.09	−0.26	0.12
Internalizing	CBCL 6–18: 42	0.36	0.25	−0.11	0.85	0.01	0.07	−0.13	0.15
Internalizing	CBCL 6–18: 5	**0.58**	0.28	0.03	1.13	−0.11	0.10	−0.32	0.09
Internalizing	CBCL 6–18: 51	0.27	0.28	−0.25	0.85	0.10	0.10	−0.10	0.30
Internalizing	CBCL 6–18: 52	0.15	0.28	−0.39	0.71	−0.02	0.10	−0.21	0.17
Internalizing	CBCL 6–18: 54	0.44	0.28	−0.09	0.99	0.00	0.10	−0.19	0.18
Internalizing	CBCL 6–18: 56d	0.17	0.30	−0.41	0.75	0.03	0.11	−0.18	0.25
Internalizing	CBCL 6–18: 56e	0.36	0.26	−0.13	0.87	0.03	0.07	−0.11	0.17
Internalizing	CBCL 6–18: 65	0.37	0.27	−0.15	0.89	−0.03	0.09	−0.21	0.14
Internalizing	CBCL 6–18: 69	0.50	0.26	−0.01	1.03	−0.03	0.08	−0.20	0.13
Internalizing	CBCL 6–18: 75	**0.54**	0.24	0.07	1.02	−0.02	0.06	−0.13	0.09
Internalizing	CBCL 6–18: 91	0.20	0.28	−0.34	0.76	−0.05	0.10	−0.26	0.16
Internalizing	CBCL 1.5–5: 19	0.00	0.23	−0.45	0.47	−0.02	0.06	−0.13	0.09

Internalizing	CBCL 1.5–5: 24	0.13	0.19	−0.24	0.52	−0.06	0.04	−0.15	0.03
Internalizing	CBCL 1.5–5: 51	0.48	0.29	−0.07	1.05	**−0.21**	0.07	−0.36	−0.08
Internalizing	CBCL 1.5–5: 52	0.36	0.21	−0.05	0.77	0.05	0.05	−0.05	0.16
Internalizing	CBCL 1.5–5: 79	0.36	0.25	−0.12	0.86	**−0.13**	0.06	−0.25	−0.02
Internalizing	CBCL 6–18: 47	**0.52**	0.25	0.03	1.01	0.01	0.06	−0.10	0.13
Internalizing	C–TRF: 12	**0.65**	0.28	0.11	1.19	−0.02	0.08	−0.18	0.14
Internalizing	C–TRF: 19	**0.65**	0.25	0.15	1.15	0.02	0.07	−0.12	0.16
Internalizing	TRF: 106	0.38	0.28	−0.17	0.93	0.07	0.11	−0.14	0.27
Internalizing	TRF: 108	**0.55**	0.27	0.02	1.09	0.06	0.10	−0.14	0.26
Internalizing	TRF: 81	0.51	0.27	−0.01	1.05	0.09	0.10	−0.11	0.29
Thought Disordered	CBCL 1.5–5: 21	**0.47**	0.20	0.07	0.86	−0.02	0.05	−0.12	0.07
Thought Disordered	CBCL 1.5–5: 23	0.29	0.21	−0.12	0.71	−0.08	0.04	−0.17	0.01
Thought Disordered	CBCL 1.5–5: 25	**0.42**	0.21	0.01	0.85	−0.07	0.05	−0.18	0.03
Thought Disordered	CBCL 1.5–5: 4	**0.55**	0.21	0.14	0.96	−0.05	0.05	−0.14	0.04
Thought Disordered	CBCL 1.5–5: 63	0.51	0.27	−0.02	1.06	−0.05	0.07	−0.19	0.10
Thought Disordered	CBCL 1.5–5: 67	0.11	0.30	−0.47	0.70	**−0.20**	0.07	−0.35	−0.06
Thought Disordered	CBCL 1.5–5: 7	0.25	0.20	−0.13	0.63	0.00	0.05	−0.10	0.09
Thought Disordered	CBCL 1.5–5: 70	**0.59**	0.26	0.11	1.10	−0.12	0.06	−0.25	0.00
Thought Disordered	CBCL 1.5–5: 76	0.29	0.19	−0.08	0.65	−0.03	0.05	−0.13	0.06
Thought Disordered	CBCL 1.5–5: 80	0.44	0.25	−0.04	0.94	−0.11	0.07	−0.24	0.02
Thought Disordered	CBCL 1.5–5: 92	0.12	0.24	−0.35	0.60	−0.09	0.06	−0.21	0.03
Thought Disordered	CBCL 1.5–5: 98	**0.55**	0.26	0.04	1.06	**−0.13**	0.06	−0.26	−0.01
Thought Disordered	CBCL 6–18: 18	0.23	0.30	−0.36	0.84	0.01	0.11	−0.21	0.22
Thought Disordered	CBCL 6–18: 40	0.14	0.31	−0.46	0.75	0.01	0.11	−0.20	0.24
Thought Disordered	CBCL 6–18: 46	0.38	0.28	−0.18	0.94	0.04	0.10	−0.16	0.24
Thought Disordered	CBCL 6–18: 58	0.31	0.25	−0.19	0.81	0.00	0.06	−0.11	0.11
Thought Disordered	CBCL 6–18: 66	0.37	0.28	−0.18	0.91	0.00	0.10	−0.19	0.18
Thought Disordered	CBCL 6–18: 70	0.27	0.29	−0.29	0.83	0.00	0.11	−0.21	0.22
Thought Disordered	CBCL 6–18: 83	**0.54**	0.28	0.02	1.10	−0.02	0.09	−0.19	0.18

Thought Disordered CBCL 6–18: 85	0.16	0.29	−0.41	0.74	−0.03	0.11	−0.25	0.19
Thought Disordered CBCL 6–18: 9	**0.83**	0.28	0.30	1.39	0.10	0.08	−0.05	0.28
Thought Disordered CBCL 6–18: 100	0.42	0.27	−0.11	0.97	−0.01	0.09	−0.20	0.17
Thought Disordered CBCL 6–18: 59	0.22	0.29	−0.35	0.81	0.00	0.11	−0.21	0.21
Thought Disordered CBCL 6–18: 60	0.37	0.28	−0.17	0.94	−0.02	0.10	−0.21	0.18
Thought Disordered CBCL 6–18: 76	**0.55**	0.26	0.05	1.10	0.03	0.07	−0.12	0.18
Thought Disordered CBCL 6–18: 92	0.41	0.27	−0.10	0.97	−0.05	0.09	−0.22	0.12

Note: Bolded coefficients represent significant differential item functioning across ages. "*SD*" = standard deviation.

Table S9: Model Fit of Latent Class Growth Analysis Models by Number of Classes

Externalizing Problems

# Classes	LL	AIC	BIC	SABIC	Entropy	LMR-LRT *p*	BLRT *p*	Smallest Class (%)
1	model did not converge				–	–	–	–
2	−2637.953	5373.906	5542.372	5387.071	0.827	.032	< .001	20.8
3	−2508.292	5122.583	5304.801	5136.823	0.815	.181	< .001	14.0
4	−2464.480	5042.961	5238.931	5058.276	0.856	.240	< .001	1.2

Internalizing Problems

# Classes	LL	AIC	BIC	SABIC	Entropy	LMR-LRT *p*	BLRT *p*	Smallest Class (%)
1	model did not converge				–	–	–	–
2	−2784.824	5667.648	5836.114	5680.813	0.679	.240	< .001	36.2
3	−2730.002	5566.004	5748.222	5580.244	0.727	.240	< .001	5.0
4	−2693.881	5501.763	5697.733	5517.078	0.717	.240	< .001	4.7

Thought-Disordered Problems

# Classes	LL	AIC	BIC	SABIC	Entropy	LMR-LRT *p*	BLRT *p*	Smallest Class (%)
1	model did not converge				–	–	–	–
2	−2720.562	5539.125	5707.590	5552.290	0.683	.240	< .001	46.1
3	−2633.318	5372.637	5554.855	5386.877	0.769	< .001	< .001	9.1
4	−2607.703	5329.406	5525.376	5344.721	0.701	.216	< .001	8.7

Note: "LMR-LRT" = Lo-Mendell-Rubin adjusted likelihood ratio test; "BLRT" = bootstrapped likelihood ratio test.

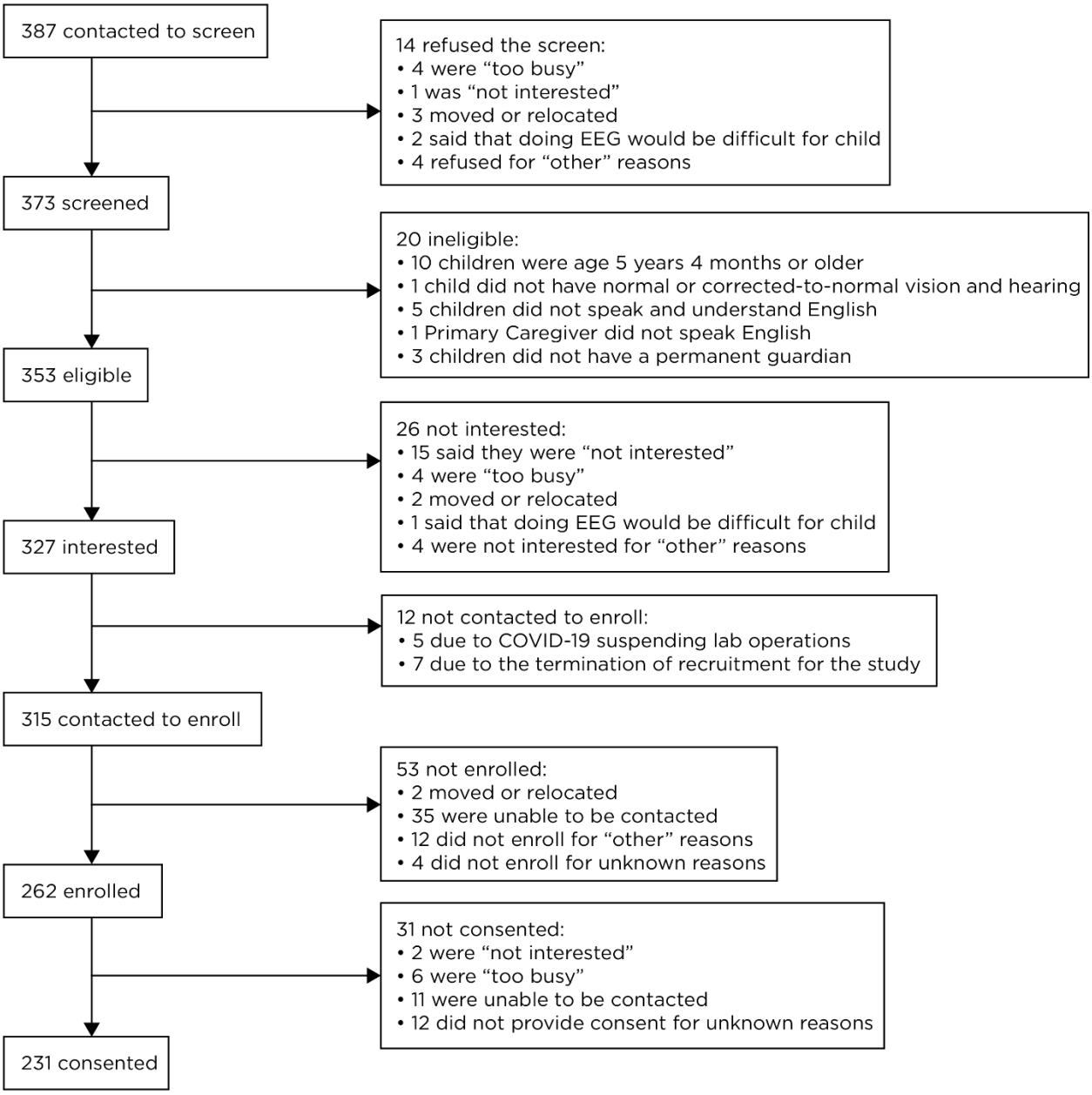
Table S10: Fisher's r-to-z Tests of Comparative Criterion-Related Validity

Correlation 1			Correlation 2					
Variable 1	Variable 2	*r*	Variable 1	Variable 2	*r*	r_{diff}	*z*	*p*
Attention to Task	Construct-Valid Items	−.24	Attention to Task	Common Items	−.16	.08	−5.54	**< .001**
Attention to Task	Construct-Valid Items (Parent-Report Only)	−.22	Attention to Task	All Possible Items (Upward/Downward Extension)	−.12	.10	−7.18	**< .001**
Attention to Task	Construct-Valid Items	−.24	Attention to Task	*T*-Scores (Age and Sex Norm-Referenced)	−.11	.13	−6.39	**< .001**
Attention to Task	Construct-Valid Items	−.24	Attention to Task	*z*-Scores (Normed Within Age)	−.12	.12	−10.30	**< .001**
Attention to Task	Developmentally Scaled Construct-Valid Items	−.28	Attention to Task	Common Items	−.16	.12	−5.22	**< .001**
Attention to Task	Developmentally Scaled Construct-Valid Items (Parent-Report Only)	−.26	Attention to Task	All Possible Items (Upward/Downward Extension)	−.12	.14	−5.76	**< .001**
Attention to Task	Developmentally Scaled Construct-Valid Items	−.28	Attention to Task	*T*-Scores (Age and Sex Norm-Referenced)	−.11	.17	−7.88	**< .001**
Attention to Task	Developmentally Scaled Construct-Valid Items	−.28	Attention to Task	*z*-Scores (Normed Within Age)	−.12	.16	−7.18	**< .001**
Attention to Task	Developmentally Scaled Construct-Valid Items	−.28	Attention to Task	Construct-Valid Items	−.24	.04	−2.25	**.025**
Compliance	Construct-Valid Items	−.22	Compliance	Common Items	−.15	.07	−5.17	**< .001**
Compliance	Construct-Valid Items (Parent-Report Only)	−.20	Compliance	All Possible Items (Upward/Downward Extension)	−.09	.11	−7.50	**< .001**
Compliance	Construct-Valid Items	−.22	Compliance	*T*-Scores (Age and Sex Norm-Referenced)	−.09	.13	−6.78	**< .001**
Compliance	Construct-Valid Items	−.22	Compliance	*z*-Scores (Normed Within Age)	−.11	.11	−10.11	**< .001**
Compliance	Developmentally Scaled Construct-Valid Items	−.25	Compliance	Common Items	−.15	.10	−4.11	**< .001**
Compliance	Developmentally Scaled Construct-Valid Items (Parent-Report Only)	−.25	Compliance	All Possible Items (Upward/Downward Extension)	−.09	.16	−5.39	**< .001**

Developmentally Scaled Construct-Valid Compliance	Items	−.25	Compliance	*T*-Scores (Age and Sex Norm-Referenced)	−.09	.16	−7.26	**< .001**
Developmentally Scaled Construct-Valid Compliance	Items	−.25	Compliance	*z*-Scores (Normed Within Age)	−.11	.14	−6.16	**< .001**
Developmentally Scaled Construct-Valid Compliance	Items	−.25	Compliance	Construct-Valid Items	−.22	.03	−1.29	.199

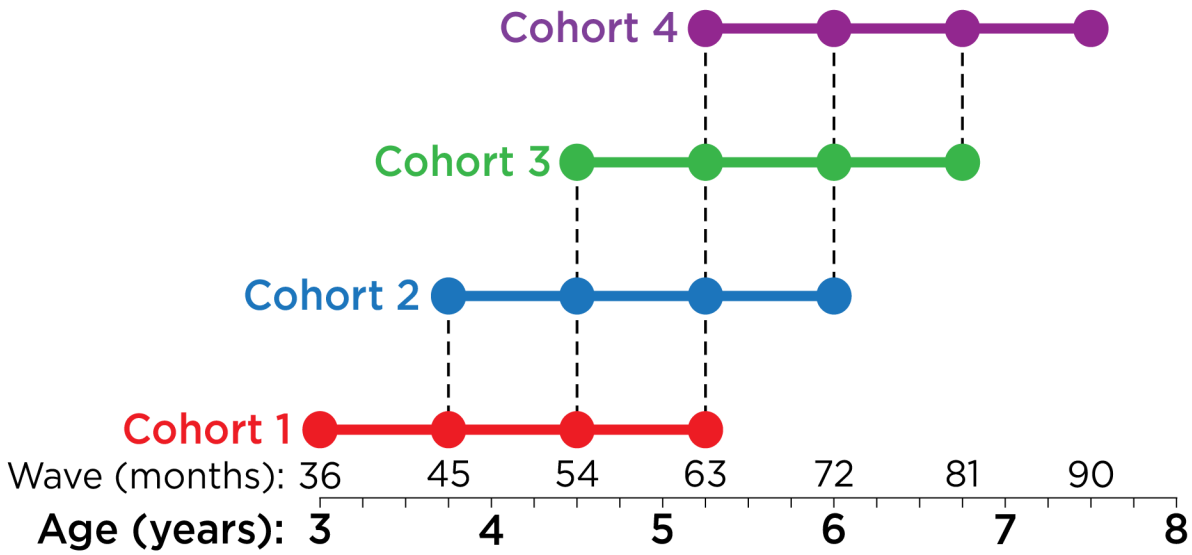
Note: Bolded *p*-values represent significant differences between correlation 1 and correlation 2 (based on Fisher's *r*-to-*z* test).

Figure S1: Participant Flow Chart



Note: "EEG" = electroencephalography.

Figure S2: Accelerated Longitudinal Design



Note: Accelerated longitudinal research design with four cohorts. The longitudinal design follows any given child for 2¼ years, with testing every nine months; the whole data set spans the ages of 3–7½ years. Circles reflect measurement points (four waves) for each cohort. Dashed lines indicate common measurement points across cohorts.

Figure S3: Latent Class Growth Analysis Trajectories