Development and Psychopathology (2025), **37**, 836–858
doi:10.1017/S0954579424000713

CAMBRIDGE UNIVERSITY PRESS

Regular Article

Reactive and control processes in the development of internalizing and externalizing problems across early childhood to adolescence

Jordan L. Harris[1], Brandon LeBeau[2] and Isaac T. Petersen[1]

[1]Department of Psychological and Brain Sciences, University of Iowa, IA, USA and [2]Department of Psychological and Quantitative Foundations, University of Iowa, IA, USA

Abstract

Reactive and control processes – e.g., negative emotionality and immediacy preference – may predict distinct psychopathology trajectories. However, externalizing and internalizing problems change in behavioral manifestation across development and across contexts, thus necessitating the use of different measures and informants across ages. This is the first study that created developmental scales for both internalizing and externalizing problems by putting scores from different informants and measures onto the same scale to examine temperament facets as risk factors. Multidimensional linking allowed us to examine trajectories of internalizing and externalizing problems from ages 2 to 15 years ($N = 1{,}364$) using near-annual ratings by mothers, fathers, teachers, other caregivers, and self report. We examined reactive and control processes in early childhood as predictors of the trajectories and as predictors of general versus specific psychopathology in adolescence. Negative emotionality at age 4 predicted general psychopathology and unique externalizing problems at age 15. Wait times on an immediacy preference task at age 4 were negatively associated with age 15 general psychopathology, and positively associated with unique internalizing problems. Findings demonstrate the value of developmental scaling for examining development of psychopathology across a lengthy developmental span and the importance of considering reactive and control processes in development of psychopathology.

Keywords: Bifactor; delay of gratification; heterotypic continuity; longitudinal; negative emotionality

(Received 13 January 2023; revised 5 March 2024; accepted 7 March 2024; First Published online 8 April 2024)

Introduction

Internalizing and externalizing problems are among the most common, costly, and burdensome issues facing children and adults across the lifespan (Forbes et al., 2016). Externalizing disorders, typically encompassing symptoms of aggression, conduct problems, and oppositionality, have a global prevalence of over 5% (Polanczyk et al., 2015). Moreover, untreated externalizing problems lead to future problems such as academic difficulties (Shi & Ettekal, 2021), social delinquency, incarceration, and substance abuse (Krueger et al., 2021; Loeber et al., 1998). Internalizing psychopathology, which encompasses anxiety and depression, has an even higher prevalence ranging from 20 to 25% in children and adolescents, and has greatly increased since the COVID-19 pandemic (Racine et al., 2021). Thus, it is crucial to identify processes that prevent the development of internalizing and externalizing psychopathology.

General psychopathology

Diagnoses from the internalizing and the externalizing spectra show a high rate of co-occurrence (Caspi et al., 2014; Clark et al., 2021; Murray et al., 2016). Internalizing and externalizing problems are most accurately modeled dimensionally rather than as categorical phenomena (Markon et al., 2011), and numerous factor analytic modeling studies have identified strong covariation among internalizing, externalizing, and thought-disordered dimensions of psychopathology (Caspi et al., 2014; Cervin et al., 2021; Choate et al., 2022; Gluschkoff et al., 2019; Lahey et al., 2012). The strong covariation between these problems has led to the hypothesis that internalizing, externalizing, and thought-disordered problems share a common cause – a higher-order factor called "*p* factor" – that accounts for their covariation (Caspi et al., 2014; Forbes et al., 2016; Krueger & Eaton, 2015). The *p* factor can be modeled in various ways, including a higher-order factor model or a bifactor model. The general factor of psychopathology reflects what is common among dimensions of psychopathology. The general factor is operationalized as the common variance among internalizing and externalizing psychopathology indicators (Ree et al., 2015; Watts, Meyer, et al., 2021; Zinbarg et al., 2005). Specific psychopathology (e.g., externalizing) in a bifactor model is represented by the residual correlations among the specific psychopathology facets that are not accounted for by the general factor.

Given the high cost and burden of internalizing and externalizing psychopathology, it is important to identify early processes that lead to development of specific internalizing or externalizing problems versus co-occurring internalizing and externalizing problems. The present study considers reactive and control processes of temperament that may help explain differing developmental pathways to psychopathology.

Corresponding author: Jordan L. Harris; Email: jordan-l-harris@uiowa.edu

Cite this article: Harris, J. L., LeBeau, B., & Petersen, I. T. (2025). Reactive and control processes in the development of internalizing and externalizing problems across early childhood to adolescence. *Development and Psychopathology* **37**: 836–858, https://doi.org/10.1017/S0954579424000713

Temperamental reactive and control processes in the development of general and specific psychopathology

Temperament is defined as constitutionally based ways in which individuals regulate and react to their environment (Rothbart & Bates, 2006). Individual differences in temperament are early appearing, biologically based, and relatively stable across development (Rothbart & Bates, 2006). The structure of temperament is broadly categorized by three relatively orthogonal dimensions – two reactivity dimensions: positive and negative emotionality; and one control dimension: self-regulation (Rothbart & Bates, 2006; Rothbart, 2011). Positive emotionality reflects the tendency to experience positive emotions, e.g., enthusiasm and joy, often reflecting extraversion (Watson et al., 1988). Negative emotionality reflects a propensity toward anger/frustration, fear, and sadness (Rothbart & Bates, 2006). Self-regulation reflects a child's ability to regulate behavior, cognition and emotions, and often includes subordinate constructs of executive functioning, attentional control, and effortful control. In the present study, we focus on facets of the negative emotionality and self-regulation dimensions of temperament given their robust associations with psychopathology (e.g., Eisenberg et al., 2005; Muris & Ollendick, 2005). One of the most widely studied predictors of children's adjustment is immediacy preference, which is a facet of self-regulation and is a control process (Krueger et al., 1996).

Immediacy preference

Immediacy preference is the selection of a smaller, immediate reward over a larger, distal one (Mischel & Ebbesen, 1970; Stephens & Anderson, 2001). The inverse of immediacy preference is delay preference (Rachlin & Jones, 2008), which is frequently called delay of gratification. Immediacy preference is often assessed with a self-imposed waiting task (Metcalfe & Mischel, 1999) that is designed to assess a person's ability or preference to resist the temptation of the immediate reward (i.e., gratification) in favor of a more motivationally salient, distal reward. In a temperament framework, delay of gratification is considered a control process because it describes a self-regulatory process by which an individual might suppress a dominant response in favor of a subordinate one (Moran et al., 2013). This conceptualization of delay of gratification as a control process of temperament is supported by prior research (Bjorklund & Kipp, 1996; Moran et al., 2013; Murray & Kochanska, 2002). We use the term "delay of gratification" when referring to the duration of waiting in a self-imposed waiting task, whereas we use the term "immediacy preference" when referring to the process as a risk factor. Immediacy preference has been shown to be associated with deficits in self-control, a common impairment in individuals with psychopathology (Kidd et al., 2013; Michaelson & Munakata, 2020).

Immediacy preference (i.e., a stronger relative preference for an immediate reward) has been associated with many facets of psychopathology, including impulsive decision making and externalizing psychopathology. Several studies have found that immediacy preference was associated with both parent- and teacher-reported aggressive and delinquent behaviors in children ranging from ages 3 to 13 years (Campbell & von Stauffenberg, 2009; Krueger et al., 1996). A 40-year follow-up study found that adults with greater immediacy preference in childhood continued to show more externalizing problems compared to their peers who had a delay preference in childhood (Casey et al., 2011). Numerous studies have identified a strong association of immediacy preference with externalizing problems and delinquent outcomes (Campbell & von Stauffenberg, 2009; Krueger et al., 1996).

Previous studies have implicated immediacy preference as a risk factor that is specific to externalizing problems, i.e., not internalizing problems (Krueger et al., 1996). However, little work has examined whether immediacy preference is associated with general psychopathology. Watts et al. (2018) used the Eunice Kennedy Shriver National Institute of Child Health and Development (NICHD) Study of Early Child Care and Youth Development (SECCYD) sample, the same sample as the present study. Importantly, Watts and colleagues found that when accounting for numerous covariates, including family background, home environment, and early cognitive ability, immediacy preference was no longer significantly associated with behavioral outcomes (Watts et al., 2018). However, the behavioral outcomes variable was a total behavior problems score computed as an average of internalizing and externalizing problems. Thus, the study did not allow for determining whether immediacy preference was associated with general versus specific psychopathology. Partitioning internalizing and externalizing problems is especially important because there might be reason to expect that immediacy preference is more strongly associated with externalizing problems compared to internalizing problems (Campbell & von Stauffenberg, 2009; Krueger et al., 1996).

Another study using the SECCYD sample (Deutz et al., 2020) examined the association between antecedent factors, including immediacy preference, and specific and general psychopathology at later ages (8 and 14 years old). Results indicated that immediacy preference was not associated with later general or specific psychopathology. However, the study did not examine immediacy preference in relation to the *development* (i.e., change over time) of behavior problems. Moreover, the study examined only mother- and self report, and did not include perspectives from other informants such as fathers.

In sum, much remains unknown regarding the association between immediacy preference and psychopathology. For instance, it is not known the degree to which immediacy preference in early childhood predicts general versus specific psychopathology in adolescence (Deutz et al., 2020). In addition to immediacy preference, the present study considers temperamental negative emotionality, a reactive process, which has been proposed as a transdiagnostic risk factor for psychopathology (Mikolajewski et al., 2013; Weissman et al., 2019). Transdiagnostic risk factors contribute to the etiology and maintenance of a broad range of emotional and behavioral difficulties (Egan et al., 2011). Transdiagnostic risk factors can occur within a given specific factor – e.g., perfectionism is a risk factor for eating disorders, anxiety, and depression (Egan et al., 2011) – or can reflect risk for general psychopathology, meaning they contribute to a broader range of behaviors (Lynch et al., 2021).

Temperamental negative emotionality

Temperamental negative emotionality is an individual's characteristic reaction to stimuli, and includes the tendency to display overt sadness, irritability, fear/withdrawal, distress, and/or anger, and may be characterized by somatic or autonomic physiological reactivity in response to stimuli (Fox, 1989; Rothbart & Derryberry, 1981). As such, negative emotionality is considered a reactive process within a temperament framework (Eisenberg et al., 1996; Moran et al., 2013). Numerous studies have found that negative emotionality is highly correlated with internalizing and externalizing psychopathology (Leaberry et al., 2019; McLaughlin & Nolen-Hoeksema, 2011; Steinberg & Drabick, 2015), and evidence suggests that negative emotionality plays a causal role in

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

the development of internalizing and externalizing psychopathology (Lilienfeld, 2003). There is some evidence that negative emotionality is associated with general psychopathology (i.e., *p* factor; Hankin et al., 2017). Negative emotionality is thought to evolve and change in its manifestation, much like symptoms of psychopathology more broadly (e.g., Aldao et al., 2016; De Los Reyes et al., 2009, 2013; Mischel & Shoda, 1995; Pettersson et al., 2018; Rutter & Sroufe, 2000).

A previous study found that negative emotionality, as observed in a frustration task, was associated with both internalizing and externalizing problems, and partially accounted for their covariation (Mikolajewski et al., 2013). Negative emotionality predicts psychopathology (Brandes et al., 2019; Briggs-Gowan et al., 2006; Hawes et al., 2020) and they overlap in behavioral indicators of distress (e.g., dysthymia and depression; Greene & Eaton, 2017), suggesting that they conceptually overlap to a degree. However, prior literature has indicated that temperament/personality constructs differ from psychopathology. Temperament is primarily concerned with the how of a behavior (e.g., how intensely a child cries), whereas psychopathology focuses on the content of the behavior (e.g., what does the child cry about; Bates et al., 2014; De Pauw & Mervielde, 2010; Lemery et al., 2002; Thomas & Chess, 1977). However, to our knowledge, no previous study has examined control and reactive temperamental processes simultaneously in predicting general and specific psychopathology in the same model.

Effortful control, an aspect of self-regulation which comprises immediacy preference, and negative emotionality are related but separate aspects of temperament (Eisenberg et al., 2005; Rothbart et al., 2001). Children with less effortful control tend to experience greater adjustment difficulties in the face of stress, thus leading them to show more negative emotionality (Moran et al., 2013; Muris & Ollendick, 2005). A prior study using the same sample as the present study found that anger was associated with higher mother-reported externalizing problems. This same study found that higher levels of effortful control at 54 months indirectly predicted lower levels of externalizing problems at age 15 years (Crockett et al., 2018). Furthermore, when examined together, negative emotionality and effortful control have shown additive effects on psychopathology (Eisenberg et al., 1996, 2000, 2005). Relatedly, one study indicated that an imbalance in approach behavior – a reactive process – and control processes was associated with externalizing problems, indicating an interaction of control and reactive processes in externalizing problems (Jonas & Kochanska, 2018). Surprisingly, few studies have simultaneously examined reactive and control systems, such as immediacy preference and negative emotionality, despite evidence that they likely influence each other (Moran et al., 2013; Rothbart & Bates, 2006). To our knowledge, no previous studies have examined these two processes simultaneously as transdiagnostic risk factors for dimensions of psychopathology across childhood to adolescence. However, one barrier to identifying early mechanisms in the development of later psychopathology is that the behavioral manifestations of psychopathology change across development.

Heterotypic continuity

A salient issue in developmental psychopathology is that the behavioral manifestations of psychopathology change across development, a phenomenon called heterotypic continuity (Cicchetti & Rogosch, 2002). Heterotypic continuity occurs when the same psychological phenomenon manifests as different behaviors across development (Petersen et al., 2018, 2020). For example, externalizing problems in children often appear as tantruming and overt oppositionality, whereas in adolescents and adults, externalizing behaviors become more covert and tend to take the form of indirect behaviors (e.g., substance use; Mikolajewski et al., 2013; Miller et al., 2009; Petersen & LeBeau, 2022). Patterson (1993) described externalizing problems using the analogy of a "chimera," a mythological creature with a goat's body that, with development, grows the head of a lion and the tail of a serpent. This was meant to highlight that although individual differences in externalizing behavior are relatively stable across time, externalizing behavior manifests in different ways across the lifespan (Patterson, 1993). This notion of changing manifestations of behavior across development has also been identified in internalizing problems (e.g., Avenevoli & Steinberg, 2001; Petersen et al., 2018; Weems, 2008). For example, separation anxiety and fear of animals is common in younger children, whereas social anxiety is more common in adolescence (Weems, 2008).

Consistent with the developmental issues framework (Sroufe, 2016), the changing behavioral manifestation of psychopathology reflects a combination of time-varying genetic and environmental factors, such as school entry transition, in combination with varying developmental tasks and greater experience-dependent capacity. For example, developmental tasks in preschool (e.g., self-regulation) differ from those in adolescence (e.g., peer acceptance), which changes how behavior problems tend to manifest. Thus, for externalizing and internalizing problems, the underlying construct persists across development, but their behavioral manifestations change. However, there are key challenges in identifying early mechanisms in the development of psychopathology.

A key challenge of identifying early mechanisms in the development of psychopathology deals with longitudinal assessment. It is difficult to examine internalizing and externalizing psychopathology across a lengthy span of development in meaningful ways because behavioral manifestations of psychopathology change across development (McElroy et al., 2018). Assessment and analysis become even more challenging when considering that internalizing and externalizing psychopathology often co-occur (Pettersson et al., 2018). Heterotypic continuity poses challenges for measurement because different measures from different informants across ages are needed to capture developmental changes (Petersen & LeBeau, 2022). If the measures do not align with the changes in the construct's manifestation, studies will yield faulty conclusions (Chen & Jaffee, 2015; Petersen et al., 2018, 2021). In addition to using different measures across ages, it is also important to consider using different informants.

Different informants across development

Before children enter schooling, the most accurate informants on the children's behavior tend to be their parents and caregivers, as reflected in the proliferation of parent- and caregiver-report measures for early childhood (Achenbach & Rescorla, 2000). Many have argued that using multiple informants is the best approach for assessing child psychopathology (De Los Reyes & Makol, 2021, 2022; Makol et al., 2020; Watts, Makol, et al., 2021). In early-to-mid-childhood, when children attend school and preschool, teachers are important informants on children's behavior, because teachers help account for children's behavior across multiple contexts (De Los Reyes & Kazdin, 2005). Context of measurement is important. Prior research has shown that parent and teacher reports of disruptive behaviors that occur in both home and school

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

tend to result in stronger convergence, whereas context-specific behaviors (i.e., disruptive only at home, not school) result in weaker convergence (De Los Reyes & Makol, 2021; Hartley et al., 2011; Kwon et al., 2012). A review on correlations between mother, father, and teacher ratings of ADHD symptoms in children and adolescents indicated that correlations between mother and father ratings of inattention and hyperactivity–impulsivity were high ($r = .67–.70$). Correlations between parent and teacher ratings of inattention and hyperactivity–impulsivity were somewhat lower ($r = .28–.47$). Taken together, evidence indicates that context matters in ratings of observed psychopathology symptoms (Martel et al., 2017).

Furthermore, when children enter their adolescent years, adolescents become more reliable reporters on their internal experience, which is particularly meaningful for internalizing symptoms such as anxiety and depression that may be less overtly visible to outside observers (Damme et al., 2022). Taken together, it is important to consider multiple informants on children's externalizing and internalizing symptoms across development to help account for (a) differing manifestations of behavior in multiple contexts (e.g., school and home) and (b) rater-specific bias.

Developmental scaling

The challenge is in how to meaningfully combine the scores from the different measures as rated by the various informants so that the scores are on a comparable metric for assessing children's change over time. The combination of heterotypic continuity and having different informants across ages poses important challenges. Heterotypic continuity requires age-differing measures to account for the changing manifestation of the construct. Similarly, informants are differentially capable of rating various aspects of the child's behavior (e.g., relational aggression versus social anxiety) in different contexts (e.g., home versus school), thus requiring different measures for different types of informants (e.g., parents, teachers, self report) and at different ages. For example, the Child Behavior Checklist 1.5–5 (Achenbach & Rescorla, 2000) and the Caregiver–Teacher Report Form (Achenbach & Rescorla, 2001) both assess children at the same age range, but the Caregiver–Teacher Report Form includes slightly different question content aimed at examining school rather than home context for behaviors. Furthermore, the Child Behavior Checklist 6–18 (Achenbach & Rescorla, 2000) includes age-differing items (compared to the ages 1.5–5 form) to maintain developmentally appropriate content, e.g., substance use, that assess problems specific to mid-childhood to adolescence. The Teacher's Report Form (Achenbach & Rescorla, 2001) also adjusts its question content to account for development. Therefore, a single informant from one context (e.g., mother in the home context) might only capture one potentially biased view of a given child's behavior (De Los Reyes & Kazdin, 2005). By contrast, including multiple informants across multiple contexts reduces the impact of informant and context-specific biases.

Traditionally, studies have largely ignored heterotypic continuity when examining children's development of psychopathology (Chen & Jaffee, 2015; Petersen et al., 2018, 2021). Many studies have addressed the challenge of differing measures and informants by using (a) only those ages where the same measure or items are assessed and (b) only those informants who provide ratings across the full age span. However, it is problematic to exclude ages because of a developmental change in the manifestation of the construct. Excluding ages due to a change in measurement would exclude important developmental periods and transitions associated with meaningful developmental change, such as the transition from preschool to school entry or the transition to adolescence. Moreover, it is problematic to exclude informants merely because an informant did not provide ratings across all ages of the study. For instance, in a study from early childhood to adolescence, this would result in the exclusion of self report, which is important for assessing internalizing problems in adolescence. It is thus crucial to leverage approaches that use all available information from all possible informants at all possible ages to get the best estimate of people's development on a comparable scale.

Developmental scaling is a recommended approach to ensure statistical equivalence of scores across different measures (Kolen & Brennan, 2014). Developmental scaling approaches have been used successfully to place scores from different measures and informants onto the same scale (Petersen & LeBeau, 2022). Developmental scaling has been widely used in educational psychology to link children's academic achievement scores across ages (e.g., Kenyon et al., 2011; McArdle, 2009; Murayama et al., 2013). However, relatively few studies have used developmental scaling approaches to study social development. For instance, few studies have used developmental scaling to study development of internalizing psychopathology (Petersen et al., 2018) and externalizing psychopathology (Petersen & LeBeau, 2022).

One approach to developmental scaling uses item response theory (IRT). A two-parameter IRT model estimates item difficulty and discrimination. Item difficulty, also called severity, is the point of median probability at which 50% of respondents endorse a given response. Item discrimination is how well the item distinguishes between the high and low levels of a given construct. Based on the items' difficulty and discrimination parameters, one can generate an item characteristic curve, which represents the expected score on the item as a function of the person's level on the latent psychopathology factor. Combining the individual items, one can generate a test characteristic curve, which represents the expected score on the measure as a function of the person's level on the latent psychopathology factor. To link any given pair of measures and raters, IRT uses scaling parameters to minimize the differences between the two test characteristic curves of the common items across the two measures. The scaling parameters are determined as the linear transformation (i.e., intercept and slope parameter) of the test characteristic curves of the common items between the two measures, that, when applied to the second measure, minimize differences between the test characteristic curves of the common items. Essentially, the scaling parameters minimize the differences in the probability of a rater endorsing the age- and rater-common items across the two measures. That is, IRT links measures' scales based on the severity and discrimination of the age- and rater-common items. IRT uses the age- and rater-common items to set the common scale. However, *all* items for a given rater at a given age are used to estimate a person's score on the common scale.

We are aware of only one prior study that has linked scores from different measures and raters across ages (Petersen & LeBeau, 2022). And, to our knowledge, no prior studies have used developmental scaling to link scores from different measures and raters for multiple dimensions of psychopathology. Performing developmental scaling of multiple dimensions of psychopathology leverages the strong covariation between internalizing and externalizing problems to obtain more accurate estimates of each. Researchers have called for studies that implement developmental scaling for multiple dimensions of psychopathology simultaneously (Tackett & Hallquist, 2022).

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

This approach of having multiple informants and measures and placing their scores on the same scale is essential to accurately estimate changes in externalizing and internalizing symptoms, given considerable differences in informant reports and measure scales (Petersen & LeBeau, 2022). Developmental scaling allows for individuals to have their informant and self-reported psychopathology mapped onto growth curves, to show people's change in psychopathology across a long development span. To our knowledge, only one study has used developmental scaling from multiple informants to account for heterotypic continuity in order to study individuals' development of psychopathology (Petersen & LeBeau, 2022).

The present study

The present study examines whether immediacy preference and negative emotionality, control and reactive processes, respectively, predict the development of externalizing and internalizing problems across early childhood to adolescence, using a large sample. We leverage multi-informant ratings from mothers, fathers, teachers, afterschool caregivers, other caregivers, and self report. This is the first study to link scores from multiple measures, raters, and psychopathology dimensions onto the same scale. Prior literature has emphasized the importance of accounting for internalizing and externalizing symptoms concurrently to be a more ecologically valid representation of psychopathology (Ruggero et al., 2019). This is the first study to account for heterotypic continuity of multiple dimensions of psychopathology simultaneously to borrow information from each in the estimation of the other, for more accurate estimates given considerable covariation between internalizing and externalizing problems.

Our study has three primary aims: 1) describe the trajectories of internalizing and externalizing problems across a lengthy developmental span. We use an IRT approach to developmental scaling that places the scores from age-differing measures and raters of internalizing and externalizing problems onto the same scale to account for heterotypic continuity and effects of informant type. 2) We aim to evaluate whether measures of reactive and control processes of temperament predict trajectories of internalizing and externalizing problems. The developmental scaling approach allows us to chart children's trajectories of internalizing and externalizing problems across ages 2–15 years, and to examine immediacy preference and negative emotionality as predictors of children's trajectories. 3) Using a bifactor model, we aim to determine whether immediacy preference and negative emotionality factors predict general versus specific psychopathology.

We hypothesize that immediacy preference will predict general psychopathology and specific externalizing problems, but not specific internalizing problems, because immediacy preference has been more strongly associated with externalizing problems compared to internalizing problems in prior work (Campbell & von Stauffenberg, 2009; Krueger et al., 1996). Little work has examined the association between immediacy preference and general psychopathology, but the strong association between immediacy preference and externalizing problems may drive this association. In addition, we hypothesize that temperamental negative emotionality will predict general psychopathology and specific internalizing problems, but not specific externalizing problems. Negative emotionality, like general psychopathology, is often thought to reflect a general liability (Forbes et al., 2019; Phillips et al., 2022). Additionally, the link between negative emotionality and general psychopathology along with internalizing problems has been established in prior literature (Castellanos-Ryan et al., 2016; Lahey et al., 2021; Lahey, 2009; Olino et al., 2014; Tackett et al., 2013).

Method

Children ($N = 1{,}364$) and their families were recruited for the NICHD SECCYD study in 1991 from 31 hospitals near one of 10 locations in the United States: Little Rock, AR; Irvine, CA; Lawrence and Topeka, KS; Boston, MA; Morganton and Hickory, NC; Charlottesville, VA; Seattle, WA; and Madison, WI. Children were recruited at birth and were followed for data collection in four total phases with multiple timepoints at each phase (Phase I, ages 0–3; Phase II, though 1st grade; Phase III, through 6th grade; Phase IV, through 9th grade) until they were 15 years old. The present study involves behavior problem ratings that were assessed near-annually from ages 2–15 years (except ages 13 and 14). The sample was 48% female, 80.4% White, 12.9% Black, 6.1% Hispanic, 1.6% Asian American, 0.4% American Indian, and 4.7% of "other" ethnicity. At intake, the mother's age ranged from 18 to 46 years of age ($M = 28.11$, $SD = 5.63$), 77% of households had fathers living in the home, and there was an average of 4.27 people living in the household ($SD = 1.17$). For more information about the study methods and participants, see NICHD Early Child Care Research Network (2005).

Exclusion criteria included: (1) The mother was younger than 18 years of age at the time of the child's birth; (2) the family did not anticipate remaining in data collection for at least three years; (3) the child had obvious disabilities at birth and/or remained in the hospital for more than seven days after birth; or (4) the mother was not able to speak conversational English. At enrollment, trained researchers visited the family homes, and families were scheduled for periodic data collection. During a given phase, research assistants visited family homes, childcare, and invited families to the laboratory playroom to collect observations and administer study measures.

Measures

Analysis scripts and a data dictionary of study variables were published at https://osf.io/yz4we/. Descriptive statistics and correlations of study variables are in Table 1.

Behavior problems

Children's externalizing and internalizing behavior problems were rated by mothers, fathers, teachers, afterschool caregivers, other caregivers (e.g., daycare workers and babysitters), and self report. Ratings from different informants and measures were used, depending on the child's age. Behavior problem ratings were completed on the following Achenbach measures: Child Behavior Checklist 2–3 (CBCL 2–3; Achenbach, 1992), Child Behavior Checklist 4–18 (CBCL 4–18; Achenbach, 1991a, 1991b), Caregiver–Teacher Report Form (C–TRF; Achenbach & Rescorla, 2000), Teacher's Report Form (TRF; Achenbach, 1991a), and Youth Self-Report (YSR; Achenbach, 1991b). The ages at which each rater provided ratings is provided in Table 2. The measures that were completed by each rater at each age are depicted in Figure 1.

Items were rated as 0, 1, or 2 corresponding to "not true," "somewhat or sometimes true," or "very true or often true," respectively. The Externalizing scale of the CBCL 2–3 includes the Aggressive Behavior and Destructive Behavior subscales. The Externalizing scale of the CBCL 4–18, TRF, and YSR includes

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

Table 1. Correlation matrix of model variables

Variables	Age	Female	African American	Hispanic	INR	Time Waited	Negative Affect	Externalizing	Internalizing
Age	—								
Female	.00	—							
African American	.00	.00	—						
Hispanic	.00	.00	−.07***	—					
INR	.00	.01***	−.22***	−.06***	—				
Time Waited	.00	.06***	−.25***	−.04***	.20***	—			
Negative Affect	.00	−.04***	.08***	.02***	−.11***	−.11***	—		
Externalizing	−.21***	−.12***	.11***	.01	−.11***	−.12***	.21***	—	
Internalizing	−.07***	.01†	.02***	.00	−.06***	−.05***	.15***	.49***	—
Data points	42,284	42,284	42,284	42,284	39,463	29,791	33,542	25,455	25,455
Missingness	0.00	0.00	0.00	0.00	6.67	29.55	20.67	39.80	39.80
M	7.90	0.48	0.13	0.06	2.86	4.48	3.97	−0.09	−0.10
SD	3.51	0.50	0.34	0.24	2.61	3.01	0.66	1.07	1.16
Min	2.00	0.00	0.00	0.00	0.08	0.00	1.69	−1.95	−2.14
Max	15.00	1.00	1.00	1.00	25.08	7.00	6.25	4.93	5.98
Skewness	0.27	0.07	2.21	3.67	2.53	−0.49	−0.03	0.42	0.74
Kurtosis	−0.42	−2.00	2.90	11.50	10.22	−1.60	−0.11	−0.26	0.45

Note. The correlations and descriptive statistics are presented from data are in long format, where each participant has multiple rows: i.e., one row for each informant-by-timepoint combination. "Age" in years; "INR" = income-to-needs ratio; "Min" = lowest score in the sample; "Max" = highest score in the sample. †$p < .10$; *$p < .05$; **$p < .01$; ***$p < .001$; all *p*s two-tailed.

Table 2. The child's age when each rater provided ratings of the child's behavior problems

	Age (years)											
Rater	2	3	4	5	6*	7	8	9	10	11	...	15
Mother*	x	x	x	x	x		x	x	x	x		x
Father					x		x	x	x	x		x
Teacher				x	x	x	x	x	x	x		
Afterschool Caregiver					x		x	x	x			
Other Caregiver	x	x	x									
Self-Report												x

Note. "x" indicates the measure was collected at the specified age. "*" indicates the referent age and rater.

the Aggressive Behavior and Delinquent Behavior subscales. The Externalizing scale of the C–TRF includes the Aggressive Behavior and Attention Problems subscales. The scales include symptoms of breaking rules, cruelty, aggression, and destruction of property.

The Internalizing scale of the CBCL 2–3 includes the Anxious/Depressed and Withdrawn subscales. The Internalizing scale of the CBCL 4–18, TRF, and YSR includes the Anxious/Depressed, Somatic Complaints, and Withdrawn subscales. The Internalizing scale of the C–TRF includes the Anxious/Obsessive, Fears, and Depressed/Withdrawn subscales.

The Achenbach scales are widely used, and the scores show strong reliability (internal consistency, test–retest reliability, and interrater reliability) and validity (content, construct, and criterion-related validity; Sattler, 2022). Internal consistency estimates by age and rater of the present study are in Supplementary Table S1.

Due to the wide age range of children and adolescents included in the study, we took steps to ensure that the same construct was assessed on the same scale, across time. The number of common items among different measures are in Table 3. To account for developmental changes in both internalizing and externalizing problems, we used an IRT approach to developmental scaling, consistent with previous research (described later; Kolen & Brennan, 2014; Petersen et al., 2018; Petersen & LeBeau, 2022). Developmental scaling linked behavior problem scores across ages and informants onto the same scale. One-year cross-time stability estimates by rater are in Supplementary Table S2. Full descriptive statistics of externalizing and internalizing problems by age and rater are in Table 4, and correlations among internalizing and externalizing problems by rater are in Table 5. Furthermore, the percentage of participants with behavior ratings, by rater type, are in Supplementary Table S3.

Predictors

Delay of gratification

Delay of gratification was assessed in the present study with a self-imposed waiting task when the participants were 54 months old. An experimenter elicited the child's preference in treats. Then, the child was told that they would engage in a game where the experimenter would leave the child in the room with the preferred treat. Further, the child was told that if they waited until the experimenter returned, the child could eat the treat and receive an additional portion of treat as reward for waiting the full length of the task. The child was also instructed that if they ring a bell, that will signal to the experimenter that the child does not want to wait, and that the child would only receive the portion of treats presented and no extra portions. The recorded experimental trial had a 7-minute ceiling, and the child's score was the total number of seconds the child waited until ringing the bell, eating the treat, or 7 minutes (whichever came first). The self-imposed waiting task is one of the most widely used performance-based tasks of inhibitory control and motivational self-regulation in psychological research.

Negative emotionality

Temperamental negative emotionality was assessed from mother- and other caregiver report at 54 months using a modified version of the Children's Behavior Questionnaire (CBQ; Rothbart et al., 2001). Raters on the CBQ were asked to rate how well each item described the child in the past 6 months using a 7-point scale where 1 = *extremely untrue* and 7 = *extremely true.* Of the 196 items in the original CBQ, mothers completed 80 items and other caregivers completed 48 items. We used the Negative Affect scale on the CBQ, which encompasses the following subscales: Anger/Frustration (10 items), Sadness (10 items), and, among mother report, Fear (10 items). Other caregivers did not complete the Fear subscale. Internal consistency estimates were $\alpha = .82$ for mothers' reports and $\alpha = .93$ for other caregivers' reports. Mothers' ratings were modestly correlated with other caregivers' ratings ($r[761] = .09$, $p = .017$). To incorporate a multi-informant perspective into the estimation of the child's negative emotionality, we averaged ratings across raters. Negative emotionality as assessed by the CBQ has been widely used and has shown strong internal consistency, rank-order stability, and construct validity (Rothbart et al., 2001).

Due to potential inflation of prediction of psychopathology outcomes due to overlap in item content between the CBQ and the numerous measures of child psychopathology, we dropped items from the mother report and other-caregiver-report CBQ that we judged to conceptually overlap with items from the Internalizing or Externalizing scales of the CBCL. We dropped two items from the mother-report CBQ: one item that assessed frustration and one item that assessed sadness. We dropped one item from the other-caregiver-report CBQ that assessed temper tantrums. Two items from the mother report and two items from the other caregiver report had similar item content but were distinct enough to retain, e.g., an item pertaining to getting irritated when making a mistake versus being afraid to make a mistake. In this case, both items conceptually assess behaviors pertaining to making a mistake, but one behavior is anticipatory whereas the other is reactive; thus, we considered them conceptually distinct.

Covariates

Several demographic characteristics were examined as covariates in the growth curve and bifactor models, including the child's sex (1 = girl, 0 = boy), race, and ethnicity, and the income-to-needs ratio of the child's family. These covariates were selected because they have been shown to robustly impact associations between predictors and child behavior problems (e.g., Petersen et al., 2021; Shi et al., 2020). Race was a dummy coded variable where African American children were compared to other races. Ethnicity was dummy coded such that Hispanic children were compared to non-Hispanic children.

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

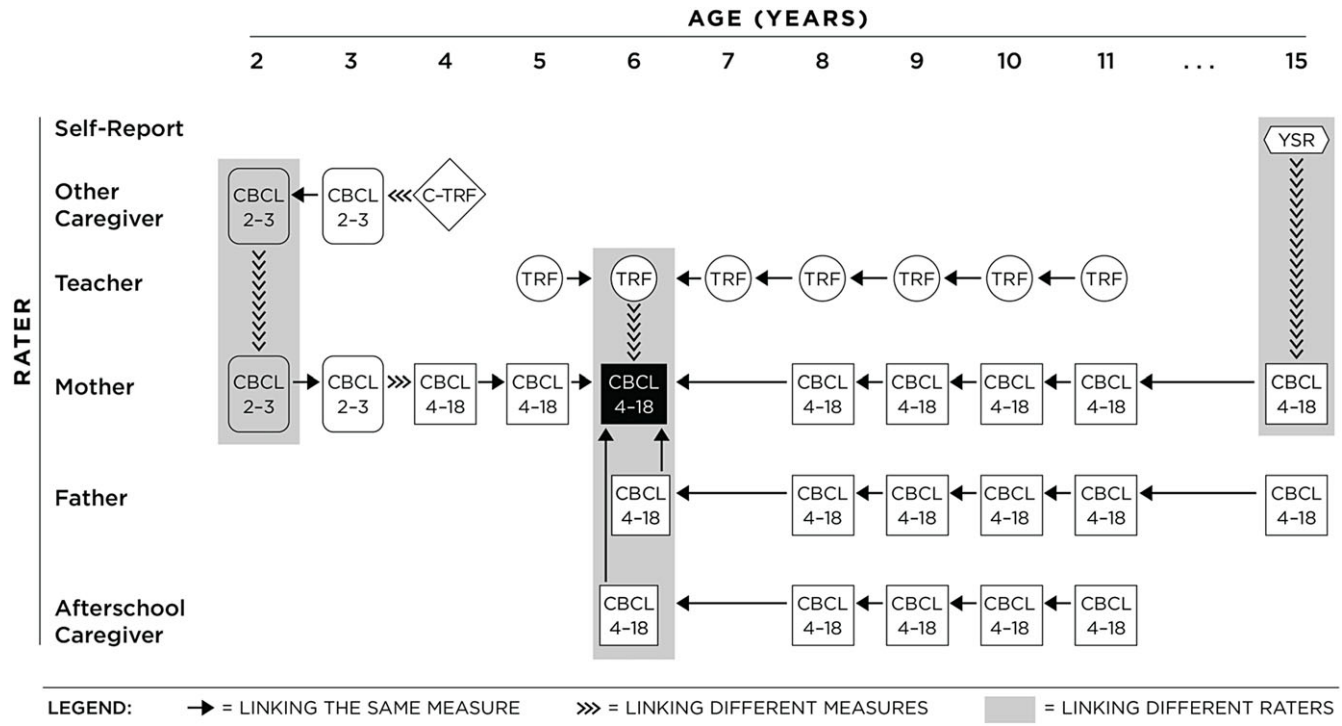


Figure 1. Depiction of how the scores from various raters and measures were linked at different ages. *Note*. Raters are depicted in the rows, and the child's age (in years) is depicted in the columns. Different shapes indicate different measures (square = Child Behavior Checklist 4–18; rounded square = Child Behavior Checklist 2–3; circle = Teacher's Report Form; diamond = Caregiver–Teacher Report Form; hexagon = Youth Self-Report). A solid arrow indicates that scores from the same measure were linked using all items (i.e., all items were common items; e.g., mothers' ratings at ages 6 and 8). A broken arrow indicates that scores from different measures were linked using the common items (e.g., mothers' ratings at ages 3 and 4). The direction of the arrow indicates the measure to which the other was linked (e.g., mothers' ratings at age 8 were linked to mothers' ratings at age 6). The solid black box indicates the referent measure (mothers' ratings at age 6) to which every other measure was linked either directly or indirectly. The gray bounding boxes indicate that scores from different raters were linked using the common items (e.g., self-report ratings at age 15 were linked to mothers' ratings at age 15).

Statistical analysis

Developmental scale of externalizing and internalizing problems

We used multidimensional IRT (M-IRT) and linking to create a single uniform developmental scale (i.e., developmental scaling) for externalizing and internalizing problems that spans multiple years of development. We conducted this linking in five steps: (1) Fit M-IRT models at each age and for each rater type separately. (2) Link the measures' scores over time within each rater type. (3) Link scores across raters. (4) Calculate latent factor scores on the linked scale. (5) Use linked factor scores in growth curve and bifactor models. We describe this procedure in detail below. Full description of linking details are in Supplementary Appendix S1.

Step 1. Fit M-IRT models at each age and for each rater type separately

We used the multidimensional graded response IRT model using the mirt package (Chalmers, 2012) in R 3.6.1 (R Core Team, 2022) to estimate item parameters. The mirt package uses a maximum likelihood expectation-maximization algorithm to estimate item parameters. The maximum likelihood estimation procedure uses all available data for each item and provides valid inferences if the data are missing at random or completely at random. The graded response model is a generalized version of the two-parameter logistic model for dichotomous outcomes, accommodating polytomous items that are ordinal in nature. The multidimensional graded response model adds the ability to include multiple latent factors (i.e., externalizing and internalizing problems) – and their covariance – in the same model. This multidimensional graded response IRT model is conceptually like a two-factor categorical confirmatory factor analysis approach (fit to ordinal data) with the internalizing and externalizing latent factors allowed to covary, and with no cross loadings. That is, internalizing and externalizing problem items were included in the same model, but they were allowed to load onto distinct latent factors. The externalizing and internalizing problem items in the current study were questionnaire items rated from 0 to 2. We used the externalizing problems latent factor as the reference group and allowed the mean and variance for internalizing problems latent factor to be estimated freely. Setting the externalizing factor as the reference group, along with linking both internalizing and externalizing items in the same model, placed the internalizing and externalizing problem scores onto the same mathematical scale across ages and raters.

There may be shifts in the externalizing or internalizing problem constructs over time due to natural developmental changes (Petersen et al., 2018). The present study spans a wide age range (ages 2–15 years). When spanning a wide age range, it is considered safer to fit a separate model at each age rather than a single model that spans all ages because a model that spans across a wide age range is more likely to violate IRT dimensionality assumptions (Kolen & Brennan, 2014). We fit two latent factors corresponding to the constructs of interest: i.e., externalizing and internalizing problems. IRT assumes that each latent factor (e.g., externalizing problems) is unidimensional, which is more likely at a single time point than across all time points in the same model.

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

Table 3. The number of common items for each pair of measures

Measure	CBCL 2–3	CBCL 4–18	C–TRF	TRF	YSR
Externalizing Problems					
CBCL 2–3	26				
CBCL 4–18	9	33			
C–TRF	18	14	40		
TRF	10	27	16	34	
YSR	8	30	14	27	30
Internalizing Problems					
CBCL 2–3	25				
CBCL 4–18	8	31			
C–TRF	17	9	34		
TRF	8	31	10	35	
YSR	8	29	8	29	31

Note. "CBCL" = Child Behavior Checklist, "C–TRF" = Caregiver–Teacher Report Form, "TRF" = Teacher's Report Form, "YSR" = Youth Self-Report. The top table presents the number of common items on the Externalizing scale. The bottom table presents the number of common items on the Internalizing scale. Numbers on the diagonal represent the total number of items in the Externalizing scale (top table) or Internalizing scale (bottom table) for that measure (e.g., the CBCL 4–18 has 33 items on the Externalizing scale and 31 items on the Internalizing scale). 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 4–18 has 6 unique externalizing items when compared with the TRF (i.e., 33 total items minus 27 common items). Conversely, the TRF has 7 unique externalizing items when compared with the CBCL 4–18 (i.e., 34 total items minus 27 common items).

Thus, we fit a separate IRT model at each age and for each rater type in the present study. This approach was also applied by Petersen et al. (2018) and by Petersen & LeBeau (2022) in their creation of a developmental scale for internalizing and externalizing problems, respectively, across a wide age range.

Step 2. Link the measures' scores over time within each rater type

After successful estimation of the individual IRT models, we used multidimensional linking methodology to create the developmental scale for externalizing and internalizing problems. Developmental scaling is a form of data harmonization that aims to place two measures that assess the same construct but differ based on severity and discrimination onto the same scale. One way to create a developmental scale is to link the two measures. The strength of the linking is enhanced if there are items that overlap across the two measures, often referred to as common items. Developmental scaling based on item parameter invariance theory assumes that any difference in item parameter estimates can be rescaled onto a single unified metric with a linear transformation across adjacent ages. Based on this assumption, the item parameters, and the resulting latent factor scores of externalizing and internalizing problems can be linked across ages by comparing and linearly transforming differences in discrimination and severity of the common items across adjacent ages.

We used multidimensional developmental scaling techniques to link the measures' scores over time within each rater type. We used the plink package (Weeks, 2010) in R to perform the linking by using the multidimensional test characteristic function procedure with an oblique Procrustes rotation (Oshima et al., 2000). The oblique rotation method allowed the latent factors – externalizing and internalizing problems – to be correlated. For linking, we used a multidimensional Stocking-Lord procedure (Stocking & Lord, 1983). The Stocking-Lord linking procedure iteratively estimates linking constants by minimizing differences in the aggregate scores across common items.

To estimate the Stocking-Lord parameters, we set the reference age at 6 years for each rater because age 6 was the first age when most rater types (except other caregivers and self report) provided ratings of the child's externalizing and internalizing problems. We set the reference rater to be the mother because the mother typically provided the most ratings across the developmental age span. The reference age and rater pair set the scale to which the item parameters at subsequent ages and for other raters were transformed. In other words, we transformed the estimated item parameters at all ages and for all raters to be on the same scale as the item parameters estimated for mothers' ratings at 6 years of age. To achieve this, we first linked the item parameters across ages within rater type. We performed the process of linking iteratively by chaining together multiple linking constants across the age span. First, for a given rater type, we estimated Stocking-Lord linking constants that linked the item parameters at age 7 to be on the same scale as that rater type's item parameters at age 6. We estimated additional linking constants between adjacent age spans, for example between 5 and 6 years of age, 7 and 8 years of age, and so on. We used two estimated scaling constants including an intercept parameter, B, and a slope parameter, A, to link the item parameters onto the reference scale.

After successfully estimating the linking constants, we then transformed all item parameters to be on the age 6 scale for the given rater. Min (2007) provides further technical details on the multivariate linking terms. To shift all item parameters to a common age 6 scale, we applied all previous adjacent scaling constants to the item parameters. For example, when shifting the item parameter estimates for 7-year-olds to the age 6 scale, we used a single set of scaling constants. However, when shifting the item parameters for 8-year-olds, we used two sets of scaling constants: first, we transformed the item parameter estimates for 8-year-olds to the scale of the 7-year-olds, and then we transformed them a second time to be on the age 6 scale. See Figure 1 for a visualization of the linking process. We performed this step of the linking process separately for each row in the figure (i.e., within rater types; horizontal arrows).

Step 3. Link scores across raters

After creating developmental scales across ages within rater types, we linked scores across raters at age 6 (except for the other caregivers' reports collected at age 2 and self report collected at age 15). As described above, we set the mother as the reference rater. We used a similar process as in step 2; we estimated Stocking-Lord linking constants to link the item parameters across raters within a single age. For example, we estimated a set of linking constants to link the item parameters of the fathers' ratings to the item parameters of mothers' ratings at age 6 to ensure that their factor scores were on the same scale. This step moved the developmental scales for fathers, teachers, and afterschool caregivers to the mothers' scale, anchored at age 6, while preserving the developmental scale created within rater types in step 2. The process of linking scores across raters is depicted in Figure 1 with the gray bounding boxes (vertical arrows).

Step 4. Calculate latent factor scores on the linked scale

After successfully placing item parameter estimates onto a single developmental scale (for all raters and ages), we calculated

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

Table 4. Descriptive statistics of externalizing and internalizing problems by age and rater

	Age (Years)										
M	2	3	4	5	6	7	8	9	10	11	15
Mother	0.73	0.61	0.62	0.07	0.00	–	−0.14	−0.30	−0.33	−0.38	−0.49
	−0.09	0.03	0.42	−0.06	0.00		0.06	0.07	0.11	0.08	0.03
Father	–	–	–	–	−0.11	–	−0.33	−0.37	−0.50	−0.42	−0.47
					−0.10		−0.08	−0.10	−0.13	−0.19	−0.22
Teacher	–	–	–	−0.28	−0.21	−0.20	−0.10	−0.20	−0.09	−0.30	–
				−0.92	−0.74	−0.75	−0.54	−0.57	−0.62	−0.65	
Afterschool Caregiver	–	–	–	–	−0.09	–	−0.27	−0.32	−0.44	–	–
					−1.22		−1.30	−1.39	−1.40		
Other Caregiver	0.41	0.30	−0.55	–	–	–	–	–	–	–	–
	1.21	1.38	1.07								
Self-Report	–	–	–	–	–	–	–	–	–	–	0.19
											0.77
	Age (Years)										
SD	2	3	4	5	6	7	8	9	10	11	15
Mother	0.79	0.80	0.79	0.89	0.93	–	0.93	0.92	0.96	0.95	0.98
	1.00	1.01	1.03	0.96	0.94		1.03	1.03	1.04	1.06	1.04
Father	–	–	–	–	0.98	–	0.94	1.01	1.04	1.04	1.04
					0.96		0.98	0.98	1.04	1.03	1.09
Teacher	–	–	–	1.13	1.15	1.15	1.22	1.17	1.18	1.20	–
				0.82	0.86	0.90	0.93	0.90	0.89	0.91	
Afterschool Caregiver	–	–	–	–	0.98	–	1.00	0.97	1.01	–	–
					0.65		0.59	0.60	0.67		
Other Caregiver	0.94	1.00	1.08	–	–	–	–	–	–	–	–
	1.34	1.29	1.36								
Self-Report	–	–	–	–	–	–	–	–	–	–	0.92
											1.18

Note. "–" indicates not applicable because the particular rater did not provide ratings at the given time point. Means and standard deviations (*SD*s) for externalizing problems are the top number in each box, whereas means and *SD*s for internalizing problems are the bottom number.

Table 5. Correlation matrix of externalizing problem scores (below diagonal) and internalizing problem scores (above diagonal) by rater

Rater	Mother	Father	Teacher	Afterschool Caregiver	Other Caregiver	Self-Report
Mother	.56***	.42***	.23***	.23***	.13***	.29***
Father	.57***	.66***	.19***	.22***	n/a	.29***
Teacher	.33***	.33***	.35***	.20***	n/a	n/a
Afterschool Caregiver	.39***	.41***	.44***	.55***	n/a	n/a
Other Caregiver	.21***	n/a	n/a	n/a	.57***	n/a
Self-Report	.32***	.33***	n/a	n/a	n/a	.48***

Note. "n/a" indicates not applicable because the two raters did not provide ratings at the same time point(s). Diagonal represents within-informant correlations between externalizing and internalizing problems scores.
$^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$; all *p*s two-tailed.

children's latent externalizing and internalizing problem scores with expected a posteriori factor scores. The linking in the previous two steps scaled the factor scores to be on the single developmental scale while retaining changes in means and variances over time and across raters. The linking constants by measure and age are in Supplementary Table S4.

In sum, the linking of scores within a rater type created a developmental scale for scores from that rater type, so each rater type had their own trajectory (see Figure 2). We then, ultimately, linked each rater type's developmental scale (directly or indirectly) to the mothers' ratings at age 6, so that each rater type's trajectory was on the same developmental scale. Examples of linked scores

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

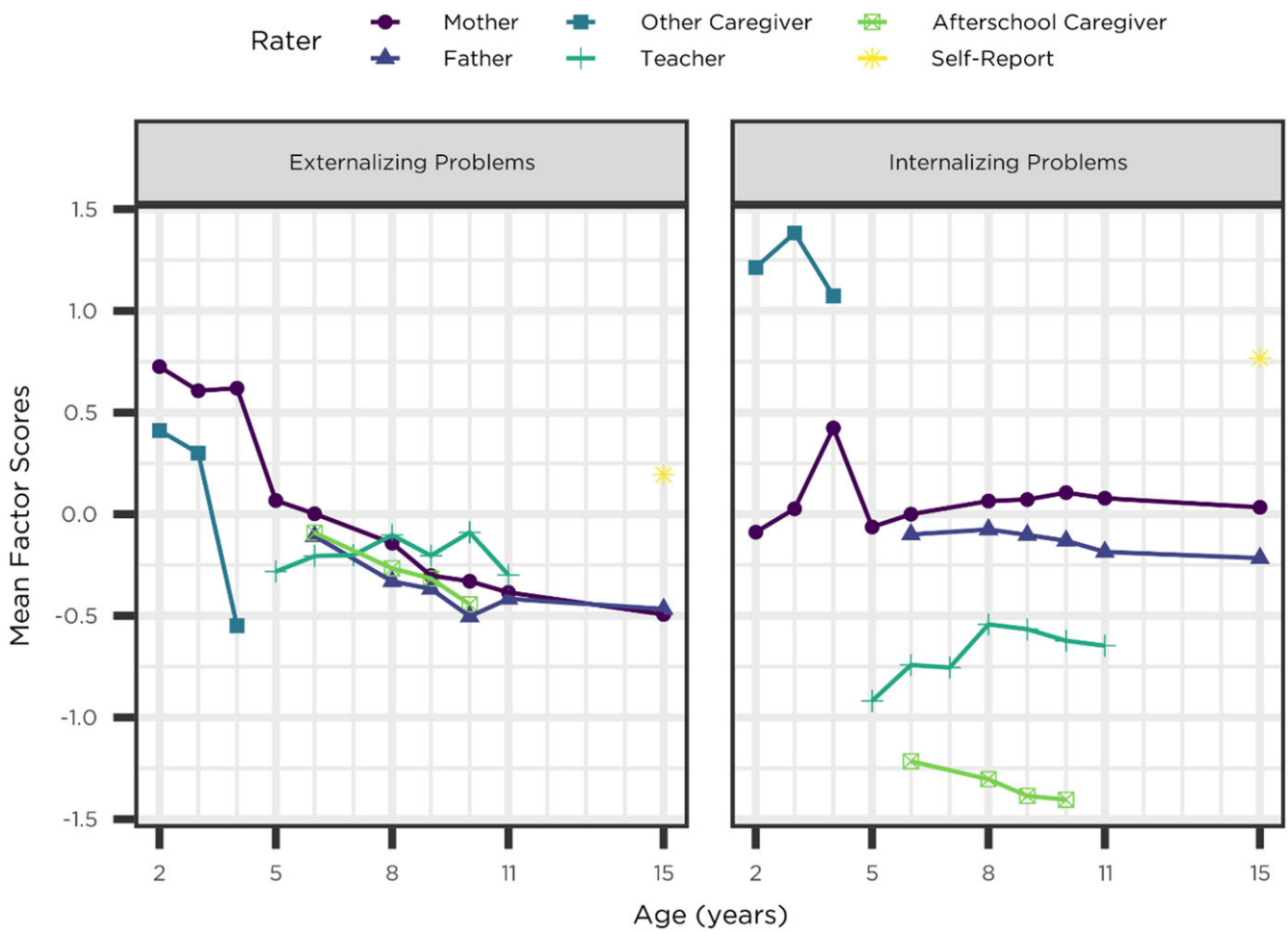


Figure 2. Mean factor scores by age, rater, and type of behavior problem.

across raters and years are depicted with test characteristic curves in Supplementary Figures S1 through S4. The test characteristic curves of the linked scores across raters and years were highly similar (and more similar than the test characteristic curves of the pre-linked scores), indicating that we successfully linked scores across raters and years to be on the same scale.

Differential item functioning. Post-linking estimates of scale-level DIF between measures used to link scores across different raters and ages are in Supplementary Table S5. Tests of differential item functioning (DIF) by age showed no major concerns at the scale level after linking (see Supplementary Appendix S2). The distribution of DIF effect size statistics between ages by rater type are in Supplementary Figure S5. We also conducted analyses to examine potential differential item functioning by sex and race (see Supplementary Appendix S3). There was some differential item functioning between males and females, and between Whites and Non-Whites. However, most instances of DIF were differences in severity (i.e., uniform DIF) rather than differences in discrimination (non-uniform DIF). Thus, covariate adjustment within the growth models should adjust for group-level differences in the factor score differences by sex and race.

Secondary analysis of aggression and delinquent behavior. As a secondary analysis, we also examined aggression and delinquent subdimensions of externalizing problems given their differing associations with risk factors (J. Murray & Farrington, 2010; Wall & Barth, 2005). Thus, we also conducted developmental scaling with aggression and delinquent behavior (see Supplementary Appendix S4).

Step 5. Use linked factor scores in growth curve and bifactor models. After linking factor scores from all raters and at all ages to be on the scale of mothers' ratings at age 6, we used the linked factor scores as the child's estimated level of behavior problems for a given rater and age in subsequent growth curve and bifactor models.

Growth curve modeling

We modeled children's trajectories of behavior problems using mixed models. We estimated mixed models using the lmer function of the lme4 package (D. Bates et al., 2015) in R. We modeled externalizing problems and internalizing problems separately given their different developmental courses. Because our goal was to predict behavior problems in adolescence, we set the intercepts to be at the last time point (age 15), consistent with prior research. First, we established the form of change of children's behavior problems over time by comparing linear, quadratic, and cubic polynomials. To reduce the potential for multicollinearity among the polynomial terms, we used orthogonal

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

polynomials calculated using the poly function in R. We used chi-square difference tests to compare nested models.

Upon establishing the form of the trajectory, we added dummy-coded fixed effects for the rater type: mother, father, teacher, afterschool caregiver, other caregiver, or self report. We set mother report to be the reference group because mothers provided the most ratings on average. Then, we added fixed effects for the covariates. After adding covariates, we then added the focal predictors of interest, the child's negative emotionality and delay of gratification. For parsimony and interpretability, we examined the predictors in relation to the intercepts and linear slopes. Model formulas are in Supplementary Appendix S5.

We determined the importance of focal predictors using R^2 statistics to evaluate how much variation in behavior problems was explained by the rater predictors, demographic predictors, and focal predictors of interest. We computed R^2 statistics defined by Nakagawa and Schielzeth (2013). In Supplementary Appendix S6, we describe tests of systematic missingness and how missing data were handled.

Bifactor model

We estimated a bifactor model at the last timepoint in the study (i.e., age 15) to determine whether negative emotionality and delay of gratification predicted general versus specific psychopathology. First, we fit a bifactor model at age 15 years with only externalizing and internalizing items and no predictors. The bifactor model included a latent factor for the general factor of psychopathology, in addition to latent factors for externalizing problems and internalizing problems. The latent factors for internalizing and externalizing problems for each participant were estimated using the developmentally scaled factor scores from all informant types who provided ratings at age 15 (mother-, father-, and self report). The latent factors were set to be uncorrelated, so the general factor represented the covariation among all externalizing and internalizing items. By contrast, the specific psychopathology factors – i.e., externalizing problems and internalizing problems – represented the covariation among the items within that dimension after extracting the variance accounted for by the general factor. The externalizing latent factor represented unique externalizing variance, and the internalizing latent factor represented unique internalizing variance. Upon establishing a well-fitting bifactor model, we examined the focal predictors of interest. Predictors were allowed to predict the three latent factors. Then, we added covariates.

Bifactor models were fit in lavaan (Rosseel, 2012) in R. Models with diagonally weighted least squares (WLSMV) were unable to be estimated due to sparse cells in some response categories for some items. Therefore, models used maximum likelihood estimation and a probit link with robust standard errors (MLR-probit) to account for the nonnormally distributed data, which has shown comparable power to WLSMV and better control for Type I error when using ordinal data (Cuhadar & Kalkan, 2023; Suh, 2015). Missing data were handled with full information maximum likelihood estimation, which uses all available data and is the gold standard approach for handling missingness when data are missing at random or completely at random. We evaluated model fit with the root mean square error of approximation (RMSEA), Robust estimate of comparative fit index (CFI), and standardized root mean square residual (SRMR). Model fit was considered good if RMSEA $\leq .05$, CFI $\geq .95$, and SRMR $\leq .08$; model fit was considered acceptable if RMSEA $\leq .08$, CFI $\geq .90$, and SRMR $\leq .10$ (Bentler & Bonett, 1980; Hu & Bentler, 1999; Schermelleh-Engel et al., 2003; Schreiber et al., 2006).

Due to the sparse cells in some response categories for some items, we structured the multi-informant ratings at age 15 in long form by rater, to leverage ratings from all raters at age 15: mothers, fathers, and self-report. That is, each participant had up to three rows (one for each rater). To account for the non-independence of multiple observations per participant, we used the participant as a cluster variable, which calculates robust standard errors using a Huber-White sandwich estimator (Huber, 1967; White, 1980).

The scale of each latent factor was set using the effects coding method (Little et al., 2006) so that the average of the items' factor loadings was equal to one. Composite reliability, indexed by coefficient omega, was estimated using the semTools package (Jorgensen et al., 2022) in R.

Sensitivity analyses

We conducted several sensitivity analyses. Methods of the sensitivity analyses are in Supplementary Appendices S7 and S8. In a secondary analysis, we also included the child's early cognitive ability as an additional covariate (see Supplementary Appendix S7).

Results

The present study sought to achieve three aims: First, we sought to describe people's trajectories of externalizing and internalizing problems from ages 2–15 years after performing developmental scaling to put the scores from the different measures, informants, and constructs onto the same scale. Second, we examined whether negative emotionality and immediacy preference predict the slopes from ages 2–15 and intercepts (i.e., ending levels) of people's trajectories of externalizing and internalizing problems at age 15. Third, we examined whether negative emotionality and immediacy preference predict general psychopathology versus specific psychopathology – i.e., unique externalizing problems or unique internalizing problems – at age 15.

Aim 1: Describe the trajectories

Unconditional means model

An unconditional means model (i.e., random intercepts) demonstrated that individual differences in intercepts accounted for 34% of the variance in children's trajectories of externalizing problems and 18% of the variance in children's trajectories of internalizing problems. A model with random intercepts and random linear slopes fit better than a model with only random intercepts and accounted for 40% of the variance in externalizing problems and 22% of the variance in internalizing problems (externalizing: $\Delta\chi2[3] = 1{,}809.63$; internalizing: $\Delta\chi2[3] = 408.11$; $ps < .001$).

Functional form

To determine the best-fitting functional form, we compared models with random linear, quadratic, and cubic slopes. A model with random linear and quadratic slopes fit better than a model with only random linear slopes (externalizing: $\Delta\chi2[4] = 1{,}543.94$; internalizing: $\Delta\chi2[4] = 1{,}937.47$; $ps < .001$). A model with random cubic slopes did not converge due to small variance of the random cubic term. Thus, we selected the quadratic model as the best-fitting functional form of growth.

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

Rater type

Then, we added the rater type (e.g., mother, father, teacher) as a dummy-coded predictor of the trajectories to account for systematic differences as a function of rater type. Mothers served as the reference rater group. The model with rater type as a predictor fit better than a model without rater type (externalizing: $\Delta\chi2[13] = 2{,}927.75$; internalizing: $\Delta\chi2[13] = 8{,}780.80$; $ps < .001$). The model with rater type predicting linear and quadratic slopes fit better than models with rater type predicting only linear slopes (externalizing: $\Delta\chi2[4] = 417.34$; internalizing: $\Delta\chi2[4] = 69.95$; $ps < .001$). Thus, for the baseline growth model, we selected the model in which the rater type predicted the intercepts, linear slopes, and quadratic slopes to allow different curvature by rater type.

Baseline growth model

Individuals' growth curves from the baseline growth model are depicted in Figure 3. Model results are in Supplementary Table S6.

For externalizing problems, the model explained 47% of the variance (fixed effects: 10%; random effects: 37%). Ratings by fathers, teachers, and afterschool caregivers had lower intercepts compared to ratings by mothers. Self-report had higher intercepts than ratings by mothers. When setting the intercepts to the first timepoint when the target informant type provided ratings, ratings by other caregivers had higher intercepts than ratings by mothers (see Supplementary Appendix S9). In summary, mothers tended to rate their child as showing more externalizing problems than did fathers, afterschool caregivers, and teachers (except for teachers' ratings after age 8; see Figure 2); however, mothers tended to rate their child as showing fewer externalizing problems than did other caregivers or self-report. On average, externalizing problems decreased across ages 2–15, but trajectories differed by rater. Mothers' ratings showed average decreases across ages 2–15, whereas teachers' ratings showed modest average increases from ages 4–8 and then stayed relatively stable with slight declines from ages 8–11.

For internalizing problems, the model explained 47% of the variance (fixed effects: 24%; random effects: 23%). Ratings by fathers, teachers, and afterschool caregivers had lower intercepts compared to ratings by mothers. Self-report had higher intercepts than ratings by mothers. When setting the intercepts to the first timepoint when the target informant type provided ratings, ratings by other caregivers had higher intercepts than ratings by mothers (see Supplementary Appendix S9). In summary, mothers tended to rate their child as showing more internalizing problems than did fathers, afterschool caregivers, and teachers; however, mothers tended to rate their child as showing fewer internalizing problems than did other caregivers or self-report. On average, internalizing problems showed decreases from ages 2–10 and increases from ages 10–15, but trajectories differed by rater. Mothers' ratings were relatively stable from ages 2–15, whereas teachers' ratings showed modest average increases from ages 4–8 and then stayed relatively stable from ages 8–11.

Demographic covariates

Then, we added demographic characteristics to the model as covariates. Model results are in Supplementary Table S7. For externalizing problems, the fixed effects explained 12% of the variance (Δ2%). Compared to girls, boys had higher intercepts but did not differ in slopes. Relative to other races, African American children had higher intercepts but did not differ in slopes. Relative to non-Hispanic children, Hispanic children did not differ in intercepts but showed marginally significant differences in slopes in terms of shallower decreases with age. A lower income-to-needs ratio was associated with higher intercepts but not different slopes. In other words, boys, African Americans, and children from families with a lower income-to-needs ratio had higher ratings of externalizing problems. Compared to non-Hispanic children, Hispanic children showed lesser decreases in ratings of externalizing problems across development.

For internalizing problems, the fixed effects in the baseline growth model explained 24% of the variance (Δ <1%). Compared to boys, girls had higher intercepts and had steeper increases over time. Relative to other races, African American children had steeper decreases over time but did not differ in intercepts. Relative to non-Hispanic children, Hispanic children did not differ in intercepts or slopes. A lower income-to-needs ratio was associated with higher intercepts but not different slopes. In other words, girls and children from families with a lower income-to-needs ratio had higher ratings of internalizing problems. Girls showed steeper increases in internalizing problems compared to boys; African Americans showed steeper decreases in internalizing problems compared to non-African Americans.

Aim 2: Predicting the trajectories

Then, we added negative emotionality and delay of gratification as predictors in the model. Model results are in Supplementary Table S8.

For externalizing problems, the fixed effects in the baseline growth model explained 17% of the variance. Thus, the predictors of interest collectively accounted for ~5% of additional variance. Higher negative emotionality was associated with higher intercepts but steeper declines in externalizing problems over time. Greater delay of gratification was associated with lower intercepts but not with differences in slopes. Prototypical growth curves of externalizing problems as a function of delay of gratification are in Figure 4.

For internalizing problems, the fixed effects in the baseline growth model explained 26% of the variance. Thus, the predictors of interest collectively accounted for ~3% of additional variance. Higher negative emotionality was associated with higher intercepts but steeper declines in internalizing problems over time. Greater delay of gratification was associated with lower intercepts but not with differences in slopes.

Sensitivity analysis results: growth curve models

For results of sensitivity analyses of growth curves, see Supplementary Appendix S9. Early cognitive ability was associated with lower intercepts of both internalizing and externalizing problems. When accounting for early cognitive ability as a covariate, its presence attenuated the previously significant associations between delay of gratification and intercepts of internalizing and externalizing problems. When examining mother-only ratings of internalizing and externalizing problems, negative emotionality and delay of gratification did not predict differences in slopes in internalizing and externalizing problems that were present when examining ratings from all informants. Results regarding intercepts did not differ significantly.

When excluding ratings prior to 54 months of age, results did not substantively change. When examining anger/frustration and fear, separately, instead of a negative emotionality composite, anger/frustration was strongly associated with higher externalizing and internalizing problem intercepts, and steeper declines in externalizing problems and, at a trend-level, internalizing problems. By contrast, fear was not associated with intercepts or

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

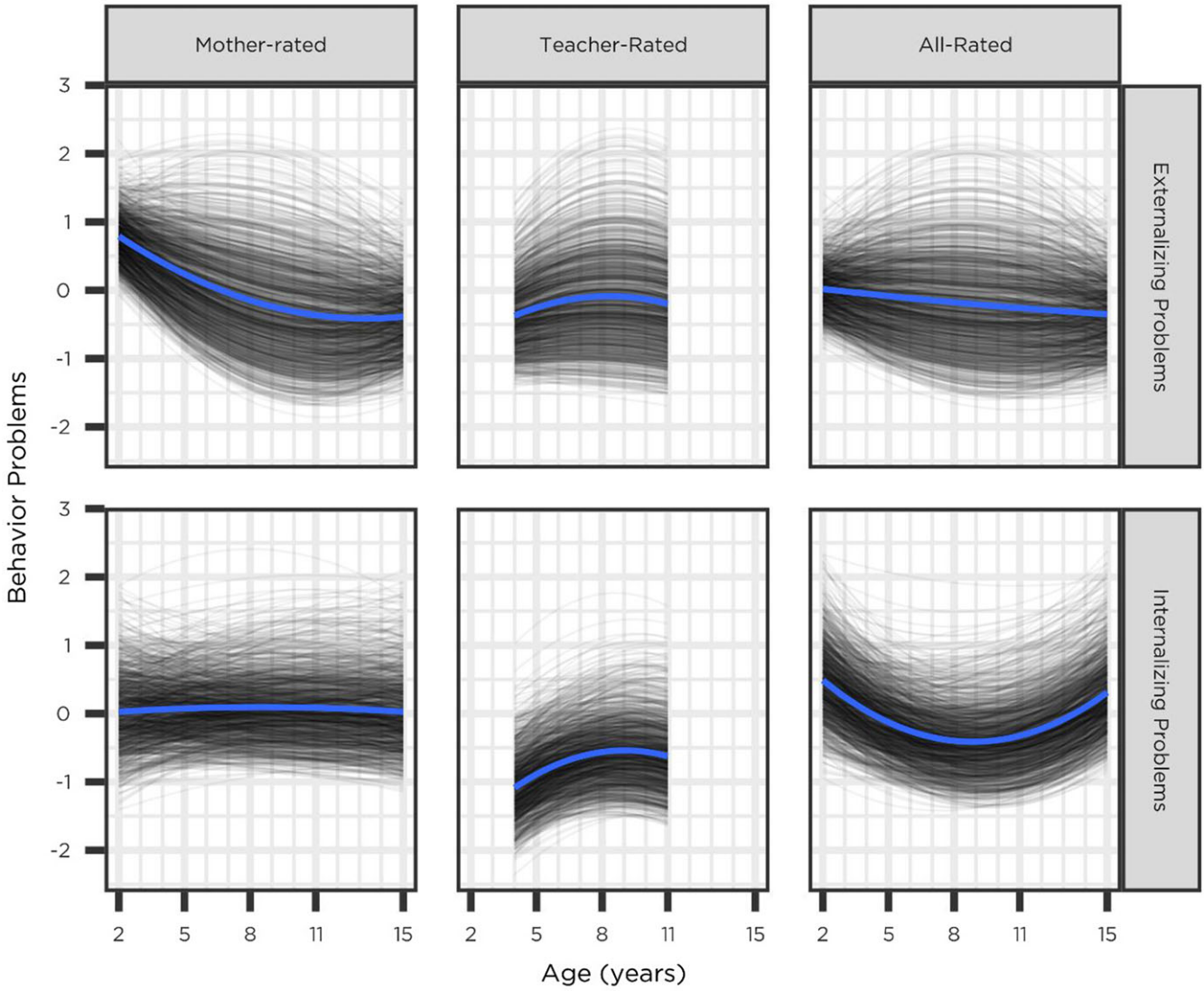


Figure 3. Children's model-implied growth curves of internalizing and externalizing problems by rater. *Note*. "All-Rated" refers to the model-implied ratings for the "average" rater.

slopes of externalizing problems but was strongly associated with higher intercepts and steeper declines in internalizing problems. When examining mother and caregiver report of negative emotionality separately, mother-reported negative emotionality was associated with higher intercepts of internalizing and externalizing problems, and steeper decreases in internalizing but not externalizing problems. Caregiver report had similar results as mother report, except that caregiver-reported negative emotionality predicted steeper declines in internalizing problems at a trend-level. Finally, when examining aggressive and delinquent behavior separately, associations with predictors and aggressive behavior factor did not substantively change from the primary analyses. Associations between predictors and delinquent behaviors also did not substantively change from the primary analyses.

Aim 3: Predicting general versus specific psychopathology at age 15

Given associations between the risk factors and ending levels of externalizing and internalizing problems at age 15, we examined the risk factors in relation to general versus specific psychopathology at age 15. The bifactor model with all items (externalizing: 33 items; internalizing: 33 items) did not fit. Thus, we modified the model by dropping items with low endorsement rates (externalizing: 7 items; internalizing: 0 items), leaving 26 externalizing items and 33 internalizing items. Then, we removed loadings onto the specific factors that were not significant and positive (externalizing: 8 items; internalizing: 1 item), leaving 18 externalizing items and 32 internalizing items that loaded onto the respective specific factor. The model fit well according to RMSEA (.051) and SRMR (.050) but did not fit well according to CFI (.796). Thus, we made model modifications.

We allowed item residuals to be correlated for which the modification index was large ($\Delta\chi 2 > 20$), indicating local non-independence of items, if the modification was also consistent with theory (i.e., both items were within the same domain). This led to 104 correlated residuals out of 649 possible within-domain correlated residuals. The model fit well according to RMSEA (.034) and SRMR (.047) and showed acceptable fit according to CFI (.907).

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

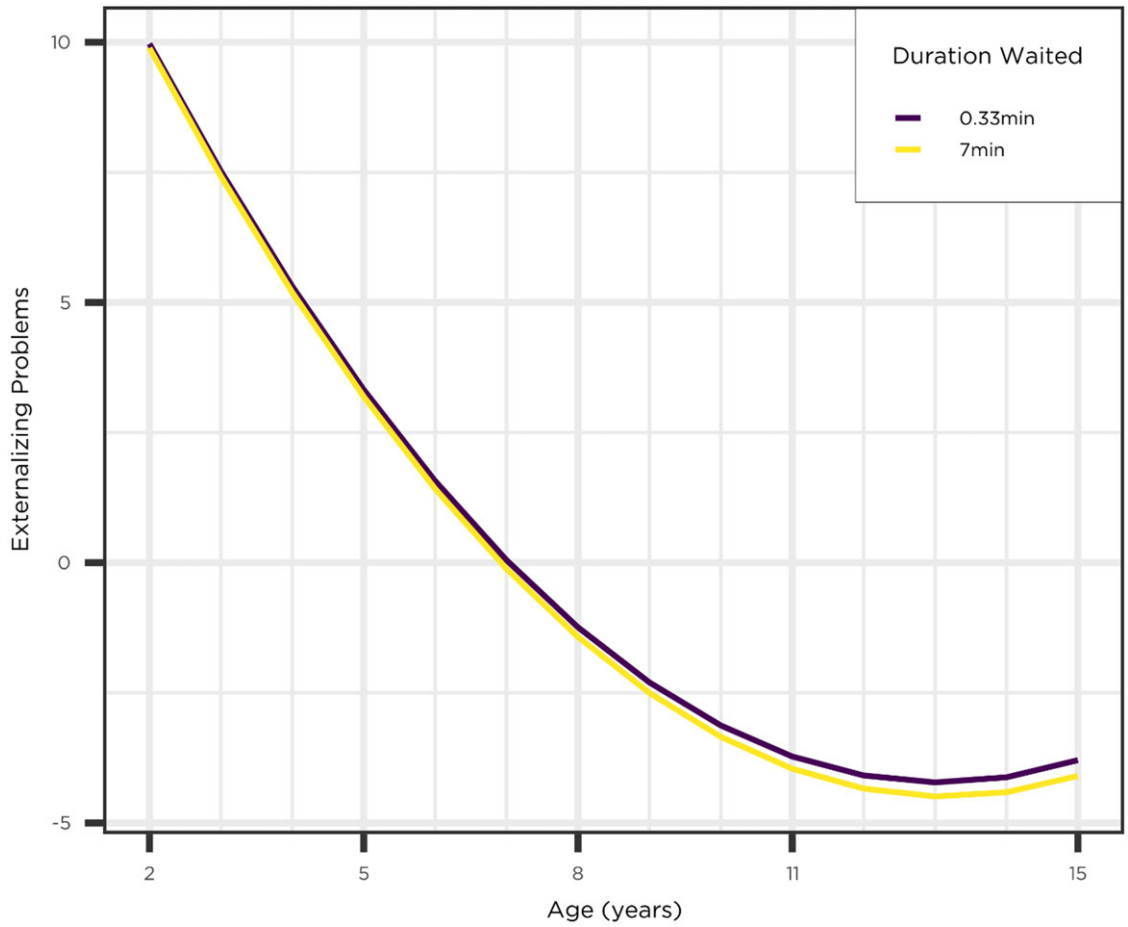


Figure 4. Children's model-implied externalizing problems trajectory as a function of duration of time waited in a self-imposed waiting task.

Items' factor loadings on the general factor and specific externalizing and internalizing factors are in Supplementary Table S9. The proportion of variance in ratings that was accounted for by the general factor (i.e., the explained common variance, ECV) was .69. The ECV of specific factors (ECVs) for the specific externalizing factor and specific internalizing factor was .09 and .22, respectively. Coefficient omega was .73, .21, and .42 for the general factor, specific externalizing factor, and specific internalizing factor, respectively. Thus, the reliability was acceptable for the general factor but was low for the specific factors, as evidenced by items' relatively weak factor loadings on the specific factors. Findings in relation to unique externalizing and unique internalizing problems should therefore be interpreted carefully.

Then, we added the predictors: negative emotionality and delay of gratification. Regression coefficients are in Supplementary Table S10. Negative emotionality was positively associated with the general factor and unique externalizing problems; it was not significantly associated with unique internalizing problems. Delay of gratification was negatively associated with the general factor but was positively associated with unique internalizing problems; it was not significantly associated with unique externalizing problems.

Then we added covariates. Regression coefficients are in Supplementary Table S11. The association between negative emotionality and the general factor remained significantly associated even when controlling for covariates. However, the association between negative emotionality and unique externalizing was attenuated to trend-level significance after controlling for covariates. The association between negative emotionality and unique internalizing problems remained nonsignificant. The association between delay of gratification and externalizing problems remained nonsignificant. However, delay of gratification was no longer significantly associated with the general factor or unique internalizing problems after controlling for covariates When compared to boys, girls showed lower ratings of general psychopathology, and higher ratings of unique externalizing and internalizing problems. When compared to non-African Americans, African Americans showed lower ratings of unique internalizing and externalizing problems and, to a trend-level, higher ratings of general psychopathology. When compared to non-Hispanics, Hispanics showed higher ratings of unique externalizing problems and, to a trend-level, general psychopathology. A higher income-to-needs ratio was associated with lower ratings of general psychopathology and higher ratings of unique internalizing problems, at a trend-level. Compared to mother report, fathers rated their child as showing greater general psychopathology, but lower unique externalizing problems unique internalizing problems, all at a trend-level. Compared to mother report, self-report showed higher ratings of general psychopathology and unique internalizing and externalizing problems.

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

Sensitivity analysis results: Bifactor models

Full results from the bifactor model sensitivity analyses are in Supplementary Appendix S10. When including early cognitive ability as an additional covariate in a bifactor framework, the association between delay of gratification and general and specific psychopathology remained nonsignificant. The association between negative emotionality and specific externalizing problems became statistically significant, after previously being associated at a trend-level. Regression coefficients, including early cognitive ability, are in Supplementary Table S12. When examining anger/frustration and fear, separately, as predictors of general and specific psychopathology, anger/frustration was not associated with unique internalizing problems, but predicted general psychopathology and unique externalizing problems more strongly than did fear. Fear was not significantly associated with general or specific psychopathology. Mother report of negative emotionality significantly predicted general and unique externalizing problems, and, at a trend-level, unique internalizing problems. Caregiver report of negative emotionality did not predict general or unique behavior problems. When examining predictors of unique aggressive and delinquent behavior, negative emotionality predicted unique delinquent behavior, but not unique aggressive behavior. Delay of gratification was not significantly associated with unique aggressive or delinquent behavior. Finally, when separately estimating bifactor models derived from mother report and self report, model fit according to CFI was poor for self-report but was acceptable for mother report. Predictors and covariates were added to these analyses and the full results are in Supplementary Appendix S10. However, we caution the interpretation of the findings from the self-report model due to poor model fit. For mother report, negative emotionality was positively associated with general psychopathology and unique externalizing problems, but was not associated with unique internalizing problems. Delay of gratification was negatively associated with general psychopathology and was positively associated with unique internalizing problems, but was not associated with unique externalizing problems.

Discussion

The present study had three goals: First, we aimed to describe people's trajectories of externalizing and internalizing problems from ages 2–15 years while accounting for heterotypic continuity via developmental scaling. Second, we examined whether negative emotionality and immediacy preference predicted the slopes from ages 2–15 and intercepts (i.e., ending levels) of people's trajectories of externalizing and internalizing problems at age 15. Third, we examined whether negative emotionality and immediacy preference predicted general psychopathology versus specific psychopathology at age 15.

Developmental trajectories

The present study is the first to chart the trajectories of externalizing and internalizing psychopathology concurrently across a lengthy developmental span while linking ratings from multiple raters. Similar to previous findings, ratings of externalizing problems decreased across development (e.g., Keiley et al., 2000; Leve et al., 2005; Petersen et al., 2015). Boys showed higher levels of externalizing problems than girls at age 15, but they did not differ in slopes.

For internalizing problems, results indicated an average decrease from ages 2 to 10 and an average increase from ages 10 to 15 years. An increase in internalizing symptoms at around 10–15 years of age maps onto pubertal development, and is expected given previous research (e.g., Hamlat et al., 2019). Mothers tended to endorse higher levels of both internalizing and externalizing problems (for the child/adolescent) compared to ratings by fathers, teachers, and afterschool caregivers; mothers tended to endorse lower rates of internalizing and externalizing problems compared to other caregivers and self-report. Mothers' ratings of internalizing problems tended to be relatively stable (on average) from 2 to 15 years of age, whereas teachers' ratings showed average increases from 4 to 8 years, and then were relatively stable (on average) from 8 to 11 years of age. Thus, developmentally scaled trajectories of internalizing and externalizing problems from ages 2 to 15 years were consistent with the expected age-related decrease in externalizing problems, and with the age-related increase in internalizing problems that aligns with pubertal development. Girls showed steeper increases in internalizing problems than boys and showed higher levels at age 15.

Reactive and control processes as predictors of trajectories

We examined only one dimension of reactive (negative emotionality) and one dimension of control (immediacy preference) processes as predictors of psychopathology trajectories. Higher informant-reported negative emotionality assessed as the average of mother and other caregiver report at 54 months was associated with steeper declines in externalizing and internalizing problems across development but remained associated with higher ending levels of externalizing and internalizing problems at age 15, controlling for covariates – though effect sizes were small. These results indicate that negative emotionality assessed in early childhood is associated with relatively high levels of internalizing and externalizing problems that endure across childhood to adolescence. In a sensitivity analysis examining the subscales of negative emotionality – fear and anger/frustration – in place of negative emotionality, they showed differential associations with trajectories of internalizing and externalizing problems. As would be expected based on prior studies (e.g., Crockett et al., 2018), anger/frustration was associated with externalizing problems but also internalizing problems, whereas fear was strongly associated with only internalizing problems. These results highlight the importance of considering multiple aspects of negative emotionality.

Immediacy preference assessed at 54 months was associated with higher ending levels of externalizing and internalizing problems at age 15, controlling for covariates. That is, a delay preference, i.e., longer wait times in a self-imposed waiting task, was associated with lower ratings of externalizing and internalizing problems at age 15 – though effect sizes were small. The association between immediacy preference and externalizing problems is consistent with previous research (e.g., Krueger et al., 1996; Mischel et al., 1989; Peake, 2017). However, immediacy preference was not associated with changes in externalizing or internalizing problems across development. These results point to the small, but potentially meaningful, association of specific facets of reactive and control processes with levels of internalizing and externalizing psychopathology across childhood and adolescence.

Predictors of mother-reported trajectories differed from predictors of trajectories derived from multiple raters, such that negative emotionality did not predict slopes of internalizing or externalizing problems, and delay of gratification was not significantly associated with lower intercepts of internalizing

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

problems. All other results were consistent with primary analyses. These results indicate that analyses including only mother ratings were just as well suited as when using multiple raters to detect that negative emotionality was associated with higher levels but not slopes of behavior problems, and that delay of gratification did not predict slopes across the developmental span. Of note, mother-reported negative emotionality was not able to detect differences in predictions of slopes, which provides further evidence for the utility of examining multiple perspectives of the child's behavior.

When accounting for early cognitive ability in a sensitivity analysis, the associations between delay of gratification and the intercepts of internalizing and externalizing problems were attenuated to non-significance and trend-level significance, respectively. Early cognitive ability such as verbal skills are thought to support executive function – whose deficits underlie externalizing problems – and delay of gratification. Given these findings, future studies should examine the role of early cognitive ability and executive functioning when examining reactive and control processes as predictors of psychopathology. Prior studies have shown that executive functioning may play a key role in the relations of reactive and control processes on behavior problems (Ursache et al., 2013; Watts et al., 2018).

General versus specific risk factors for psychopathology at age 15

In a bifactor model, negative emotionality at 54 months was significantly associated with general psychopathology and unique externalizing problems, but not unique internalizing problems at age 15 years. When accounting for covariates, the association between negative emotionality and unique externalizing was attenuated to trend-level significance. Sensitivity analysis elucidated that among symptoms of externalizing problems, delinquent behavior, but not aggressive behavior, appears to be most strongly related to negative emotionality. Furthermore, anger/frustration, not fearfulness, was strongly associated with general psychopathology and unique externalizing problems. A possible interpretation is that negative emotionality as assessed by the CBQ is overt and observable in nature, and thus may reflect more externalizing problems (Eisenberg et al., 2009). Furthermore, when accounting for covariates, the association between negative emotionality and general psychopathology remained significant. Interestingly, negative emotionality is often considered associated with internalizing psychopathology (Greene & Eaton, 2017), but our findings indicate that when accounting for general psychopathology, there was no association between childhood negative emotionality and later internalizing problems. The exception was that when examining mother- vs. caregiver-reported negative emotionality as predictors, only mother-reported negative emotionality was significantly associated with general psychopathology, unique externalizing problems, and, at a trend-level, unique internalizing problems.

One potential explanation for the consistently strong relation between negative emotionality and general psychopathology is that general psychopathology might be functionally interpreted as negative emotionality, such that temperamental negative emotionality may partially reflect a trait-like version of what is shared between internalizing and externalizing problems. An implication of these findings is that negative emotionality assessed at young ages may be a clinically relevant factor to consider when assessing psychopathology, as previous work has indicated (Forbes et al., 2019).

In a sensitivity analysis in which early cognitive ability was included as an additional covariate along with demographic characteristics, the association between negative emotionality and general psychopathology was attenuated to trend-level significance and the association with unique externalizing problems returned to statistical significance with the inclusion of early cognitive ability as a covariate. These findings suggest that other variables, potentially early cognitive ability, or demographic or environmental features, might partially account for the association between negative emotionality and psychopathology, which otherwise show strong associations.

Immediacy preference, i.e., shorter wait time, assessed at 54 months was associated with higher general psychopathology and with *lower* ratings of unique internalizing problems at age 15 years. Surprisingly, there was no association between immediacy preference and unique externalizing problems. Although the association between shorter wait times and lower ratings of internalizing problems was not hypothesized, previous research has observed such an association (Ho et al., 2022). The authors interpreted the association as possibly reflecting that choosing an immediate reward over a distal one is adaptive in reducing feelings of anxiety or depression. Alternatively, the finding could reflect that internalizing problems may be characterized by over-regulation (Murray & Kochanska, 2002). The finding that there was no association between immediacy preference and unique externalizing problems was contrary to what many studies have found (e.g., Ip et al., 2019; Krueger et al., 1996). However, these studies have not examined the association between immediacy preference and *unique* externalizing problems by controlling for the general factor of psychopathology. Nevertheless, previous studies have shown that after accounting for the general factor, self-regulation and executive functioning are not strongly associated with unique psychopathology (Bloemen et al., 2018). However, consistent with Watts et al. (2018), when controlling for covariates, the associations between immediacy preference and psychopathology outcomes were attenuated. This may reflect that individual differences in delay of gratification may be partially accounted for by demographic and background factors, including sex, culture, and socioeconomic status.

Watts et al. (2018) examined the association between immediacy preference and internalizing and externalizing psychopathology. However, Watts et al., assessed a summation of internalizing and externalizing symptoms, whereas we derived general psychopathology using bifactor models, which allowed us to distinguish shared and unique psychopathology. This difference is important because we found that the signs of the association of delay of gratification with general psychopathology (positive) versus unique internalizing problems (negative) were in the opposite direction. The sign difference in the association with immediacy preference would not have been observable with a total behavior problems score, as was used in Watts et al. (2018), or by examining total externalizing problems or total internalizing problems, as we separately examined in the present study. Others have argued against the use of behavior composites because they do not allow for partitioning variance between levels (Michaelson & Munakata, 2020). Therefore, accounting for the shared variance of internalizing and externalizing psychopathology may help identify potentially divergent effects of immediacy preference on various forms of psychopathology.

In a separate model including predictors and covariates together, immediacy preference was no longer associated with unique internalizing problems or the general factor. These results

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

indicate that much of the variance in the association of immediacy preference with internalizing and general psychopathology may be accounted for demographic and socioeconomic factors. This result was similar to those found by Watts et al. (2018), suggesting that immediacy preference may partially reflect environmental and dispositional processes. Although Watts et al. (2018) focused primarily on children of mothers without college degrees, we leveraged a larger sample while linking scores from multiple informants to obtain a more robust estimate.

The numerous sensitivity analyses (growth curve models: Supplementary Appendix S9; bifactor models: Supplementary Appendix S10) indicated utility in examining subdimensions of constructs. Negative emotionality, in particular anger/frustration, was a robust predictor of externalizing problems across many sensitivity analyses. It was also useful to examine subdimensions of externalizing problems, including delinquent versus aggressive behavior; negative emotionality was more strongly associated with delinquent than aggressive behavior. Furthermore, when examining the association between temperament and psychopathology, accounting for covariates such as early cognitive ability provided potential alternative explanations for future work to consider. The effect sizes of findings in the sensitivity analyses were generally small in magnitude ($\beta = -.01$–$.23$); therefore, we caution that smaller samples might be underpowered to detect these associations.

Explained common variance

Partitioning unique and general psychopathology also allows estimating the proportion of variance in psychopathology ratings that was accounted for by general and specific psychopathology factors in the sample. The general factor accounted for ~69% of the reliable variance in ratings of psychopathology at age 15 while specific externalizing problems accounted for ~9% and specific internalizing problems accounted for ~22%. This indicates that the general factor was substantially stronger than the specific factors at age 15. Because internalizing problems increased whereas externalizing problems decreased from childhood to adolescence, it is unsurprising that the specific internalizing factor had a larger ECVs compared to the specific externalizing factor. These results are important to consider, because predictors of unique externalizing psychopathology at age 15 may not be as robust given that there is little strength in ratings of unique externalizing problems at this age. By contrast, the unique internalizing and general factors were much stronger at age 15.

Comparison to prior studies using SECCYD dataset

Prior researchers have studied internalizing and externalizing problems using the same dataset as the present study. Fanti & Heinrich (2010) examined pure and co-occurring mother-reported internalizing and externalizing problems from age 2 to 12. In their study, "pure" referred to a latent class where individuals endorsed high levels in externalizing or internalizing problems but low levels in the other. "Co-occurring" referred to a latent class where individuals endorsed high levels of both. This study found that children with a difficult temperament were more likely to be classified in a co-occurring group compared to "pure" externalizing or internalizing problems, which is consistent with our findings with general psychopathology. They also found similar results where early cognitive development difficulties were associated most strongly with pure chronic externalizing problems compared to other groups, and were less likely to be associated with pure internalizing problems (Fanti & Henrich, 2010).

Cao et al. (2021) examined various inhibitory control tasks, including immediacy preference, as mediators between early tobacco smoke exposure and an average composite of mother- and father-reported internalizing and externalizing problems at the earlier time point of 6th grade. They found that delay of gratification did not predict internalizing or externalizing problems in 6th grade, which differed from our results. Wang & Liu (2021) examined internalizing and externalizing problems at multiple timepoints of 1st, 3rd, 4th, 5th, and 6th grade from teacher ratings. Unlike our linking approach, they estimated a latent factor score for intercepts and slopes of internalizing and externalizing problems composed of scores from each grade as manifest variables. In this study, executive functioning and social competence measured at 1st grade were the predictors of interest (Wang & Liu, 2021). The researchers found that poor executive functioning and poor social competence both predicted higher intercepts of internalizing and externalizing problems (Wang & Liu, 2021).

In comparison to prior studies, the present study uses multidimensional linking to chart trajectories across multiple raters and uses robust methods of bifactor modeling to evaluate predictors of general and specific psychopathology. Furthermore, like Cao et al. (2021) who examined "hot" and "cool" inhibitory control as risk factors, our study examines different facets of a given construct, i.e., reactive and control processes.

To our knowledge, there are two prior studies, Deutz et al. (2020) and McElroy et al. (2018), that examined factor analytic models of general psychopathology using the SECCYD dataset. Deutz et al. (2020) found that immediacy preference did not predict separately analyzed mother- or self-reported general psychopathology after controlling for covariates, similar to our study; however, we also included father report, and we linked scores across raters. Furthermore, we examined these questions for general and specific psychopathology with bifactor modeling. McElroy et al. (2018) examined phenotypic stability in the general and specific factors across all measurements of mother-reported psychopathology from ages 2 to 14, but did not examine predictors of general versus specific psychopathology.

Strengths

The present study had key strengths. First, we examined children's internalizing and externalizing problems from multiple informants. Second, we examined a lengthy span of development in a large and diverse sample. Third, we used a robust IRT approach to developmental scaling to estimate trajectories of internalizing and externalizing psychopathology on the same scale using ratings at different ages, from numerous raters, and from different measures. Fourth, we examined aspects of reactive and control processes – negative emotionality and immediacy preference – along with demographic characteristics in association with general versus specific psychopathology. The present study provides novel and important contributions to the theory of reactive and control processes of temperament as predictors and the measurement and analysis of psychopathology across a lengthy developmental span.

Limitations

The present study also had weaknesses. First, we are unable to make causal inferences due to the observational design of the study. Although we examined negative emotionality and immediacy

https://doi.org/10.1017/S0954579424000713 Published online by Cambridge University Press

preference at 54 months in relation to psychopathology at age 15, we are unable to rule out the possibility of the reverse direction of effect, or the possibility of unmeasured third variables. Second, the reliability of the specific psychopathology factors was relatively low, which may have hindered our ability to detect associations with unique externalizing or internalizing problems. It will be important for future work to identify ways of assessing specific psychopathology that yield greater internal consistency.

Third, the linking approach used in the study assumes that item parameters and factor scores are linearly related across measures, raters, and periods of measurements. However, the results indicated that the approach to linking was successful across measures, raters, and timepoints (see Supplement Figures S2–S5). The linking assumption of linearity is between item parameters at *adjacent* ages; the linking does not assume that the change is linear across the entire age span – the changes can be larger or smaller at various ages as needed to adjust for differences. Our developmental scaling approach linked scores across the age- and rater-common items at the aggregate construct level rather than at the item level. Thus, our approach would not be well-suited for interpreting scores for any individual behavior/item across time or raters. Moreover, our approach would also not be appropriate if there is not a common factor (e.g., externalizing problems) that influences scores across the home and school contexts. Our approach may also potentially overlook meaningful qualitative differences that can occur from one year to the next (e.g., transition from elementary to middle school; Shi & Ettekal, 2021).

Fourth, negative emotionality and psychopathology were both derived from the same method (i.e., informant report), whereas immediacy preference was assessed using a laboratory task. Shared method variance might inflate associations between ratings of negative emotionality and psychopathology, when compared to more context-dependent performance on an immediacy preference task (Brock et al., 2014; Makol et al., 2020). Finally, there are potential limitations to the sample. Data collection for the NICHD SECCYD study began approximately 30 years ago, which may impact the relevance of the findings to current understanding of temperament and psychopathology. Moreover, the sample is not nationally representative in terms of race or ethnicity, thus limiting the potential generalizability of the findings to the U.S. population. However, the sample is diverse economically and geographically.

Conclusion

The findings from the present study extend previous work (Petersen & LeBeau, 2022) that has linked ratings of psychopathology from multiple informants across a wide age range. Very few studies have used developmental scaling to study the development of psychopathology across a lengthy span (Petersen et al., 2018; Petersen & LeBeau, 2022). The present study demonstrates a useful approach to link ratings of internalizing and externalizing psychopathology across time and across raters. The approach was valuable because it allowed us to chart children's development in internalizing and externalizing problems across a lengthy developmental span and to examine early risk factors of their trajectories. Additionally, findings demonstrate the importance of partitioning general versus specific psychopathology.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0954579424000713.

References

Achenbach, T. M. (1991a). *Manual for the Teacher's Report Form and 1991 profile*. University of Vermont, Department of Psychiatry. https://aseba.org/.

Achenbach, T. M. (1991b). *Manual for the Youth Self-Report and 1991 profile*, University of Vermont, Department of Psychiatry. https://aseba.org/.

Achenbach, T. M. (1992). *Manual for the Child Behavior Checklist/2-3 and 1992 profile: Profile for boys and girls*. University of Vermont, Department of Psychiatry. https://aseba.org/.

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms and profiles (Vol. 30)*. University of Vermont, University of Vermont, Department of Psychiatry. https://aseba.org/.

Achenbach, T. M., & Rescorla, L. A. (2001, *Manual for the ASEBA school-age forms & profiles*. University of Vermont, Department of Psychiatry. https://aseba.org/.

Aldao, A., Gee, D. G., De Los Reyes, A., & Seager, I. (2016). Emotion regulation as a transdiagnostic factor in the development of internalizing and externalizing psychopathology: Current and future directions. *Development and Psychopathology*, *28*(4pt1), 927–946. https://doi.org/10.1017/S0954579416000638

Avenevoli, S., & Steinberg, L. (2001). The continuity of depression across the adolescent transition. *Advances in Child Development and Behavior*, *28*, 139–173. https://doi.org/10.1016/s0065-2407(02)80064-7

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, J. E., Schermerhorn, A. C., & Petersen, I. T. (2014). Temperament concepts in developmental psychopathology. In M. Lewis, & K. D. Rudolph (Eds.), *Handbook of developmental psychopathology* (pp. 311–329). Springer US, https://doi.org/10.1007/978-1-4614-9608-3_16

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Bjorklund, D. F., & Kipp, K. (1996). Parental investment theory and gender differences in the evolution of inhibition mechanisms. *Psychological Bulletin*, *120*(2), 163–188.

Bloemen, A., Oldehinkel, A. J., Laceulle, O. M., Ormel, J., Rommelse, N., & Hartman, C. A. (2018). The association between executive functioning and psychopathology: General or specific? *Psychological Medicine*, *48*(11), 1787–1794. https://doi.org/10.1017/S0033291717003269

Brandes, C. M., Herzhoff, K., Smack, A. J., & Tackett, J. L. (2019). The p factor and the n factor: Associations between the general factors of psychopathology and neuroticism in children. *Clinical Psychological Science*, *7*(6), 1266–1284. https://doi.org/10.1177/2167702619859332

Briggs-Gowan, M. J., Carter, A. S., Bosson-Heenan, J., Guyer, A. E., & Horwitz, S. M. (2006). Are infant-toddler social-emotional and behavioral problems transient? *Journal of the American Academy of Child & Adolescent Psychiatry*, *45*(7), 849–858. https://doi.org/10.1097/01.chi.0000220849.48650.59

Brock, L. L., Rimm-Kaufman, S. E., & Wanless, S. B. (2014). Delay of gratification in first grade: The role of instructional context. *Learning and Individual Differences*, *29*, 81–88. https://doi.org/10.1016/j.lindif.2013.10.012

Campbell, S. B., & von Stauffenberg, C. (2009). Delay and inhibition as early predictors of ADHD symptoms in third grade. *Journal of Abnormal Child Psychology*, *37*(1), 1–15. https://doi.org/10.1007/s10802-008-9270-4

Cao, H., Liang, Y., & Zhou, N. (2021). Early tobacco smoke exposure, preschool cool/hot inhibitory control, and young adolescents' externalizing/internalizing problems. *Journal of Family Psychology*, *35*(3), 311–323.

Casey, B. J., Somerville, L. H., Gotlib, I. H., Ayduk, O., Franklin, N. T., Askren, M. K., Jonides, J., Berman, M. G., Wilson, N. L., Teslovich, T., Glover, G., Zayas, V., Mischel, W., & Shoda, Y. (2011). Behavioral and neural correlates of delay of gratification 40 years later. *Proceedings of the National Academy of Sciences of The United States of America*, *108*(36), 14998–15003. https://doi.org/10.1073/pnas.1108561108

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure

of psychiatric disorders? *Clinical Psychological Science*, *2*(2), 119–137. https://10.1177/2167702613497473

Castellanos-Ryan, N., Brière, F. N., O'Leary-Barrett, M., Banaschewski, T., Bokde, A., Bromberg, U., Büchel, C., Flor, H., Frouin, V., Gallinat, J., Garavan, H., Martinot, J.-L., Nees, F., Paus, T., Pausova, Z., Rietschel, M., Smolka, M. N., Robbins, T. W., Whelan, R., Schumann, G., Conrod, P., The IMAGEN Consortium (2016). The structure of psychopathology in adolescence and its common personality and cognitive correlates. *Journal of Abnormal Psychology*, *125*(8), 1039–1052.

Cervin, M., Norris, L. A., Ginsburg, G., Gosch, E. A., Compton, S. N., Piacentini, J., Albano, A. M., Sakolsky, D., Birmaher, B., Keeton, C., Storch, E. A., & Kendall, P. C. (2021). The p factor consistently predicts long-term psychiatric and functional outcomes in anxiety-disordered youth. *Journal of the American Academy of Child & Adolescent Psychiatry*, *60*(7), 902–912.e5. https://doi.org/10.1016/j.jaac.2020.08.440

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, F. R., & Jaffee, S. R. (2015). The heterogeneity in the development of homotypic and heterotypic antisocial behavior. *Journal of Developmental and Life-Course Criminology*, *1*(3), 269–288. https://doi.org/10.1007/s40865-015-0012-3

Choate, A. M., Bornovalova, M. A., Hipwell, A. E., Chung, T., & Stepp, S. D. (2022). The general psychopathology factor (p) from adolescence to adulthood: Exploring the developmental trajectories of p using a multi-method approach. *Development and Psychopathology*, *1-19*(4), 1775–1793. https://doi.org/10.1017/S0954579422000463

Cicchetti, D., & Rogosch, F. A. (2002). A developmental psychopathology perspective on adolescence. *Journal of Consulting and Clinical Psychology*, *70*(1), 6–20. https://doi.org/10.1037/0022-006X.70.1.6

Clark, D. A., Hicks, B. M., Angstadt, M., Rutherford, S., Taxali, A., Hyde, L., Weigard, A. S., Heitzeg, M. M., & Sripada, C. (2021). The general factor of psychopathology in the Adolescent Brain Cognitive Development (ABCD) study: A comparison of alternative modeling approaches. *Clinical Psychological Science*, *9*(2), 169–182. https://doi.org/10.1177/2167702620959317

Crockett, L. J., Wasserman, A. M., Rudasill, K. M., Hoffman, L., & Kalutskaya, I. (2018). Temperamental anger and effortful control, teacher-child conflict, and externalizing behavior across the elementary school years. *Child Development*, *89*(6), 2176–2195.

Cuhadar, I., & Kalkan, Ö.K. (2023). Performance of estimation methods in bifactor models with ordered categorical data. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(2), 329–339. https://doi.org/10.1080/10705511.2023.2247567

Damme, K. S. F., Norton, E. S., Briggs-Gowan, M. J., Wakschlag, L. S., & Mittal, V. A. (2022). Developmental patterning of irritability enhances prediction of psychopathology in preadolescence: Improving RDoC with developmental science. *Journal of Psychopathology and Clinical Science*, *131*(6), 556–566. https://doi.org/10.1037/abn0000655

De Los Reyes, A., Bunnell, B. E., & Beidel, D. C. (2013). Informant discrepancies in adult social anxiety disorder assessments: Links with contextual variations in observed behavior. *Journal of Abnormal Psychology*, *122*(2), 376–386. https://doi.org/10.1037/a0031150

De Los Reyes, A., Henry, D. B., Tolan, P. H., & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Child Psychology*, *37*(5), 637–652. https://doi.org/10.1007/s10802-009-9307-3

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*(4), 483–509. https://doi.org/10.1037/0033-2909.131.4.483

De Los Reyes, A., & Makol, B. A. (2021). Interpreting convergences and divergences in multi-informant, multimethod assessment. In Joni L. Mihura (Eds.), *The oxford handbook of personality and psychopathology assessment* (2nd. ed.). Oxford Academic,

De Los Reyes, A., & Makol, B. A. (2022). Informant reports in clinical assessment. In G. J. G. Asmundson (Eds.), *Comprehensive clinical psychology* (2nd ed. pp. 105–122). Elsevier, https://doi.org/10.1016/B978-0-12-818697-8.00113-8

De Pauw, S. S. W., & Mervielde, I. (2010). Temperament, personality and developmental psychopathology: A review based on the conceptual dimensions underlying childhood traits. *Child Psychiatry & Human Development*, *41*(3), 313–329. https://doi.org/10.1007/s10578-009-0171-8

Deutz, M. H. F., Geeraerts, S. B., Belsky, J., Deković, M., van Baar, A. L., Prinzie, P., & Patalay, P. (2020). General psychopathology and dysregulation profile in a longitudinal community sample: Stability, antecedents and outcomes. *Child Psychiatry and Human Development*, *51*(1), 114–126. https://doi.org/10.1007/s10578-019-00916-2

Egan, S. J., Wade, T. D., & Shafran, R. (2011). Perfectionism as a transdiagnostic process: A clinical review. *Clinical Psychology Review*, *31*(2), 203–212.

Eisenberg, N., Fabes, R. A., Guthrie, I. K., Murphy, B. C., Maszk, P., Holmgren, R., & Suh, K. (1996). The relations of regulation and emotionality to problem behavior in elementary school children. *Development and Psychopathology*, *8*(1), 141–162. https://doi.org/10.1017/S095457940000701X

Eisenberg, N., Guthrie, I. K., Fabes, R. A., Shepard, S., Losoya, S., Murphy, B., Jones, S., Poulin, R., & Reiser, M. (2000). Prediction of elementary school children's externalizing problem behaviors from attentional and behavioral regulation and negative emotionality. *Child Development*, *71*(5), 1367–1382. https://doi.org/10.1111/1467-8624.00233

Eisenberg, N., Sadovsky, A., Spinrad, T. L., Fabes, R. A., Losoya, S. H., Valiente, C., Reiser, M., Cumberland, A., & Shepard, S. A. (2005). The relations of problem behavior status to children's negative emotionality, effortful control, and impulsivity: Concurrent relations and prediction of change. *Developmental Psychology*, *41*(1), 193–211. https://doi.org/10.1037/0012-1649.41.1.193

Eisenberg, N., Valiente, C., Spinrad, T. L., Cumberland, A., Liew, J., Reiser, M., Zhou, Q., & Losoya, S. H. (2009). Longitudinal relations of children's effortful control, impulsivity, and negative emotionality to their externalizing, internalizing, and co-occurring behavior problems. *Developmental Psychology*, *45*(4), 988–1008. https://doi.org/10.1037/a0016213

Fanti, K. A., & Henrich, C. C. (2010). Trajectories of pure and co-occurring internalizing and externalizing problems from age 2 to age 12: Findings from the National Institute of Child Health and Human Development Study of Early Child Care. *Developmental Psychology*, *46*(5), 1159–1175. https://doi.org/10.1037/a0020659.

Forbes, M. K., Rapee, R. M., & Krueger, R. F. (2019). Opportunities for the prevention of mental disorders by reducing general psychopathology in early childhood. *Behaviour Research and Therapy*, *119*, 103411–103411. https://doi.org/10.1016/j.brat.2019.103411

Forbes, M. K., Tackett, J. L., Markon, K. E., & Krueger, R. F. (2016). Beyond comorbidity: Toward a dimensional and hierarchical approach to understanding psychopathology across the life span. *Development and Psychopathology*, *28*(4pt1), 971–986. https://doi.org/10.1017/S0954579416000651

Fox, N. A. (1989). Psychophysiological correlates of emotional reactivity during the first year of life. *Developmental Psychology*, *25*(3), 364–372. https://doi.org/10.1037/0012-1649.25.3.364

Gluschkoff, K., Jokela, M., & Rosenström, T. (2019). The general psychopathology factor: Structural stability and generalizability to within-individual changes. *Frontiers in Psychiatry*, *10*, 594. https://doi.org/10.3389/fpsyt.2019.00594

Greene, A. L., & Eaton, N. R. (2017). The temporal stability of the bifactor model of comorbidity: An examination of moderated continuity pathways. *Comprehensive Psychiatry*, *72*, 74–82. https://doi.org/10.1016/j.comppsych.2016.09.010

Hamlat, E. J., Snyder, H. R., Young, J. F., & Hankin, B. L. (2019). Pubertal timing as a transdiagnostic risk for psychopathology in youth. *Clinical Psychological Science*, *7*(3), 411–429. https://doi.org/10.1177/2167702618810518

Hankin, B. L., Davis, E. P., Snyder, H., Young, J. F., Glynn, L. M., & Sandman, C. A. (2017). Temperament factors and dimensional, latent bifactor models of child psychopathology: Transdiagnostic and specific

associations in two youth samples. *Psychiatry Research*, *252*, 139–146. https://doi.org/10.1016/j.psychres.2017.02.061

Hartley, A. G., Zakriski, A. L., & Wright, J. C. (2011). Probing the depths of informant discrepancies: Contextual influences on divergence and convergence. *Journal of Clinical Child & Adolescent Psychology*, *40*(1), 54–66. https://doi.org/10.1080/15374416.2011.533404

Hawes, M. T., Carlson, G. A., Finsaas, M. C., Olino, T. M., Seely, J. R., & Klein, D. N. (2020). Dimensions of irritability in adolescents: Longitudinal associations with psychopathology in adulthood. *Psychological Medicine*, *50*(16), 2759–2767. https://doi.org/10.1017/S0033291719002903.

Ho, H., Dang, H. M., Odum, A. L., DeHart, W. B., & Weiss, B. (2022). Sooner is better: Longitudinal relations between delay discounting, and depression and anxiety symptoms among Vietnamese adolescents. *Research on Child and Adolescent Psychopathology*, *51*(1), 133–147. https://doi.org/10.1007/s10802-022-00959-5.https://doi.org/10.1007/s10802-022-00959-5

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huber, P. J. The behavior of maximum likelihood estimates under nonstandard conditions. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, University of California Press, 1967

Ip, K. I., Jester, J. M., Sameroff, A., & Olson, S. L. (2019). Linking Research Domain Criteria (RDoC) constructs to developmental psychopathology: The role of self-regulation and emotion knowledge in the development of internalizing and externalizing growth trajectories from ages 3 to 10. *Development and Psychopathology*, *31*(4), 1557–1574. https://doi.org/10.1017/S0954579418001323

Jonas, K., & Kochanska, G. (2018). An imbalance of approach and effortful control predicts externalizing problems: Support for extending the dual-systems model into early childhood. *Journal of Abnormal Child Psychology*, *46*(8), 1573–1583. https://doi.org/10.1007/s10802-018-0400-3

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. *semTools: Useful tools for structural equation modeling*. [Computer software], 2022. https://CRAN.R-project.org/package=semTools.

Keiley, M. K., Bates, J. E., Dodge, K. A., & Pettit, G. S. (2000). A cross-domain growth analysis: Externalizing and internalizing behaviors during 8 years of childhood. *Journal of Abnormal Child Psychology*, *28*(2), 161–179. https://doi.org/10.1023/a:1005122814723

Kenyon, D. M., MacGregor, D., Li, D., & Cook, H. G. (2011). Issues in vertical scaling of a K-12 English language proficiency test. *Language Testing*, *28*(3), 383–400. https://doi.org/10.1177/0265532211404190

Kidd, C., Palmeri, H., & Aslin, R. N. (2013). Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. *Cognition*, *126*(1), 109–114. https://doi.org/10.1016/j.cognition.2012.08.004

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed., pp. xxvi, 566). https://doi.org/10.1007/978-1-4939-0317-7

Krueger, R. F., Caspi, A., Moffitt, T. E., White, J., & Stouthamer-Loeber, M. (1996). Delay of gratification, psychopathology, and personality: Is low self-control specific to externalizing problems? *Journal of Personality*, *64*(1), 107–129. https://doi.org/10.1111/j.1467-6494.1996.tb00816.x

Krueger, R. F., & Eaton, N. R. (2015). Transdiagnostic factors of mental disorders. *World Psychiatry*, *14*(1), 27–29. https://doi.org/10.1002/wps.20175

Krueger, R. F., Hobbs, K. A., Conway, C. C., Dick, D. M., Dretsch, M. N., Eaton, N. R., Forbes, M. K., Forbush, K. T., Keyes, K. M., Latzman, R. D., Michelini, G., Patrick, C. J., Sellbom, M., Slade, T., South, S C., Sunderland, M., Tackett, J., Waldman, I., Waszczuk, M. A., Wright, A. G. C., Zald, D. H., Watson, D., Kotov, R., HiTOP Utility Workgroup (2021). Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): II. Externalizing superspectrum. *World Psychiatry*, *20*(2), 171–193. https://doi.org/10.1002/wps.20844

Kwon, K., Kim, E. M., & Sheridan, S. M. (2012). A contextual approach to social skills assessment in the peer group: Who is the best judge? *School Psychology Quarterly*, *27*(3), 121–133. https://doi.org/10.1037/a0028696

Lahey, B. B. (2009). Public health significance of neuroticism. *American Psychologist*, *64*(4), 241–256.

Lahey, B. B., Applegate, B., Hakes, J. K., Zald, D. H., Hariri, A. R., & Rathouz, P. J. (2012). Is there a general factor of prevalent psychopathology during adulthood? *Journal of Abnormal Psychology*, *121*(4), 971–977. https://doi.org/10.1037/a0028355

Lahey, B. B., Moore, T. M., Kaczkurkin, A. N., & Zald, D. H. (2021). Hierarchical models of psychopathology: Empirical support, implications, and remaining issues. *World Psychiatry*, *20*(1), 57–63. https://doi.org/10.1002/wps.20824

Leaberry, K. D., Rosen, P. J., Slaughter, K. E., Reese, J., & Fogleman, N. D. (2019). Temperamental negative affect, emotion-specific regulation, and concurrent internalizing and externalizing pathology among children with ADHD. *ADHD Attention Deficit and Hyperactivity Disorders*, *11*(3), 311–324. https://doi.org/10.1007/s12402-019-00294-8

Lemery, K. S., Essex, M. J., & Smider, N. A. (2002). Revealing the relation between temperament and behavior problem symptoms by eliminating measurement confounding: Expert ratings and factor analyses. *Child Development*, *73*(3), 867–882. https://doi.org/10.1111/1467-8624.00444

Leve, L. D., Kim, H. K., & Pears, K. C. (2005). Childhood temperament and family environment as predictors of internalizing and externalizing trajectories from ages 5 to 17. *Journal of Abnormal Child Psychology*, *33*(5), 505–520. https://doi.org/10.1007/s10802-005-6734-7

Lilienfeld, S. O. (2003). Comorbidity between and within childhood externalizing and internalizing disorders: Reflections and directions. *Journal of Abnormal Child Psychology*, *31*(3), 285–291. https://doi.org/10.1023/a:1023229529866

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(1), 59–72. https://doi.org/10.1207/s15328007sem1301_3

Loeber, R., Farrington, D. P., Stouthamer-Loeber, M., & Van Kammen, W. B. (1998). Multiple risk factors for multiproblem boys: Co-occurrence of delinquency, substance use, attention deficit, conduct problems, physical aggression, covert behavior, depressed mood, and shy/withdrawn behavior. In R. Jessor (Eds.), *New perspectives on adolescent risk behavior* (pp. 90–149). Cambridge University Press. https://doi.org/10.1017/CBO9780511571138.005

Lynch, S. J., Sunderland, M., Newton, N. C., & Chapman, C. (2021). A systematic review of transdiagnostic risk and protective factors for general and specific psychopathology in young people. *Clinical Psychology Review*, *87*, 102036. https://doi.org/10.1016/j.cpr.2021.102036

Makol, B. A., Youngstrom, E. A., Racz, S. J., Qasmieh, N., Glenn, L. E., & De Los Reyes, A. (2020). Integrating multiple informants' reports: How conceptual and measurement models may address long-standing problems in clinical decision-making. *Clinical Psychological Science*, *8*(6), 953–970. https://doi.org/10.1177/2167702620924439

Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological Bulletin*, *137*(5), 856–879. https://doi.org/10.1037/a0023678

Martel, M. M., Markon, K., & Smith, G. T. (2017). Research review: Multi-informant integration in child and adolescent psychopathology diagnosis. *Journal of Child Psychology and Psychiatry*, *58*(2), 116–128. https://doi.org/10.1111/jcpp.12611

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*(1), 577–605. https://doi.org/10.1146/annurev.psych.60.110707.163612

McElroy, E., Belsky, J., Carragher, N., Fearon, P., & Patalay, P. (2018). Developmental stability of general and specific factors of psychopathology from early childhood to adolescence: Dynamic mutualism or p-differentiation? *Journal of Child Psychology and Psychiatry*, *59*(6), 667–675. https://doi.org/10.1111/jcpp.12849

McLaughlin, K. A., & Nolen-Hoeksema, S. (2011). Rumination as a transdiagnostic factor in depression and anxiety. *Behaviour Research and Therapy*, *49*(3), 186–193. https://doi.org/10.1016/j.brat.2010.12.006

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, *106*(1), 3–19. https://doi.org/10.1037/0033-295X.106.1.3

Michaelson, L. E., & Munakata, Y. (2020). Same data set, different conclusions: Preschool delay of gratification predicts later behavioral outcomes in a

preregistered study. *Psychological Science*, *31*(2), 193–201. https://doi.org/10.1177/0956797619896270

Mikolajewski, A. J., Allan, N. P., Hart, S. A., Lonigan, C. J., & Taylor, J. (2013). Negative affect shares genetic and environmental influences with symptoms of childhood internalizing and externalizing disorders. *Journal of Abnormal Child Psychology*, *41*(3), 411–423. https://doi.org/10.1007/s10802-012-9681-0

Miller, J. L., Vaillancourt, T., & Boyle, M. H. (2009). Examining the heterotypic continuity of aggression using teacher reports: Results from a national Canadian study. *Social Development*, *18*(1), 164–180. https://doi.org/10.1111/j.1467-9507.2008.00480.x

Min, K.-S. (2007). Evaluation of linking methods for multidimensional irt calibrations. *Asia Pacific Education Review*, *8*(1), 41–55. https://doi.org/10.1007/BF03025832

Mischel, W., & Ebbesen, E. B. (1970). Attention in delay of gratification. *Journal of Personality and Social Psychology*, *16*(2), 329–337. https://doi.org/10.1037/h0029815

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*(2), 246–268. https://doi.org/10.1037/0033-295x.102.2.246

Mischel, W., Shoda, Y., & Rodriguez, M. (1989). Delay of gratification in children. *Science*, *244*(4907), 933–938. https://doi.org/10.1126/science.2658056

Moran, L. R., Lengua, L. J., & Zalewski, M. (2013). The interaction between negative emotionality and effortful control in early social-emotional development. *Social Development*, *22*(2), 340–362. https://doi.org/10.1111/sode.12025

Murayama, K., Pekrun, R., Lichtenfeld, S., & Vom Hofe, R. (2013). Predicting long-term growth in students' mathematics achievement: The unique contributions of motivation and cognitive strategies. *Child Development*, *84*(4), 1475–1490. https://doi.org/10.1111/cdev.12036

Muris, P., & Ollendick, T. H. (2005). The role of temperament in the etiology of child psychopathology. *Clinical Child and Family Psychology Review*, *8*(4), 271–289. https://doi.org/10.1007/s10567-005-8809-y

Murray, A. L., Eisner, M., & Ribeaud, D. (2016). The development of the general factor of psychopathology 'p factor' through childhood and adolescence. *Journal of Abnormal Child Psychology*, *44*(8), 1573–1586. https://doi.org/10.1007/s10802-016-0132-1

Murray, J., & Farrington, D. P. (2010). Risk factors for conduct disorder and delinquency: Key findings from longitudinal studies. *The Canadian Journal of Psychiatry*, *55*(10), 633–642.

Murray, K. T., & Kochanska, G. (2002). Effortful control: Factor structure and relation to externalizing and internalizing behaviors. *Journal of Abnormal Child Psychology*, *30*(5), 503–514. https://doi.org/10.1023/a:1019821031523

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

NICHD Early Child Care Research Network 2005). *Child care and child development: Results from the NICHD study of early child care and youth development*. Guilford Press.

Olino, T. M., Dougherty, L. R., Bufferd, S. J., Carlson, G. A., & Klein, D. N. (2014). Testing models of psychopathology in preschool-aged children using a structured interview-based assessment. *Journal of Abnormal Child Psychology*, *42*(7), 1201–1211. https://doi.org/10.1007/s10802-014-9865-x

Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement*, *37*(4), 357–373. http://www.jstor.org/stable/1435246

Patterson, G. R. (1993). Orderly change in a stable world: The antisocial trait as a chimera. *Journal of Consulting and Clinical Psychology*, *61*(6), 911–919. https://doi.org/10.1037/0022-006X.61.6.911

Peake, P. K. (2017). Delay of gratification: Explorations of how and why children wait and its linkages to outcomes over the life course. *Nebraska Symposium on Motivation*, *64*, 7–60. https://doi.org/10.1007/978-3-319-51721-6_2

Petersen, I. T., Bates, J. E., Dodge, K. A., Lansford, J. E., & Pettit, G. S. (2015). Describing and predicting developmental profiles of externalizing problems from childhood to adulthood. *Development and Psychopathology*, *27*(3), 791–818. https://doi.org/10.1017/S0954579414000789

Petersen, I. T., Choe, D. E., & LeBeau, B. (2020). Studying a moving target in development: The challenge and opportunity of heterotypic continuity. *Developmental Review*, *58*, 100935. https://doi.org/10.1016/j.dr.2020.100935

Petersen, I. T., & LeBeau, B. (2022). Creating a developmental scale to chart the development of psychopathology with different informants and measures across time. *Journal of Psychopathology and Clinical Science*, *131*(6), 611–625. https://doi.org/10.1037/abn0000649

Petersen, I. T., LeBeau, B., & Choe, D. E. (2021). Creating a developmental scale to account for heterotypic continuity in development: A simulation study. *Child Development*, *92*(1), e1–e19. https://doi.org/10.1111/cdev.13433

Petersen, I. T., Lindhiem, O., LeBeau, B., Bates, J. E., Pettit, G. S., Lansford, J. E., & Dodge, K. A. (2018). Development of internalizing problems from adolescence to emerging adulthood: Accounting for heterotypic continuity with vertical scaling. *Developmental Psychology*, *54*(3), 586–599. https://doi.org/10.1037/dev0000449

Pettersson, E., Lahey, B. B., Larsson, H., & Lichtenstein, P. (2018). Criterion validity and utility of the general factor of psychopathology in childhood: Predictive associations with independently measured severe adverse mental health outcomes in adolescence. *Journal of the American Academy of Child and Adolescent Psychiatry*, *57*(6), 372–383. https://doi.org/10.1016/j.jaac.2017.12.016

Phillips, E. M., Brock, R. L., James, T. D., Nelson, J. M., Espy, K. A., & Nelson, T. D. (2022). Empirical support for a dual process model of the p-factor: Interaction effects between preschool executive control and preschool negative emotionality on general psychopathology. *Journal of Psychopathology and Clinical Science*, *131*(8), 817–829. https://doi.org/10.1037/abn0000777

Polanczyk, G. V., Salum, G. A., Sugaya, L. S., Caye, A., & Rohde, L. A. (2015). Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *Journal of Child Psychology and Psychiatry*, *56*(3), 345–365. https://doi.org/10.1111/jcpp.12381

R Core Team (2022). *R: A language and environment for statistical computing*.

Rachlin, H., & Jones, B. A. (2008). Social discounting and delay discounting. *Journal of Behavioral Decision Making*, *21*(1), 29–43. https://doi.org/10.1002/bdm.567

Racine, N., McArthur, B. A., Cooke, J. E., Eirich, R., Zhu, J., & Madigan, S. (2021). Global prevalence of depressive and anxiety symptoms in children and adolescents during COVID-19: A meta-analysis. *JAMA Pediatrics*, *175*(11), 1142–1150. https://doi.org/10.1001/jamapediatrics.2021.2482

Ree, M. J., Carretta, T. R., & Teachout, M. S. (2015). Pervasiveness of dominant general factors in organizational measurement. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *8*(3), 409–427. https://doi.org/10.1017/iop.2015.16

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rothbart, M., & Derryberry, D. (1981). Development of individual differences in temperament. In *Advances in developmental psychology erlbaum*. vol. *1*, pp. 33–86.

Rothbart, M. K. (2011). *Becoming who we are: Temperament and personality in development*. Guilford Press.

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. *Child Development*, *72*(5), 1394–1408. https://doi.org/10.1111/1467-8624.00355

Rothbart, M. K., & Bates, J. E. (2006). Temperament. In *Handbook of child psychology: Social, emotional, and personality development*. vol. *3*, 6th ed. pp. 99–166). John Wiley & Sons, Inc.

Ruggero, C. J., Kotov, R., Hopwood, C. J., First, M., Clark, L. A., Skodol, A. E., Mullins-Sweatt, S. N., Patrick, C. J., Bach, B., Cicero, D. C., Docherty, A., Simms, L. J., Bagby, R. M., Krueger, R. F., Callahan, J. L., Chmielewski, M., Conway, C. C., De Clercq, B., Dornbach-Bender, A., Eaton, N. R., Forbes, M. K., Forbush, K. T., Haltigan, J. D., Miller, J. D., Morey, L. C., Patalay, P., Regier, D. A., Reininghaus, U., Shackman, A. J., Waszczuk, M. A., Watson, D., Wright, A. G. C., Zimmermann, J. (2019). Integrating the Hierarchical Taxonomy of Psychopathology (HiTOP) into clinical

practice. *Journal of Consulting and Clinical Psychology*, *87*(12), 1069–1084. https://doi.org/10.1037/ccp0000452.

Rutter, M., & Sroufe, L. A. (2000). Developmental psychopathology: Concepts and challenges. *Development and Psychopathology*, *12*(3), 265–296. https://doi.org/10.1017/s0954579400003023

Sattler, J. M. (2022). *Foundations of behavioral, social, and clinical assessment of children*. Jerome M. Sattler, Publisher, Inc.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, *8*(2), 23–74.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Shi, Q., & Ettekal, I. (2021). Co-occurring trajectories of internalizing and externalizing problems from grades 1 to 12: Longitudinal associations with teacher-child relationship quality and academic performance. *Journal of Educational Psychology*, *113*(4), 808–829. https://doi.org/10.1037/edu0000525

Shi, Q., Ettekal, I., Deutz, M. H. F., & Woltering, S. (2020). Trajectories of pure and co-occurring internalizing and externalizing problems from early childhood to adolescence: Associations with early childhood individual and contextual antecedents. *Developmental Psychology*, *56*(10), 1906–1918. https://doi.org/10.1037/dev0001095

Sroufe, L. A. (2016). The place of attachment in development. In J. Cassidy, & P. Shaver (Eds.), *Handbook of attachment: Theory, research, and clinical applications*. vol. *3*, p. 997–1011). Guilford Press.

Steinberg, E. A., & Drabick, D. A. G. (2015). A developmental psychopathology perspective on ADHD and comorbid conditions: The role of emotion regulation. *Child Psychiatry & Human Development*, *46*(6), 951–966. https://doi.org/10.1007/s10578-015-0534-2

Stephens, D. W., & Anderson, D. (2001). The adaptive value of preference for immediacy: When shortsighted rules have farsighted consequences. *Behavioral Ecology*, *12*(3), 330–339. https://doi.org/10.1093/beheco/12.3.330

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Suh, Y. (2015). The performance of maximum likelihood and weighted least square mean and variance adjusted estimators in testing differential item functioning with nonnormal trait distributions. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 568–580. https://doi.org/10.1080/10705511.2014.937669

Tackett, J. L., & Hallquist, M. (2022). The need to grow: Developmental considerations and challenges for modern psychiatric taxonomies. *Journal of Psychopathology and Clinical Science*, *131*(6), 660–663. https://doi.org/10.1037/abn0000751

Tackett, J. L., Lahey, B. B., van Hulle, C., Waldman, I., Krueger, R. F., & Rathouz, P. J. (2013). Common genetic influences on negative emotionality and a general psychopathology factor in childhood and adolescence. *Journal of Abnormal Psychology*, *122*(4), 1142–1153.

Thomas, A., & Chess, S. (1977). *Temperament and development*. Brunner/Mazel.

Ursache, A., Blair, C., Stifter, C., & Voegtline, K. (2013). Emotional reactivity and regulation in infancy interact to predict executive functioning in early childhood. *Developmental Psychology*, *49*(1), 127–137. https://doi.org/10.1037/a0027728

Wall, A. E., & Barth, R. P. (2005). Aggressive and delinquent behavior of maltreated adolescents: Risk factors and gender differences. *Stress, Trauma, and Crisis*, *8*(1), 1–24. https://doi.org/10.1080/15434610490888081

Wang, Y., & Liu, Y. (2021). The development of internalizing and externalizing problems in primary school: Contributions of executive function and social competence. *Child Development*, *92*(3), 889–903.

Watson, D., Clark, L. A., & Carey, G. (1988). Positive and negative affectivity and their relation to anxiety and depressive disorders. *Journal of Abnormal Psychology*, *97*(3), 346–53.

Watts, A. L., Makol, B. A., Palumbo, I. M., De Los Reyes, A., Olino, T. M., Latzman, R. D., DeYoung, C. G., Wood, P. K., & Sher, K. J. (2021). How robust is the p factor? Using multitrait-multimethod modeling to inform the meaning of general factors of youth psychopathology. *Clinical Psychological Science*, *10*(4), 640–661. https://doi.org/10.1177/21677026211055170

Watts, A. L., Meyer, F. A. C., Greene, A. L., Wood, P., Trull, T., Steinley, D., & Sher, K. (2021). Spurious empirical support for the p-factor arises with the inclusion of undiagnosed cases. *PsyArXiv*. Preprint: https://doi.org/10.31234/osf.io/4tazx

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science*, *29*(7), 1159–1177. https://doi.org/10.1177/0956797618761661

Weeks, J. P. (2010). plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software*, *35*(12), 1–33. https://doi.org/10.18637/jss.v035.i12

Weems, C. F. (2008). Developmental trajectories of childhood anxiety: Identifying continuity and change in anxious emotion. *Developmental Review*, *28*(4), 488–502. https://doi.org/10.1016/j.dr.2008.01.001

Weissman, D. G., Bitran, D., Miller, A. B., Schaefer, J. D., Sheridan, M. A., & McLaughlin, K. A. (2019). Difficulties with emotion regulation as a transdiagnostic mechanism linking child maltreatment with the emergence of psychopathology. *Development and Psychopathology*, *31*(3), 899–915. https://doi.org/10.1017/S0954579419000348

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. https://doi.org/10.2307/1912934

Zinbarg, R., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7

Supplementary Appendix S1. Details of Developmental Scaling (Linking Scores Across Informants, Measures, and Ages).

We used multidimensional IRT (M-IRT) and linking to create a single uniform developmental scale (i.e., developmental scaling) for externalizing and internalizing problems that spans multiple years of development. We conducted this linking in five steps: (1) Fit M-IRT models at each age and for each rater type separately. (2) Link the measures' scores over time within each rater type. (3) Link scores across raters. (4) Calculate latent factor scores on the linked scale. (5) Use linked factor scores in growth curve and bifactor models. We describe this procedure in detail below.

Step 1. Fit M-IRT models at each age and for each rater type separately.

We used the multidimensional graded response IRT model using the mirt package (Chalmers, 2012) in R 3.6.1 (R Core Team, 2022) to estimate item parameters. The mirt package uses a maximum likelihood expectation-maximization algorithm to estimate item parameters. The maximum likelihood estimation procedure uses all available data for each item and provides valid inferences if the data are missing at random or completely at random. The graded response model is a generalized version of the two-parameter logistic model for dichotomous outcomes, accommodating polytomous items that are ordinal in nature through a series of cumulative comparisons. The multidimensional graded response model adds the ability to include multiple latent factors (i.e., externalizing and internalizing problems)—and their covariance—in the same model. That is, internalizing and externalizing problem items were included in the same model, but they were allowed to load onto distinct latent factors. The externalizing and internalizing problem items in the current study were questionnaire items rated from 0 to 2. The multivariate graded response model takes the following general form:

$$P(X_{ni} = x_{ni} \,|\theta_n) = P^*_{x_{ni}}(\theta_n) - \, P^*_{x_{ni}+1}(\theta_n) \quad (1)$$

where:

$$P^*_{x_{ni}}(\theta_n) = P(X_{ni} \geq x_{ni}|\theta_n) = \frac{1}{1+ \mathrm{e}^{a_i(\theta_n \pm b_{ic})}}. \quad (2)$$

In this model, three parameters are of primary interest: a_i is a vector of item-specific discrimination parameter estimate for each latent factor; b_{ic} is an item-specific severity parameter (commonly referred to as difficulty in educational measurement literature); and θ_n is a subject-specific vector representing the child's level of externalizing and internalizing problems. In the above model, *i* represents unique items, *c* represents different categories that are rated, and *n* represents unique children. Because the respondent rates each item from 0 to 2, there are two b_{ic} item-specific severity terms reflecting the category boundary locations: b_{i0}, b_{i1}. The category boundary locations reflect the point at which the probability of being in category *c* or lower compared to the categories above *c* is 50%. For example, if an externalizing item has a severity estimate of $b_{i1} = 1.2$, there is a 50% probability of being in category 0 or 1 (i.e., category *c* or lower) compared to category 2 (i.e., categories above *c*) at this value, 1.2, on the externalizing problems scale. We used the externalizing problems latent factor as the reference group and allowed the mean and variance for internalizing problems latent factor to be estimated freely. Setting the externalizing factor as the reference group, along with linking both internalizing and externalizing items in the same model, placed the internalizing and externalizing problem scores onto the same mathematical scale across ages and raters. This multidimensional graded response IRT model is conceptually like a two-factor categorical confirmatory factor analysis approach (fit to ordinal data) with the internalizing and externalizing factors allowed to covary, and with no cross loadings.

There may be shifts in the externalizing or internalizing problem constructs over time due

to natural developmental changes (Petersen et al., 2018). The present study spans a wide age range (ages 2–15 years). When spanning a wide age range, it is considered safer to fit a separate model at each age rather than a single model that spans all ages because a model that spans across a wide age range is more likely to violate IRT dimensionality assumptions (Kolen & Brennan, 2014). We fit two latent factors corresponding to the constructs of interest: i.e., externalizing and internalizing problems. IRT assumes that each latent factor (e.g., externalizing problems) is unidimensional, which is more likely at a single time point than across all time points in the same model. Thus, we fit a separate IRT model at each age and for each rater type in the present study. This approach was also applied by Petersen et al. (2018) and by Petersen & LeBeau (2022) in their creation of a developmental scale for internalizing and externalizing problems, respectively, across a wide age range.

Step 2. Link the measures' scores over time within each rater type.

After successful estimation of the individual IRT models, we used multidimensional linking methodology to create the developmental scale for externalizing and internalizing problems. Developmental scaling (aka vertical scaling) is a form of data harmonization that aims to place two measures that assess the same construct but differ based on severity and discrimination onto the same scale. One way to create a developmental scale is to link the two measures. The strength of the linking is enhanced if there are items that overlap across the two measures, often referred to as common items or anchor items in educational measurement. When linking any pair of measures in the present study, some items were shared across measures (i.e., common items) and some items were not shared (i.e., unique items). We used M-IRT to link the scores across informants, measures, and ages based on their common items. The M-IRT approach to linking minimizes differences between the probability of a person endorsing the

common items across the two given measures to be linked. That is, we linked measures' scores so that their common items had similar severity and discrimination at the scale level by minimizing the differences in their test characteristic curves of the common items (i.e., lessening the gap between the two curves; see Figures S2–S5). See Figure 1 for a visualization of the measure to which each other measure was linked. Developmental scaling based on item parameter invariance theory assumes that any difference in item parameter estimates is able to be rescaled onto a single unified metric with a linear transformation. Based on this assumption, the item parameters and the resulting latent factor scores of externalizing and internalizing problems can be linked across ages by comparing and linearly transforming differences in discrimination and severity of the common items across ages. We created the developmental scale by linking scores across ages and raters with four steps described in detail below:

(1) As described above, we fit M-IRT models at each age and for each rater type separately, resulting in 31 M-IRT models (see Table 2 for the 31 rater-by-age instances). For example, we fit a separate M-IRT model for mothers' ratings at age 5 and mothers' ratings at age 6. Each IRT model estimated latent factor scores that represented a child's level of externalizing and internalizing problems. We then linked externalizing and internalizing problem scores across informants, measures, and ages to be on the same scale. As an example, we linked mothers' ratings at age 3 on the Child Behavior Checklist (CBCL) 2–3 to mothers' ratings at age 4 on the CBCL 4–18 using the common items of the CBCL 2–3 and CBCL 4–18. Common items across the CBCL 2–3 and CBCL 4–18 included items such as "destroys own things." When we linked scores across ages or informants from the same measure, all items were common items[1]. For example, we linked mothers' ratings at age 5 on the CBCL 4–18 to mothers' ratings at age 6 on the

CBCL 4–18 using all of their items (all of their items were common items because the items came from the same measure). The number of common items for each pair of measures to be linked is in Table 2.

(2) We used multidimensional developmental scaling techniques to link the measures' scores over time within each rater type. We used the plink package (Weeks, 2010) in R to perform the linking by using the multidimensional test characteristic function procedure with an oblique Procrustes rotation (Oshima et al., 2000). The oblique rotation method allowed the latent factors, externalizing and internalizing problems, to be correlated. For linking, we used a multidimensional Stocking-Lord procedure (Stocking & Lord, 1983). The Stocking-Lord linking procedure iteratively estimates linking constants by minimizing differences in the aggregate scores across common items. We used the Stocking-Lord linking procedure as opposed to other linking procedures (e.g., Haebara) because we were interested in construct-level (i.e., externalizing and internalizing problems) scores and were less interested in the response to a single item. Nevertheless, there has been little empirical difference shown between the two characteristic curve linking methods, Stocking-Lord and Haebara. As an empirical test, we used multidimensional least squares linking as a comparison and found little empirical difference between the linking parameters and resulting factor scores. For example, the correlations between the factor scores using least squares linking compared to the Stocking-Lord linking were typically $r \geq .99$ for both externalizing and internalizing problems.

To estimate the Stocking-Lord parameters, we set the reference age to be 6 years of age for each rater because age 6 was the first age when most rater types (except other caregivers and self-report) provided ratings of the child's externalizing and internalizing problems. We set the reference rater to be the mother because the mother typically provided the most ratings across the

developmental age span. The reference age and rater pair set the scale to which the item parameters at subsequent ages and for other raters were transformed. In other words, we transformed the estimated item parameters at all ages and for all raters to be on the same scale as the item parameters estimated for mothers' ratings at 6 years of age. To achieve this, we first linked the item parameters across ages within rater type. To perform the linking of scores from two measures, we estimated the test characteristic curve for the common items of each of the pair of measures to be linked. The test characteristic curve represents the probability of endorsing the items (i.e., the expected proportion out of the total possible score) as a function of a child's latent level of externalizing or internalizing problems. Because we used a confirmatory M-IRT model where we directly specify which items load onto the externalizing or internalizing problems latent factor, we simplified each dimension to a curve instead of a response surface. We specified loadings of externalizing problem items to be zero for the internalizing problems dimension and vice versa. Next, we estimated scaling parameters to make the test characteristic curves of the common items of each measure more similar. We estimated scaling parameters as the linear transformation (i.e., intercept and slope parameter) that, when applied to the second measure (see Equations 3–4), minimizes differences between the probability of a person endorsing the common items across the two measures. The scaling parameters that we used to link each pair of measures are in Supplementary Table S4. We describe an example below.

See Figures S2 through S5 for examples of test characteristic curves of the common items of mother- and teacher-rated externalizing and internalizing problems at age 6. The left panel of the figure illustrates the test characteristic curves for the common items before the linking process (i.e., the model-implied proportion out of total possible scores on the common items as a function of the latent externalizing problems score for mothers' and teachers' ratings at age 6).

The right panel of the figure illustrates the test characteristic curves for the common items after the linking process. The gap between the mother- and teacher-rated test characteristic curves (depicted by gray shading) indicates different probabilities of endorsing the common items across the measures (i.e., different severity and/or discrimination of the common items), where larger differences reflect scores that are less comparable. Discrimination is depicted by the steepness of the slope at the inflection point of the test characteristic curve. Severity is represented by the value on the x-axis at the inflection point of the test characteristic curve. Linking uses linear scaling parameters to minimize differences between the discrimination and severity of the common items. We estimated scaling parameters to minimize the differences in the mothers' and teachers' test characteristic curves at age 6. The scaling parameters to link teachers' ratings on the TRF at age 6 to mothers' ratings on the CBCL 4–18 at age 6 are shown in Supplementary Table S4. The left panel of Figure S4 and S5 indicate that, prior to linking, mothers' ratings showed somewhat lower discrimination than teachers' ratings at age 6. When linking developmental scales between mothers and teachers (Figures S4 and S5), the non-uniform DIF shown by the crossing test characteristic curves prior to linking (left panel) was adjusted to remove the non-uniform DIF in the linked scores (right panel). The right panel shows considerably smaller differences between the two test characteristic curves, which provides empirical evidence that the linking successfully placed the latent externalizing problem scores across raters on a more comparable scale (i.e., more similar discrimination and severity of the common items). In general, we observed successful linking across ages and raters (see Figures S2–S5).

To link scores across ages for a given rater type, we estimated Stocking-Lord linking constants that linked the item parameters at a given age to be on the same scale as the item

parameters at an adjacent age for that rater type. For example, we estimated linking constants between adjacent age spans for mothers' ratings, for example between 5 and 6 years of age, 7 and 8 years of age, and so on. We used two estimated scaling constants including an intercept parameter, B, and a slope parameter, A, to link the item parameters onto the reference scale. We performed the process of linking iteratively by chaining together multiple linking constants across the age span. We linked all measures directly or indirectly to the scale of mothers' ratings at age 6. For example, we linked mothers' ratings at age 5 directly to mothers' ratings at age 6 because they were at adjacent ages. By contrast, we linked mothers' ratings at age 4 indirectly to mothers' ratings at age 6 via mothers' ratings at age 5, using a process of linking and chaining. To do this, we first linked mothers' ratings at age 4 to the scale of mothers' ratings at age 5, and then linked the mothers' ratings at age 4 on the age 5 scale to the age 6 scale. As an example of linking across raters, teachers' ratings at age 5 were indirectly linked to mothers' ratings at age 6 via teacher's ratings at age 6 (see Figure 1 for the broader linking design). We first linked scores within rater type (see Equation 5), and then linked scores across raters to link scores to mothers' ratings (see Equation 6).

After successfully estimating the linking constants, we then transformed all item parameters to be on the age 6 scale for the given rater. The transformations took the following form:

$$\boldsymbol{a}\left(\text{age}_i\right) = (\boldsymbol{A^{-1}})\boldsymbol{a}(age_j), \tag{3}$$

$$\boldsymbol{b}\left(\text{age}_i\right)_c = \boldsymbol{b}\left(\text{age}_j\right)_c - \boldsymbol{a}\left(age_j\right)' \boldsymbol{A^{-1}B}, \tag{4}$$

where $\boldsymbol{a}\left(\text{age}_i\right)$ and $\boldsymbol{a}(\text{age}_j)$ are vectors of discrimination parameter estimates for the common items at adjacent ages *i* and *j* respectively; $\boldsymbol{b}\left(\text{age}_i\right)_c$ and $\boldsymbol{b}\left(\text{age}_j\right)_c$ are severity parameter

estimates for the common items at adjacent ages *i* and *j* respectively for category *c*; ***A*** is a rotation matrix which is 2 x 2 in the present study due to the two latent factors, and ***B*** represents a translation vector. Min (2007) provides further technical details on the multivariate linking terms. To shift all item parameters to a common age 6 scale, we applied all previous adjacent scaling constants to the item parameters. For example, when shifting the item parameter estimates for 7-year-olds to the age 6 scale, we used a single set of scaling constants. However, when shifting the item parameters for 8-year-olds, we used two sets of scaling constants: first, we transformed the item parameter estimates for 8-year-olds to the scale of the 7-year-olds, and then we transformed them a second time to be on the age 6 scale. See Figure 1 for a visualization of the linking process. We performed this step of the linking process separately for each row in the figure (i.e., within rater types; horizontal arrows).

Step 3. Link scores across raters.

(3) After creating developmental scales across ages within rater types, we linked scores across raters at age 6 (except for the other caregivers' reports collected at age 2 and self-report collected at age 15). As described above, we set the mother as the reference rater. Percentage of participants with scores on behavior problem ratings across time points are in Supplementary Table S3. We used a similar process as in step 2; we estimated Stocking-Lord linking constants to link the item parameters across raters within a single age. For example, we estimated a set of linking constants to link the item parameters of the fathers' ratings to the item parameters of mothers' ratings at age 6 to ensure that their factor scores were on the same scale. This step moved the developmental scales for fathers, teachers, and afterschool caregivers to the mothers' scale, anchored at age 6, while preserving the developmental scale created within rater types in step 2. The process of linking scores across raters is depicted in Figure 1 with the gray bounding

boxes (vertical arrows).

Step 4. Calculate latent factor scores on the linked scale.

(4) After successfully placing item parameter estimates onto a single developmental scale (for all raters and ages), we calculated children's latent externalizing and internalizing problem scores with expected a posteriori (EAP) factor scores. The linking in the previous two steps scaled the factor scores to be on the single developmental scale while retaining changes in means and variances over time and across raters. The factor scores are assumed to be linearly related based on the following equation:

$$\boldsymbol{\theta}(\text{age } 6) = \boldsymbol{A\theta}\left(\text{age}_j\right) + \boldsymbol{B} \tag{5}$$

where $\boldsymbol{\theta}(\text{age } 6)$ represents a vector of factor scores at age 6 (the reference scale) and $\boldsymbol{\theta}\left(\text{age}_j\right)$ represents a vector of factor scores at subsequent measurement occasions. The chaining description referenced with the linking applies here, as well. For example, the factor scores at age 8 used two sets of linking constants to transform them to the age 6 reference age: one between ages 6 and 7 and another between ages 7 and 8. Finally, after creating the developmental scale within each rater type, we then linked each rater to the age 6 mother scale using a similar equation to above, except only a single transformation was used across each rater.

$$\boldsymbol{\theta}\left(\text{age } 6_{\text{mother}}\right) = \boldsymbol{A\theta}\left(\text{age } 6_r\right) + \boldsymbol{B} \tag{6}$$

where $\boldsymbol{\theta}\left(\text{age } 6_{\text{mother}}\right)$ represents the vector of factor scores at age 6 for the mother rater and $\boldsymbol{\theta}\left(\text{age } 6_r\right)$ represents the vector of factor scores at age 6 for the r rater types including fathers, teachers, caregivers, and afterschool caregivers. For transforming the other caregivers' scores at age 2 to mothers' ratings, we linked the scores with a similar equation, however we used the transformed mothers' ratings at age 2 as the reference group (see Figure 1). For transforming the self-reported scores at age 15 to mothers' ratings, we used the transformed mothers' ratings at

age 15 as the reference group (see Figure 1). The linking constants by measure and age are in Supplementary Table S4. Post-linking estimates of scale-level DIF between measures used to link scores across different raters and ages are in Supplementary Table S5. Tests of differential item functioning (DIF) by age showed no major concerns at the scale level after linking (see Supplementary Appendix S2). Distribution of DIF effect size statistics between ages by rater type are in Supplementary Figure S1.

In sum, the linking of scores within a rater type created a developmental scale for scores from that rater type, so each rater type had their own trajectory (see Figure 2). We then, ultimately, linked each rater type's developmental scale (directly or indirectly) to the mothers' ratings at age 6, so that each rater type's trajectory was on the same developmental scale. Examples of linked scores across raters and years are depicted with test characteristic curves in Supplementary Figures S2 through S5. The test characteristic curves of the linked scores across raters and years were highly similar (and more similar than the test characteristic curves of the pre-linked scores), indicating that we successfully linked scores across raters and years to be on the same scale. As a secondary analysis, we also examined aggression and delinquent subdimensions of externalizing problems given their differing associations with risk factors (Murray & Farrington, 2010; Wall & Barth, 2005). Thus, we also conducted developmental scaling with aggression and delinquent behavior (see Supplementary Appendix S3).

Step 5. Use linked factor scores in growth curve and bifactor models.

After linking factor scores from all raters and at all ages to be on the scale of mothers' ratings at age 6, we used the linked factor scores as the child's estimated level of behavior problems for a given rater and age in subsequent growth curve and bifactor models.

Supplementary Appendix S2. Tests of Differential Item Functioning by Age and Rater.

Method

After fitting multidimensional item response theory (M-IRT) models, we examined whether there was differential item functioning (DIF) across ages and raters (comparable to tests of longitudinal measurement/factorial invariance). Lack of DIF across ages and raters for individual items is not an assumption of the linking procedure we used because the linking was performed at the scale level of the common items (rather than at the item level). Nevertheless, we examined the extent of DIF to evaluate the degree to which linking across ages and raters was likely to be successful with the common items. DIF examines whether the likelihood of endorsing a particular item differs between groups (in this case, between two ages or raters) for people with the same levels on the construct. To evaluate the extent to which the linking would be successful with the common items, we examined potential item-level and scale-level DIF using the common items between adjacent ages and between raters at ages when we linked raters' scores. We expected some but modest item-level DIF of the common items across ages prior to linking, consistent with a construct that shows theoretically expected changes in its manifestation across development (heterotypic continuity). The Stocking-Lord multivariate linking procedure with an oblique rotation we used to link scores across measures, informants, and years minimizes scale-level latent factor differences rather than item-level differences (that would be minimized by the least-square multivariate linking procedure). Thus, we expected some items to continue to show DIF even after linking, but we expected that the item-level DIF would be offset by other items on the aggregate. By contrast, we expected that the scale-level DIF would improve (i.e., decrease) after linking (because the Stocking-Lord linking procedure minimizes scale-level DIF).

To evaluate DIF, we used effect size measures following strategies discussed by Raju (1988) and Meade (2010) that mitigate the multiple testing problems that would occur from testing DIF across hundreds of items (i.e., many items across many ages and multiple raters) in a hypothesis testing framework. The effect size measure computes the difference in the expected scores (i.e., model-implied scores) for an individual item for the focal and reference groups (e.g., age 4 compared to age 5) at specific values of the latent externalizing and internalizing problems scale. The multiple differences are then averaged across the latent externalizing and internalizing problems scale. The effect size is interpreted as the average difference in the expected scores on the item across the two groups. There are two versions of this computation, a signed and unsigned difference. The unsigned difference takes the absolute value of the difference in expected scores whereas the signed difference does not. The primary benefit of computing the two statistics is to detect uniform versus non-uniform DIF. Uniform DIF occurs when one group systematically has higher or lower expected scores compared to the other group. Non-uniform DIF occurs when the expected scores change in sign; for example, one group has higher expected scores at lower latent factor scores but has lower expected scores at higher latent factor scores. If unsigned differences are present and signed differences are similar in magnitude to the unsigned differences, uniform DIF is present. If unsigned differences are present and signed differences are smaller than unsigned differences, non-uniform is present. Uniform DIF reflects differences in difficulty (i.e., severity) between groups, whereas non-uniform DIF reflects differences in discrimination (and possibly severity) between groups. Differences in discrimination could indicate that an item is not construct-valid for a particular rater at a given age, so non-uniform DIF is considered more potentially problematic than uniform DIF.

We used a similar approach to examine common item scale-level differences, consistent

with the approach we used to examine item-level differences. However, when examining common item scale-level differences, the expected scores would be the expected scores at the latent factor-level (of the common items) instead of at the item-level. The expected scores at the latent factor-level are equivalent to a sum of the item-level expected scores for the common items. We standardized the expected scores (for the purposes of testing DIF) to remove the effect of a different number of common items used for linking at adjacent ages. As an example, for externalizing problems, we used 26 common items to link mothers' ratings between ages 2 and 3, but we used only 9 common items to link mothers' ratings between ages 3 and 4 (see Supplementary Table 2).

We conducted DIF analysis for externalizing and internalizing problems separately due to the confirmatory nature of the multivariate IRT model. We assumed simple structure for the multivariate IRT model; each item was specified to load (i.e., a discrimination term was estimated) on one and only one of the latent factors as designed by the test developers. For example, an item was assumed to load on either externalizing or internalizing problems, not both. This simple structure approach allowed for the DIF analysis to independently evaluate the extent to which the multidimensional linking was successful on the externalizing and internalizing scales separately.

There is not strong guidance for interpreting effect sizes of DIF. We selected effect size cutoffs that would help identify potentially important DIF while not focusing on negligible differences. At both the item level and scale level, we selected effect size cutoffs a priori consistent with prior work (Petersen & LeBeau, 2022) so that minor DIF would represent a 5% difference in expected scores, whereas moderate DIF would represent a 10% difference in expected scores. To achieve this, for determining the effect size of item-level DIF, we used effect

sizes thresholds of 0.1 and 0.2 for evidence of minor and moderate DIF, respectively. For instance, an effect size of 0.1 would indicate that the expected scores for one group are on average 0.1 score points different from the expected scores of the other group. The expected score range is from 0 to 2, so an effect size of 0.1 would indicate a 5% difference in expected scores (i.e., 0.1 / 2 = 5%). For scale-level DIF, we used effect size thresholds of 0.05 and 0.1 for minor and moderate DIF, respectively. We used more stringent effect size thresholds for scale-level DIF because we standardized the expected scores to range from 0 to 1 instead of ranging from 0 to the total number of score points (i.e., the total number of score points on the scale would reflect the number of items times two, with two reflecting the total number of score points on a single item). The effect size cutoffs were half the size for scale-level DIF compared to the effect size cutoffs for the individual items due to the standardization, ranging from 0 to 1 for the scale level, compared to ranging from 0 to 2 for the individual items. Thus, effect size cutoffs for both item-level and scale-level DIF were comparable such that minor DIF would represent a 5% difference in expected scores, whereas moderate DIF would represent a 10% difference in expected scores.

Results

DIF Between Ages

Item-level DIF. Out of the 1,377 common items from creating the developmental scales within rater type across externalizing and internalizing problems, 1 item showed evidence of DIF in terms of discrimination and 114 items (8%) showed evidence of DIF in terms of severity. The percentage of items showing DIF (i.e., had effect size measures greater than 0.1) between ages ranged from 6% to 21% across raters, although most of these items showed only minor levels of DIF. Rates of moderate DIF between ages ranged from 0% to 6% across raters. Afterschool

caregivers' ratings showing the highest rates of minor and moderate DIF between ages after linking, with about 16% and 6% of the 140 common items showing evidence of minor and moderate DIF, respectively. There were four items that showed DIF across three pairs of ages: two items for the mother and teacher developmental scales. For these items, there was no evidence of systematic item-level DIF in the same direction. The severity shift in the signed metric was positive or negative with no apparent pattern. In addition, the items for the teacher scale did not show evidence of DIF between consecutive ages. Supplementary Figure S1 shows the distribution of unsigned effect size statistics between ages by rater type both before and after linking. The figure illustrates that most items showed no evidence of DIF across ages. For the items that showed evidence of DIF across ages, we also examined non-uniform DIF. We flagged items that showed unsigned effect sizes greater than 0.1 and had signed effect size statistics less than 0.05 in absolute value. Before linking, two items for the mother showed evidence of non-uniform DIF across ages. After linking, only one of those items remained as showing evidence of non-uniform DIF across ages and the linking reduced the magnitude of DIF by approximately 25%.

Supplementary Figure S1 also shows differences based on if the item assessed internalizing or externalizing problems. Before linking, internalizing problem items showed greater DIF than externalizing problem items for reports by teachers, mothers, afterschool caregivers, and other caregivers. These differences were greatly reduced after linking.

Scale-level DIF. We also evaluated DIF at the scale-level to determine the extent to which the developmental scales were placed on the same scale within a rater. Scale-level DIF estimates are in Supplementary Table S5. Of all five raters where a developmental scale was created and a total of 50 linkages examined across externalizing and internalizing problems,

there were five linkages that showed evidence of scale-level DIF after linking. Four of the five total instances of scale level DIF were for internalizing problems and one was for externalizing problems. Of the five total scales that showed some evidence of DIF, four of the five had an effect size statistic less than 0.1, indicating minor scale-level DIF. The one that was larger, had an effect size statistic of 0.11, indicating moderate DIF for afterschool caregivers' ratings of externalizing problems between ages 6 and 8. We proceeded with the developmental scale for afterschool caregivers for at least two reasons. First, there was no evidence of DIF in discrimination, which would indicate more problematic DIF. The DIF in this scale was due to severity differences, which may occur due to heterotypic continuity or may reflect challenges that afterschool caregivers have in rating children's externalizing problems. Second, compared to ratings by other informants, there was relatively less variation in the ratings by afterschool caregiver responses, which makes IRT model estimation more difficult.

DIF Between Raters

Item-level DIF. Finally, we also explored potential DIF between raters. The percentage of items that showed some level of DIF between raters ranged from 18% to 76% across rater comparisons prior to linking and this percentage ranged from 12% to 84% across rater comparisons after linking. Even though some items showed some level of DIF, most of these were minor DIF across raters shown by the percentage of items that were minor DIF out of the total DIF items, ranging from 29% to 85%. The one linking that had more items showing DIF was between mothers and other caregivers, a linking that was performed at age 2. Of the items that showed DIF, 11 of 271 items showed non-uniform DIF prior to linking, and five items showed non-uniform DIF after linking. Furthermore, there was evidence that externalizing problem items showed greater evidence of DIF between mothers and afterschool caregivers

compared to internalizing problem items (93% externalizing versus 73% internalizing). By contrast, internalizing problem items showed greater evidence of DIF between mothers and other caregivers, where 88% of internalizing items showed evidence of DIF compared to only 15% of externalizing problems. Therefore, although there was evidence of item-level DIF, the linking improved the magnitude of DIF and removed over half of the instances of items showing non-uniform DIF.

Scale-level DIF. We also examined potential scale-level DIF between raters over a total of ten linkages. Scale-level DIF estimates are in Supplementary Table S5. There was evidence of minor DIF for three of the scales and moderate DIF for one scale prior to linking between mothers' and afterschool caregivers' and caregivers' ratings. After linking, three of those scales still showed minor DIF, with no moderate DIF present. The effect size reduction for those that showed evidence of DIF was between 25% and 50%, indicating a strong reduction in the amount of scale-level DIF after linking.

Discussion

In summary, we observed some evidence of DIF but generally observed that linking successfully smoothed out the DIF at the scale-level, which provides support that our procedure for linking scores across ages and raters was successful. We observed some item-level DIF, but relatively few items showed DIF for a given rater at a given age. Moreover, where item-level DIF was observed, the effect sizes tended to be small, suggesting negligible DIF. The greatest number of instances of DIF at the item and scale level occurred when linking afterschool caregivers' ratings between ages 6 and 8. In particular, items showed evidence of DIF related to severity, but not discrimination. This uniform DIF is less problematic than non-uniform DIF. In general, linking appeared to be successful across both ages and raters, especially for mothers'

ratings from ages 2–15, fathers' ratings from ages 6–15, teachers' ratings from ages 5–11, other caregivers' ratings from ages 2–4, and self-report at age 15. Given the number of links that were established, both within and across raters when creating the developmental scale, the reduction in DIF after linking was substantial and represents a strong improvement in terms of placing the measures' scores onto to the same scale.

Differences in severity are expected across a lengthy developmental span and are unlikely to be serious threats to measuring the same construct. Compared to differences in severity, differences in discrimination are potentially more serious because they may reflect that an item does not reflect the same construct for some raters at some ages. However, changes in discrimination may instead reflect meaningful developmental shifts in the construct (heterotypic continuity) even though the items still reflect the theoretical content of the construct, as was likely the case in the present study given the strong empirical basis and content validity of the measures we used. Nevertheless, most of the DIF we observed reflected differences in severity (uniform DIF) rather than differences in discrimination (non-uniform DIF). We observed very little evidence of non-uniform DIF at the item level (only six items after linking), and no instances of non-uniform DIF at the scale level, further supporting that we were measuring the same construct at all ages.

Despite considerable research on DIF and measurement invariance, there is not clear guidance in the literature on how to proceed in the case of DIF (or failed measurement invariance) because there is no test to determine whether the difference reflects a change in the manifestation of the construct (i.e., heterotypic continuity), changes in the functioning of the measures, or some combination of the two. Nevertheless, we examined the effect size of DIF and it was modest in all situations except one (afterschool caregivers between ages 6 and 8). Our

developmental scaling approach accounted for DIF by estimating a separate IRT model at each age and for each rater, thus allowing items' parameters to change over time and to differ across raters, and using scaling parameters to link the scores across ages and raters to "smooth out" the DIF at the construct level. In sum, there are theoretical and empirical considerations when determining whether we measured the same construct in an equivalent way over time, and the totality of the evidence suggests that we did.

Supplementary Appendix S3. Tests of Differential Item Functioning by Sex and Ethnicity

We conducted tests of differential item functioning (DIF; i.e., measurement non-invariance) by sex and ethnicity. We conducted the DIF analysis using the two-factor IRT models with externalizing and internalizing problems constructs. For the sex comparison, male respondents were compared to female respondents across age and rater combinations. For the ethnicity comparison, White respondents were compared to those who were not White. It was necessary to combine the non-White racial groups for purposes of DIF testing due to the modest sample size of participants who were not White or Black.

The DIF analysis procedure mimicked the DIF examined across ages and raters at the scale level (see Supplementary Appendix S2). DIF at the item level was not examined, because the scale-level scores were the focus in the present study. The two-factor IRT model was fit to the subgroup data (i.e., combination of age, rater, and sex/race group) separately. Then, the difference in the test characteristic curves across the subgroups were compared in the effect size metric defined by Meade (2010). Due to small subgroups, some items were not included in the IRT models for specific ages or raters due to a lack of variation in a given item for a given subgroup. For example, all mothers at age 4 endorsed a value of 0 for one of the items; therefore, this item would represent a constant for which the model estimates cannot be obtained. Also, due to smaller sample sizes, the convergence criterion was increased from 0.0001 (the default value), to 0.001 to aid in convergence for small subgroups. We did not perform developmental scaling for making the comparisons; instead, we compared scores across sex/race groups within an age and rater combination to evaluate how much the male/female and White/non-White subgroups differed. Finally, we compared scores for mothers, fathers, and teacher raters. For caregiver and self-report ratings, the models could only readily converge by dropping many items.

Comparisons of scores by sex and ethnicity with caregivers and self-report raters would have been too different from the combined models to be of usefulness to evaluate the impact of the DIF for these subgroups. Thus, we did not compare scores by sex and ethnicity for caregiver and self-report ratings. However, we expect that results would be qualitatively similar to those described below for the sex and ethnicity subgroups.

DIF results showed that there were some differences in both the sex and ethnicity subgroups. When exploring the differences between male and female sub-groups, the effect sizes ranged from close to zero to 4 or 5 score point difference for the teacher rater. There was evidence of greater DIF for teacher raters compared to mothers and fathers. Using similar cut scores for small, medium, and large DIF of 0.05, 0.10, and greater than 0.10, respectively, 2 comparisons showed evidence of small magnitude DIF, 2 comparisons showed evidence of medium magnitude DIF, and the remaining 48 were large DIF. Females had higher scores for 18 of the 52 comparisons and males had higher scores for the remaining 34. Furthermore, only 3 of the 20 comparisons for the teacher had higher scores for females and were for the internalizing problems construct. Similarly, 4 of the 11 comparisons made for the father rater had higher scores for females and were all for internalizing problems. Mother raters had similar numbers that had higher scores for females and males; similar to the other two raters, the majority of instances were for internalizing problems. Only 5 comparisons showed discrimination differences (non-uniform DIF); the remaining differences were in severity (uniform DIF). This provides evidence that the covariate adjustment within the growth models should adjust for group-level differences in the factor scores for male and female individuals.

DIF results for the ethnicity group showed similar results to that for sex. DIF effect size statistics ranged from 0.1 to a high of 5.5. Of the 52 comparisons, 2 showed moderate DIF, and

the remaining were large DIF. Teachers had evidence of having more DIF compared to mothers and fathers. When comparing White to Non-White children, mother raters tended to show larger DIF effect sizes than father raters. One comparison showed higher scores for White individuals for mother raters at age 4 for the internalizing problems construct. The remaining 51 comparisons showed that Non-white children had higher scores than White individuals. Similar to the sex DIF evaluation, only 3 comparisons showed discrimination differences (non-uniform DIF); the remaining differences were in severity (uniform DIF). Of those 3, 2 had small DIF effect sizes suggesting this effect was smaller. This provides evidence, similar to results of the sex DIF exploration, that the covariate adjustment in the growth models should provide adequate adjustment for group-level differences in the factor scores for White and Non-White individuals.

Supplementary Appendix S4. Developmental Scaling of Externalizing Problem Subdimensions: Aggression and Delinquent Behavior

As a secondary analysis, we also conducted developmental scaling of aggression and delinquent behavior subdimensions of externalizing problems. Three-factor IRT models that included latent factors for aggression, delinquent behavior, and internalizing problems, were fit for ages 4 through 15. For ages 2 and 3, the CBCL measure includes a subscale of destructive behavior rather than delinquent behavior. Similar to the two-factor model, separate IRT models were fit for each age and rater combination. Upon model convergence, we performed linking to create developmental scales across ages within a rater, as described in Supplementary Appendix S1. Then, we linked raters at a single age. This allowed the developmental scale for the three-factor models to adjust for any scale-level differences across ages and raters.

Each IRT model converged; however, linking could not be adequately performed between ages 2/3 and 4, because the linking between destructive behavior (ages 2–3) and destructive behavior (ages 4–15) was unsuccessful. This was likely due, in part, due to minimal item overlap of the differing subscales and to less variance in the item-level responses for ages 2/3 on the destructive subscale compared to responses on the delinquent externalizing problems subscale at older ages.

To see how similar results were for developmentally scaled factors scores from the two- and three-factor models, we evaluated the correlations between the factor scores for the two-factor and three-factor models. Internalizing problem factor scores were highly correlated across the two- and three-factor models; correlations were generally greater than $r = .99$ for all rater and age combinations where linking was successful. Aggression factor scores from the three-factor model were highly correlated with externalizing problem factor scores from the two-factor

model; correlations were greater than $r = .95$ for all age and rater combinations where the linking was successful. The association between delinquent behavior factor scores from the three-factor model and externalizing problems factor scores from the two-factor model were somewhat smaller, but were still strongly associated, ranging between $.70 < r < .85$. The high correlations between factor scores from the two- and three-factor models provides additional confidence in the stability of the linking procedure and suggests that findings examining aggression and delinquent behavior are likely to be similar with those of general externalizing problems. Results from growth curve models examining aggression and delinquent behavior are in Supplementary Appendix S9. Results from bifactor models examining aggression and delinquent behavior are in Supplementary Appendix S10.

Supplementary Appendix S5. Growth Curve Model Formulas

$$Y_{ij} = \beta_0 + b_{00i} + \epsilon_{ij}$$

$$Y_{ij} = \beta_0 + \beta_1\left(\text{age}_{ij} - 15\right) + \beta_2\left(\text{age}_{ij} - 15\right)^2 + b_{00i} + b_{10i}\left(\text{age}_{ij} - 15\right) + \epsilon_{ij}$$

$$Y_{ij} = \beta_0 + \beta_1\left(\text{age}_{ij} - 15\right) + \beta_2\left(\text{age}_{ij} - 15\right)^2 + \beta_3\text{rater}_{ij} + b_{00i} + b_{10i}\left(\text{age}_{ij} - 15\right) + \epsilon_{ij}$$

$$Y_{ij} = \beta_0 + \beta_1\left(\text{age}_{ij} - 15\right) + \beta_2\left(\text{age}_{ij} - 15\right)^2 + \beta_3\text{rater}_{ij} + b_{00i} + b_{10i}\left(\text{age}_{ij} - 15\right) + b_{20i}\left(\text{age}_{ij} - 15\right)^2 + \epsilon_{ij}$$

$$Y_{ij} = \beta_0 + \beta_1\left(\text{age}_{ij} - 15\right) + \beta_2\left(\text{age}_{ij} - 15\right)^2 + \beta_3\text{rater}_{ij} + \beta_4\left(\text{age}_{ij} - 15\right) \times \text{rater}_{ij} + b_{00i} + b_{10i}\left(\text{age}_{ij} - 15\right) + \epsilon_{ij}$$

$$Y_{ij} = \beta_0 + \beta_1\left(\text{age}_{ij} - 15\right) + \beta_2\left(\text{age}_{ij} - 15\right)^2 + \beta_3\text{rater}_{ij} + \beta_4\left(\text{age}_{ij} - 15\right) \times \text{rater}_{ij} + \beta_k\text{Demographics}_{ik} + b_{00i} + b_{10i}\left(\text{age}_{ij} - 15\right) + \epsilon_{ij}$$

$$Y_{ij} = \beta_0 + \beta_1\left(\text{age}_{ij} - 15\right) + \beta_2\left(\text{age}_{ij} - 15\right)^2 + \beta_3\text{rater}_{ij} + \beta_4\left(\text{age}_{ij} - 15\right) \times \text{rater}_{ij} + \beta_k\text{Demographics}_{ik} + \beta_5\text{NegEmot}_i + \beta_6\text{NegEmot}_i \times \left(\text{age}_{ij} - 15\right) + \beta_7\text{Delay}_i + \beta_8\text{Delay}_i \times \left(\text{age}_{ij} - 15\right) + b_{00i} + b_{10i}\left(\text{age}_{ij} - 15\right) + \epsilon_{ij}$$

Note. Y_{ij} is the behavior problems factor score for person *i* at time *j*. $\beta_0, \ldots \beta_k$ are fixed-effect terms representing the unstandardized estimate of the association between the predictor and behavior problems. b_{0i}, b_{1i}, and b_{2i} are random effects representing person-specific deviations from the intercept, linear slope, and quadratic slope respectively. ϵ_{ij} are within-person error terms for person *i* at time *j*. Demographics_{ik} represents a set of *k* demographic covariates used to account for potential differences as a function of sex, ethnicity, and income-to-needs ratio. The focal predictors of interest were β_5, β_6, β_7, β_8 representing the association of negative emotionality and delay of gratification with intercepts and slopes, respectively, of behavior problems.

Supplementary Appendix S6. Tests of Systematic Missingness and How Missing Data Were Handled.

Tests of Systematic Missingness

We observed some systematic missingness of behavior problem scores as a function of demographic and socioeconomic factors. The number of time points that a child had ratings of behavior problems differed as a function of the child's sex and ethnicity, and the family's income-to-needs ratio. Girls had more time points of ratings on average compared to boys (t[1,360.70] = -2.05, p = .040). African Americans (t[214.89] = 3.28, p = .001) but not compared Hispanics (t[92.03] = 0.63, p = .532) had fewer ratings than other racial/ethnic groups. The children's number of time points of ratings was positively associated with the families' income-to-needs ratio (r[1,271] = .12, p < .001). Therefore, we included the child's sex, the child's ethnicity, and the family's income-to-needs ratio as covariates in the final growth curve models. We also observed some systematic missingness of behavior problem scores as a function of a predictor, delay of gratification. Delay of gratification was positively associated with children's number of time points of behavior problem ratings such that children with greater delay of gratification had more time points of behavior problem ratings (r[959] = .07, p = .038). However, the child's number of time points of behavior problem ratings was not associated with their negative emotionality.

How We Handled Missing Data

We modeled behavior problem trajectories using mixed models. Mixed models analyze data in long format, where each participant has multiple rows: i.e., one row for each informant-by-timepoint combination. Therefore, the analyses use all available data on each child across the measurement occasions when they have scores on the predictors. For example, if a child drops

out of the study after the first two measurement occasions, mixed models still use the child's data for the first two measurement occasions. Mixed models assume that the data are missing at random or completely at random. We did not use multiple imputation because multiple imputation can lead to unstable results when fitting mixed models (Twisk et al., 2013).

Supplementary Appendix S7. Sensitivity Analysis Methods.

Mother-Reported Trajectories

We conducted a sensitivity analysis to examine trajectories using only those ratings by the informant type who provided the most ratings on average, i.e., mothers. To assess trajectories of children's behavior problems from mother report, we used the same mixed methods approach as the primary analyses, using the lmer function of the lme4 package (Bates et al., 2015) in R.

Setting Intercepts to the First Timepoint When Informant Type Provided Ratings

In the original models, we set the intercepts to the last time point (age 15). We included dummy-coded variables in the models to examine whether particular informant types (e.g., fathers) differed in their ratings on average compared to the reference informant type (i.e., mothers). However, some informant types did not provide ratings at age 15. Thus, to determine whether there were mean-level differences in ratings by informant type, we conducted sensitivity analyses in which we set the intercepts to the first timepoint when the target informant type provided ratings (father: age 6; teacher: age 5; afterschool caregiver: age 6; other caregiver: age 2). For instance, for the model comparing mother- versus teacher-report, we conducted a sensitivity analysis to set the intercepts of behavior problems to age 5.

Excluding Ratings Before 54 Months

The first timepoint that the outcome variables (ratings of internalizing and externalizing problems) were assessed was at 24 months of age, whereas the predictor variables (delay of gratification and negative emotionality) were assessed at later ages (54 months of age). To avoid reverse prediction (e.g., variables at age 54 months of age as "predictors" of outcomes at 24 months), we conducted a sensitivity analysis that excluded ratings of psychopathology prior to 54 months of age. Doing so placed the starting point of outcomes and predictors at the same age,

to reduce the likelihood that associations reflected effects of earlier levels of the outcomes. Intercepts of growth curves of internalizing and externalizing problems were set to 15 years of age, the same intercept as the primary analyses.

Early Cognitive Ability

As a sensitivity analysis, we examined early cognitive ability as a potential confound in the association between delay of gratification and negative emotionality on specific and general psychopathology. The child's cognitive ability at age 24 months was assessed using the Bayley Scales of Infant Development (Bayley, 1969). The Bayley Scales consist of play tasks. Raw scores were converted to age-normed standard scores of cognitive and mental development relative to same-aged peers.

Anger/Frustration vs. Fear

Prior research on negative emotionality has found that subdimensions of negative emotionality, including anger and fear, differentially predict internalizing and externalizing psychopathology outcomes (e.g., Dollar et al., 2022; Stifter & Dollar, 2016). In a sensitivity analysis, we examined the anger/frustration subscale and the fear subscale of the CBQ as predictors in separate analyses to determine their association with growth curves of internalizing and externalizing problems.

Mother vs. Caregiver Report of Negative Emotionality

Given the modest association between mothers' and caregivers' ratings of children's negative emotionality, we conducted sensitivity analyses examining them separately. Items and subscales that compose the negative emotionality scale from the CBQ differed slightly between caregiver and mother forms. For example, negative emotionality from mother report was derived from the Anger/Frustration, Fear, and Sadness subscales, whereas caregiver report did not

include the Fear subscale. Prior research has noted the importance of assessing multiple informants from different contexts (Kramer et al., 2023). We conducted a sensitivity analysis examining mother versus caregiver report of negative emotionality to determine how the context of behavior, i.e., informant type, influences the association between ratings of negative emotionality on internalizing and externalizing growth curves.

Aggressive vs. Delinquent Behavior

Prior research has noted that heterogeneity in externalizing problems can be parsed by separating aggressive behaviors (e.g., physically attacks others) from nonaggressive rule-breaking behaviors (e.g., lying or running away from home; Harden et al., 2015). Furthermore, there is evidence that risk and protective factors have differing associations with these subdimensions of externalizing behavior(e.g., Harden et al., 2015; Mann et al., 2018). Thus, we conducted sensitivity analyses to examine the Aggressive and Delinquent Behavior subscales separately.

Mother vs. Self-Report Bifactor Models

Prior research has noted that bifactor models derived from multiple versus single informants have differences in model fit and in the interpretation of general psychopathology (A. L. Watts et al., 2021). Our primary analyses include assessments of psychopathology from multiple informants at age 15 years. To examine differences in results as a function of the informant of the child's psychopathology at age 15 years, we separately estimated bifactor models from mother- and self-report.

Estimation of separate models mirrored the primary analyses. First, we fit a bifactor model at age 15 years with only externalizing and internalizing problem items and no predictors. The latent factors were set to be uncorrelated, so the general factor represented the covariation

among all externalizing and internalizing items. We allowed item residuals to be correlated for which the modification index was large ($\Delta\chi^2 > 20$), indicating local non-independence of items, if the modification was also consistent with theory (i.e., both items were within the same domain). After adding the covariance terms among item residuals, we added predictors. Predictors were allowed to predict the three latent factors. Then, we added covariates.

Supplementary Appendix S8. Exploratory Factor Analysis of CBQ Negative Affectivity Items.

Method

To assess whether a one-factor model is the best representation of negative emotionality items, we conducted an exploratory factor analysis (EFA). EFA was conducted using the efa() function in lavaan (Rosseel, 2012) in R. In the EFA, we examined the factor structure of the items that were included in the higher-order Negative Affectivity scale of the Children's Behavior Questionnaire (CBQ). To account for potential correlations among factors, we used a geomin oblique rotation.

Results

Results from the EFA with mother-reported items indicated that all items on the Negative Affectivity scale loaded significantly onto a single negative emotionality factor. Standardized factor loadings ranged from .11 to .59. The single factor accounted for 13% of the variance. When including a second factor, the second factor accounted for only 6% of variance. Furthermore, a number of items showed significant cross-loadings. Thus, a second factor did not explain substantial additional variance and led to complications in interpretation.

Results from the caregiver-report showed even more confidence in a single factor. All items on the Negative Affectivity scale loaded significantly onto a negative emotionality factor. Standardized factor loadings ranged from .26 to .76. A single negative emotionality factor accounted for 31% of variance. A second factor accounted for only 6% of the variance, with many items showing significant cross-loadings.

Taken together, these findings suggest that—like most psychological data—these data are not truly unidimensional. However, a single factor accounted for a substantial portion of the

variance. A second factor did not account for substantial additional variance and led to complications in interpretation due to significant cross-loadings. Thus, given our goals to examine overall negative emotionality, we examined a composite of general negative emotionality across all items. However, to examine potential distinct effects of fear versus anger subdimensions, we also conducted sensitivity analyses that examined fear and anger subscales of the CBQ separately (see Supplementary Appendices S7, S9–10).

Supplementary Appendix S9. Sensitivity Analysis Results: Growth Curve Models

Early Cognitive Ability

Higher early cognitive ability was associated with lower intercepts of externalizing problems (β = -.01, SE = .02, $p < .001$), but not with differences in slope (β = -.00, SE = .01, p = .815). When accounting for early cognitive ability, negative emotionality was associated with higher intercepts (β = .18, SE = .02, $p <.001$) and steeper declines in externalizing problems (β = -.03, SE = .01, $p <.001$), which did not differ from primary analyses. The significant association between poorer delay gratification and higher intercepts of externalizing problems in the primary analyses was attenuated to trend-level significance when accounting for early cognitive problems (β = -.03, SE = .02, p = .079). The slope remained nonsignificant (β = .00, SE = .01, p = .565).

Higher early cognitive ability was associated with lower intercepts of internalizing problems (β = -.08, SE = .02, $p < .001$), but not with differences in slope (β = .00, SE = .01, p = .624). When accounting for early cognitive ability, negative emotionality was associated with higher intercepts (β = .14, SE = .01, $p < .001$), and steeper declines in slope of internalizing problems (β = -.02, SE = .01, p = .025), which did not differ from primary analyses. The significant association between greater delay of gratification and lower intercepts of internalizing problems was no longer significant when accounting for early cognitive ability (β = -.01, SE = .02, p = .491). The slope remained nonsignificant (β = .00, SE = .01, p = .968). Taken together, these results indicate that early cognitive ability accounts for a significant portion of variance in the association between delay of gratification and the intercepts, but not slopes, of internalizing and externalizing problems**.**

Mother-Reported Trajectories

When examining trajectories of mother-reported externalizing problems, higher negative

emotionality was associated with higher intercepts (β = .21, SE = .02, *p* <.001), but not with differences in slope (β = -.00, SE = .01, *p* = .997). These results were consistent with primary analyses. Greater delay of gratification was marginally significantly associated with lower intercepts (β = -.04, SE = .02, *p* = .050), but not with differences in slope (β = -.00, SE = .01, *p* = .634). These results differ slightly from the primary analyses such that greater delay of gratification was marginally significantly associated with lower intercepts.

Predicting internalizing problems, higher negative emotionality was associated with higher intercepts (β = .20, SE = .02, *p* <.001), but not with differences in slope (β = -.00, SE = .01, *p* = .946). By contrast, primary analyses indicated that higher negative emotionality was associated with steeper declines in slope. Delay of gratification was also not associated with differences in intercept (β = -.00, SE = .02, *p* = .901) or slope (β = .01, SE = .01, *p* = .240). By contrast, the primary analyses indicated that delay of gratification was associated with lower intercepts of internalizing problems. Findings in predicting slopes were consistent with primary analyses.

Setting Intercepts to Informant's First Rating

Mother

Because mother report was the reference group in all models, we did not fit additional models for mother report to set the intercepts at the first timepoint they provided ratings.

Father

We fit a model with intercepts set to age 6, the first timepoint when fathers provided ratings. Compared to mothers' ratings, fathers' ratings showed lower intercepts of externalizing and internalizing problems.

Teacher

We fit a model with intercepts set to age 5, the first timepoint when teachers provided ratings. Compared to mothers' ratings, teachers' ratings showed lower intercepts of externalizing and internalizing problems.

Afterschool Caregiver

We fit a model with intercepts set to age 6, the first timepoint when afterschool caregivers provided ratings. Compared to mothers' ratings, afterschool caregivers' ratings showed lower intercepts of externalizing and internalizing problems.

Other Caregiver

We fit a model with intercepts set to age 2, the first timepoint when other caregivers provided ratings. Compared to mothers' ratings, other caregivers' ratings showed higher intercepts of externalizing and internalizing problems. In the original model (with intercepts set to age 15), other caregivers' ratings showed *lower* intercepts than mother's ratings, but this was an artifact of setting intercepts to ages when other caregivers did not provide ratings. In sum, compared to mothers, other caregivers tended to rate children as showing higher levels of internalizing and externalizing problems.

Self-Report

Intercepts in the main models were already set to the first timepoint when adolescents provided self-reported ratings (age 15). Therefore, we did not fit additional models for self-report.

Excluding Ratings Before 54 Months

When excluding behavior problem ratings before age 54 months, negative emotionality was associated with higher intercepts ($\beta = .21$, SE = .02, $p < .001$) and steeper declines ($\beta = -.05$, SE = .01, $p < .001$) in externalizing problems over time. Delay of gratification was associated

with lower intercepts of externalizing problems (β = -.04, SE = .02, p <.001), but not with differences in slopes (β = .01, SE = .01, p = .150). Results excluding ratings before 54 months did not change results from primary analyses.

When excluding behavior problem ratings before age 54 months, negative emotionality was associated with higher intercepts (β = .14, SE = .02, p <.001) and steeper declines (β = -.02, SE = .01, p = .038) in internalizing problems over time. Delay of gratification was associated with lower intercepts at a trend level (β = -.01, SE = .01, p = .056), and was not significantly associated with differences in slopes of internalizing problems (β = .01, SE = .01, p = .503). These results excluding ratings before 54 months also did not change the results from primary analyses.

Anger/Frustration vs. Fear

Anger/Frustration

Anger/frustration was associated with higher intercepts (β = .23, SE = .02, p <.001) and steeper declines (β = -.03, SE = .01, p <.001) in externalizing problems over time. These results replicated findings from prior research with the same sample that anger at 54 months predicted intercepts (β = .34) and slopes (β = -.08) of mother-reported externalizing problems (Crockett et al., 2018). These findings aligned with the primary analyses. Anger/frustration was also associated with higher intercepts (β = .15, SE = .01, p <.001) and steeper declines (β = -.02, SE = .01, p = .051) at a trend level in internalizing problems over time.

Fear

Fear was not significantly associated with intercepts (β = -.01, SE = .02, p = .709) or slopes (β = -.00, SE = .01, p = .775) of externalizing problems. By contrast, fear was associated with higher intercepts (β = .05, SE = .02, p = .003) and steeper declines (β = -.02, SE = .01, p =

.003) in internalizing problems over time. Taken together, the subscales of negative emotionality—fear and anger/frustration—show differential associations with trajectories of internalizing and externalizing problems. As would be expected based on theory, anger/frustration was more strongly associated with externalizing problems, whereas fear was more strongly associated with internalizing problems. These results highlight the importance of assessing the different facets of negative emotionality.

Mother versus Caregiver Report of Negative Emotionality

Mother-Reported

When examining mother-reported negative emotionality, negative emotionality was associated with higher intercepts (β = .13, SE = .02, p <.001), and was not associated with slopes (β = -.01, SE = .01, p = .112) in externalizing problems over time. These results were similar to primary results, but the association with slopes was attenuated to non-significance when examining mother-reported negative emotionality. Delay of gratification was associated with lower intercepts (β = -.07, SE = .02, p <.001), but not with differences in slopes (β = .01, SE = .01, p = .365) in externalizing problems.

Mother-reported negative emotionality was associated with higher intercepts (β = .13, SE = .01, p <.001), and with a steeper decrease in slopes (β = -.01, SE = .01, p = .021) in internalizing problems. Unlike with externalizing problems, negative emotionality was significantly associated with slopes of internalizing problems when examining mother-reported negative emotionality.

Caregiver-Reported

When examining caregiver-reported negative emotionality, negative emotionality was associated with higher intercepts (β = .15, SE = .02, p <.001) and steeper declines (β = -.04, SE =

.01, p <.001) in externalizing problems over time. These results align with the results of the primary analyses. Delay of gratification was associated with lower intercepts (β = -.06, SE = .02, p = .005) but not with differences in slopes (β = .01, SE = .01, p = .261) in externalizing problems.

Caregiver-reported negative emotionality was associated with higher intercepts (β = 08, SE = .02, p <.001) and steeper declines at a trend level (β = -.01, SE = .01, p = .070) in internalizing problems. Results of the intercepts were the same as the primary results, but the association with slopes of internalizing problems was attenuated to trend-level significance when examining caregiver-reported negative emotionality.

Aggressive vs. Delinquent Behavior

Aggressive Behaviors

Negative emotionality was associated higher intercepts (β = .18, SE = .02, p <.001) and steeper declines (β = -.03, SE = .01, p <.001) in aggressive behaviors over time. Delay of gratification was associated with lower intercepts (β = -.07, SE = .02, p <.001) but not differences in slopes (β = .01, SE = .01, p = .150) in aggressive behaviors over time. These results align with the primary analyses.

Delinquent Behaviors

Negative emotionality was associated with higher intercepts (β = .09, SE = .02, p <.001) and steeper declines (β = -.02, SE = .01, p = .016) in delinquent behaviors over time. These results align with the primary analyses. Delay of gratification was associated with lower intercepts (β = -.07, SE = .02, p <.001) but not with differences in slopes (β = .01, SE = .01, p = .289) of delinquent behaviors over time. These results were the same as the primary analyses.

Supplementary Appendix S10. Sensitivity Analysis Results: Bifactor Models.

Early Cognitive Ability

Regression coefficients of the model including early cognitive ability as a covariate are in Supplementary Table S12. When controlling for early cognitive ability and demographic characteristics, negative emotionality was not associated with unique internalizing problems (β = .01, *p* = .857), but was significantly associated with the general factor (β = .09, *p* = .021) and unique externalizing problems (β = .08, *p* = .048). Delay of gratification was not associated with general psychopathology (β = -.02, *p* = .557), or unique internalizing (β =.04, *p* = .269) and externalizing problems (β = -.03, *p* = .294). Early cognitive ability was negatively associated with the general factor at a trend level (β = -.07, *p* = .073), but was not associated with unique internalizing (β = -.04, *p* = .245) or externalizing problems (β = -.02, *p* = .557).

The results indicated that when controlling for early cognitive ability, in addition to other demographic characteristics, associations between negative emotionality and specific and general psychopathology did not differ, with one exception: its association with specific externalizing problems became statistically significant. The nonsignificant association between delay of gratification and specific and general behavior problems remained when controlling for early cognitive ability. Results contradict prior findings that implicate early cognitive ability as a potential common cause between unique externalizing problems and delay of gratification (Ursache et al., 2013; T. W. Watts et al., 2018).

Anger/Frustration vs. Fear

Anger/Frustration

Anger/frustration was positively associated with general psychopathology (β = 0.13, *p* = .001) and unique externalizing problems (β = 0.09, *p* = .044), but not with unique internalizing

problems (β = 0.01, *p* = .815).

Fear

Fear was not significantly associated with general psychopathology (β = -0.03, *p* = .447, unique externalizing problems (β = 0.02, *p* = .514), or unique internalizing problems (β = 0.04, *p* = .307).

Mother versus Caregiver Report of Negative Emotionality

Mother-Reported

Mother-reported negative emotionality was positively associated with general psychopathology (β = 0.08, *p* = .005), unique externalizing problems (β = 0.10, *p* = .004), and with unique internalizing problems at a trend level (β = 0.05, *p* = .071).

Caregiver-Reported

Caregiver-reported negative emotionality was not significantly associated with general psychopathology (β = 0.06, *p* = .271), unique externalizing problems (β = 0.05, *p* = .321), or unique internalizing problems (β = - 0.04, *p* = .361).

Aggressive vs. Delinquent Behavior

When separating aggressive from delinquent behavior into separate factors to replace the externalizing problems factor, negative emotionality was not associated with unique aggressive behavior (β = .06, *p* = .206), but was significantly associated with general psychopathology (β = .08, *p* = .023) and unique delinquent behavior (β = .10, *p* = .006). Delay of gratification was not significantly associated with general psychopathology (β = -.04, *p* = .238), unique aggressive behavior (β = -.03, *p* = .476), or unique delinquent behavior (β = -.03, *p* = .280).

Mother vs. Self-Report Bifactor Models

Mother Report

The mother-report model fit well according to RMSEA (.040) and SRMR (.043) and had acceptable fit according to CFI (.922). Therefore, we added predictors to the measurement model, then separately added predictors, and finally added covariates. Negative emotionality was positively associated with general psychopathology (β = .22, *p* = .016) and unique externalizing problems (β = .23, *p* = .016), but not with unique internalizing problems (β = .11, *p* = .115). Delay of Gratification was negatively associated with general psychopathology (β = -.12, *p* = .008), positively associated with unique internalizing problems (β = .15, *p* = .002), but was not associated with unique externalizing problems (β = .02, *p* = .724).

Upon adding covariates, negative emotionality was no longer significantly associated with general psychopathology (β = .11, *p* = .103), but was now significantly associated with unique internalizing problems at a trend level (β = .12, *p* = .090). Delay of gratification was also no longer significantly associated with general psychopathology after controlling for covariates (β = -.05, *p* = .286). Children who had lower early cognitive abilities had higher general psychopathology at a trend level. When compared to non-African Americans, African Americans showed lower ratings of unique internalizing and externalizing problems. Females, compared to males were associated with higher internalizing problems.

Self-Report

The self-report model fit well according to RMSEA (.036) and SRMR (.052) but did not fit well according to CFI (.893), even when adding correlated residuals based on modification indices. We caution interpretation of these findings due to the model fit; nonetheless, we added predictors to the measurement model, then separately added predictors, and finally covariates. Negative emotionality was positively associated with general psychopathology (β = .08, *p* = .087) at a trend level, but was not associated with unique externalizing problems (β = .07, *p* =

.203) or unique internalizing problems (β = -.02, *p* = .661). Delay of gratification was negatively associated with general psychopathology (β = -.10, *p* = .032), but was not associated with unique externalizing problems (β = .00, *p* = .969) or unique internalizing problems (β = .03, *p* = .527).

Upon adding covariates, negative emotionality was no longer significantly associated with general psychopathology (β = .04, *p* = .417). Delay of gratification was also no longer associated with general psychopathology (β = -.00, *p* = .953). Females, compared to males, showed higher general psychopathology and lower unique internalizing and externalizing problems. When compared to non-African Americans, African Americans had lower general psychopathology, at a trend level, and lower unique externalizing problems, but they showed higher unique internalizing problems. When compared to non-Hispanics, Hispanics showed lower general psychopathology. A higher income-to-needs ratio was associated with lower general psychopathology and higher unique internalizing problems.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bayley, N. (1969). *Manual for the Bayley Scales of Infant Development*. The Psychological Corporation.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Crockett, L. J., Wasserman, A. M., Rudasill, K. M., Hoffman, L., & Kalutskaya, I. (2018). Temperamental anger and effortful control, teacher–child conflict, and externalizing behavior across the elementary school years. *Child Development*, *89*(6), 2176–2195. https://doi.og/10.1111/cdev.12910

Dollar, J. M., Perry, N. B., Calkins, S. D., Shanahan, L., Keane, S. P., Shriver, L., & Wideman, L. (2022). Longitudinal associations between specific types of emotional reactivity and psychological, physical health, and school adjustment. *Development and Psychopathology*, 1–15. Cambridge Core. https://doi.org/10.1017/S0954579421001619

Harden, K. P., Patterson, M. W., Briley, D. A., Engelhardt, L. E., Kretsch, N., Mann, F. D., Tackett, J. L., & Tucker-Drob, E. M. (2015). Developmental changes in genetic and environmental influences on rule-breaking and aggression: Age and pubertal development. *Journal of Child Psychology and Psychiatry*, *56*(12), 1370–1379. https://doi.org/10.1111/jcpp.12419

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices, 3rd ed.* (pp. xxvi, 566). Springer Science + Business Media. https://doi.org/10.1007/978-1-4939-0317-7

Kramer, E., Willcutt, E. G., Peterson, R. L., Pennington, B. F., & McGrath, L. M. (2023). Processing speed is related to the general psychopathology factor in youth. *Research on Child and Adolescent Psychopathology*. https://doi.org/10.1007/s10802-023-01049-w

Mann, F. D., Tackett, J. L., Tucker-Drob, E. M., & Harden, K. P. (2018). Callous-unemotional traits moderate genetic and environmental influences on rule-breaking and aggression: Evidence for gene× trait interaction. *Clinical Psychological Science*, *6*(1), 123–133. https://doi.org/10.1177/2167702617730889

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*(4), 728. https://doi.org/10.1037/a0018966

Min, K.-S. (2007). Evaluation of linking methods for multidimensional irt calibrations. *Asia Pacific Education Review*, *8*(1), 41–55. https://doi.org/10.1007/BF03025832

Murray, J., & Farrington, D. P. (2010). Risk factors for conduct disorder and delinquency: Key findings from longitudinal studies. *The Canadian Journal of Psychiatry*, *55*(10), 633–642. https://doi.org/10.1177/070674371005501003

Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement*, *37*(4), 357–373. JSTOR.

Petersen, I. T., & LeBeau, B. (2022). Creating a developmental scale to chart the development of psychopathology with different informants and measures across time. *Journal of*

Psychopathology and Clinical Science, *131*, 611–625. https://doi.org/10.1037/abn0000649

Petersen, I. T., Lindhiem, O., LeBeau, B., Bates, J. E., Pettit, G. S., Lansford, J. E., & Dodge, K. A. (2018). Development of internalizing problems from adolescence to emerging adulthood: Accounting for heterotypic continuity with vertical scaling. *Developmental Psychology*, *54*(3), 586–599. https://doi.org/10.1037/dev0000449

R Core Team. (2022). *R: A language and environment for statistical computing*.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Stifter, C., & Dollar, J. (2016). Temperament and developmental psychopathology. In D. Cicchetti (Ed.), *Developmental psychopathology: Risk, resilience, and intervention* (pp. 546–607). John Wiley & Sons, Inc.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Twisk, J., de Boer, M., de Vente, W., & Heymans, M. (2013). Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *Journal of Clinical Epidemiology*, *66*(9), 1022–1028. https://doi.org/10.1016/j.jclinepi.2013.03.017

Ursache, A., Blair, C., Stifter, C., & Voegtline, K. (2013). Emotional reactivity and regulation in infancy interact to predict executive functioning in early childhood. *Developmental Psychology*, *49*(1), 127–137. APA PsycArticles. https://doi.org/10.1037/a0027728

Wall, A. E., & Barth, R. P. (2005). Aggressive and delinquent behavior of maltreated adolescents: Risk factors and gender differences. *Stress, Trauma, and Crisis*, *8*(1), 1–24. https://doi.org/10.1080/15434610490888081

Watts, A. L., Makol, B. A., Palumbo, I. M., De Los Reyes, A., Olino, T. M., Latzman, R. D., DeYoung, C. G., Wood, P. K., & Sher, K. J. (2021). How robust is the p factor? Using multitrait-multimethod modeling to inform the meaning of general factors of youth psychopathology. *Clinical Psychological Science*, 21677026211055170. https://doi.org/10.1177/21677026211055170

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science*, *29*(7), 1159–1177. https://doi.org/10.1177/0956797618761661

Weeks, J. P. (2010). plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software*, *35*, 1–33.

Supplementary Table S1

Internal Consistency Estimates by Age and Rater

	Age (Years)										
Cronbach's Alpha	2	3	4	5	6	7	8	9	10	11	15
Mother	.88 .82	.89 .84	.88 .83	.88 .83	.88 .81	–	.89 .86	.89 .85	.89 .85	.89 .87	.91 .87
Father	–	–	–	–	.88 .84	–	.87 .85	.90 .85	.91 .88	.91 .88	.92 .90
Teacher	–	–	–	.94 .86	.94 .85	.94 .88	.95 .87	.95 .85	.95 .86	.95 .87	–
Afterschool Caregiver	–	–	–	–	.92 .87	–	.92 .83	.91 .86	.91 .88	–	–
Other Caregiver	.90 .87	.92 .86	.96 .89	–	–	–	–	–	–	–	–
Self-Report	–	–	–	–	–	–	–	–	–	–	.86 .89

	Age (Years)										
Omega	2	3	4	5	6	7	8	9	10	11	15
Mother	.88 .82	.89 .84	.89 .83	.90 .84	.90 .82	–	* .86	* .86	.90 .85	* .87	.92 .87
Father	–	–	–	–	* .84	–	* .85	.91 .86	* .88	.92 .88	.92 .91
Teacher	–	–	–	* 88	.95 .86	.95 .88	* .87	* .85	* .87	.96 .88	–
Afterschool Caregiver	–	–	–	–	* .87	–	* .84	* .87	* *	–	–
Other Caregiver	.90 .87	.92 .86	.96 .89	–	–	–	–	–	–	–	–

Self-Report	–	–	–	–	–	–	–	–	–	–	.86 .89

Note. "–" indicates not applicable because the particular rater did not provide ratings at the given time point; * = unable to be estimated. Internal consistency estimates for externalizing problems are the top number in each box, whereas internal consistency estimates for internalizing problems are the bottom number.

Supplementary Table S2

One-Year Cross-Time Rank-Order Stability Estimates (r-value) by Rater

Externalizing Problems

Informant	Mean	Min	Max
Mother	0.73	0.63	0.80
Father	0.76	0.75	0.76
Teacher	0.63	0.53	0.68
Afterschool Caregiver	0.63	0.56	0.69
Other Caregiver	0.39	0.39	0.39

Internalizing Problems

Informant	Mean	Min	Max
Mother	0.67	0.52	0.75
Father	0.67	0.64	0.69
Teacher	0.27	0.14	0.33
Afterschool Caregiver	0.52	0.45	0.56
Other Caregiver	0.33	0.33	0.33

Supplementary Table S3

Percentage of Participants with Scores on Behavior Problems by Rater Type at Different Numbers of Time Points

	# of Time Points											
Rater	0	1	2	3	4	5	6	7	8	9	10	11
Any	7.6	2.1	4.8	1.7	1.7	2.8	1.8	2.5	2.0	4.4	13.0	55.7
Mother	8.1	2.2	4.8	1.7	2.2	3.9	2.2	3.8	4.2	11.4	55.6	n/a
Father	26.0	9.0	5.9	7.0	8.8	13.6	29.8	n/a	n/a	n/a	n/a	n/a
Teacher	17.2	1.8	3.0	3.7	5.6	8.6	20.3	40.0	n/a	n/a	n/a	n/a
Afterschool Caregiver	67.2	15.2	8.1	6.2	3.4	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Other Caregiver	27.3	25.1	20.7	26.8	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Self-Report	29.8	70.2	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Note: “n/a” indicates not applicable because, across the timeframe of the present study, the rater type was not given the opportunity to provide ratings that number of times.

Supplementary Table S4

Linking Constants for Linking Scores from Different Raters and at Different Ages

Rater linked from	Rater linked to	Age linked from	Age linked to	*A*	*B*
Afterschool Caregiver	–	8	6	1.150, 0.003 -0.000, 1.470	-0.202, 2.981
Afterschool Caregiver	–	9	8	0.958, -0.010 -0.001, 0.578	-0.088, -1.680
Afterschool Caregiver	–	10	9	1.181, 0.005 -0.008, 0.825	-0.164, -2.339
Father	–	8	6	1.008, -0.002 0.006, 1.131	-0.251, 0.047
Father	–	9	8	1.131, 0.003 -0.006, 0.963	-0.073, -0.013
Father	–	10	9	1.098, -0.001 -0.001, 1.153	-0.190, -0.116
Father	–	11	10	0.949, 0.011 -0.016, 0.985	0.097, -0.058
Father	–	15	11	1.052, 0.017 -0.024, 1.218	-0.086, -0.134
Mother	–	2	3	0.988, -0.003 -0.001, 0.737	0.160, -0.521
Mother	–	3	4	1.014, 0.006 -0.004, 1.059	-0.030, 0.081
Mother	–	4	5	0.835, -0.005 0.003, 0.946	0.641, 0.486
Mother	–	5	6	0.910, 0.004, -0.002, 1.350	0.176, -0.279
Mother	–	8	6	1.009, 0.004 -0.004, 1.083	-0.122, -0.035
Mother	–	9	8	1.038, -0.019	-0.194, 0.009

				0.024, 1.000	
Mother	–	10	9	1.090, 0.008 -0.014, 0.923	-0.065, 0.116
Mother	–	11	10	1.008, 0.003 -0.002, 1.117	-0.059, -0.133
Mother	–	15	11	1.109, 0.017 -0.016, 1.869	-0.178, 0.007
Other Caregiver	–	2	3	1.096, 0.001 -0.001, 0.574	-0.136, 2.458
Other Caregiver	–	3	4	1.296, 0.001 -0.002, 1.054	-1.016, -0.792
Teacher	–	5	6	0.969, 0.001 0.000, 1.616	-0.145, -0.368
Teacher	–	7	6	1.065, 0.001 -0.001, 0.874	0.033, -0.829
Teacher	–	8	7	1.043, -0.003 0.007, 1.073	0.072, 0.922
Teacher	–	9	8	0.970, -0.001 -0.001, 0.810	-0.170, -1.486
Teacher	–	10	9	0.981, 0.003 -0.008, 0.941	0.190, 0.196
Teacher	–	11	10	1.216, -0.014 0.024, 1.456	-0.288, 1.033
Other Caregiver	Mother	2	–	0.981, -0.050 0.020, 2.788	0.497, -0.672
Father	Mother	6	–	1.092, 0.000 0.000, 0.898	-0.143, -0.164
Afterschool Caregiver	Mother	6	–	5.977, 0.000 0.000, 1.942	-1.815, 1.211
Teacher	Mother	6	–	1.606, 0.000 0.000, 1.109	-0.596, -0.045
Self-Report	Mother	15	–	0.973, -0.001 0.004, 1.845	0.231, 0.253

Note. "–" indicates that scores were linked to the same rater role or age. "A" = slope linking matrix. "B" = intercept linking vector.

Supplementary Table S5

Estimates of (Post Linking) Scale-Level Differential Item Functioning (DIF) Between Measures That Were Used to Link Scores Across Different Raters and Ages

Rater linked from	Rater linked to	Aged linked from	Age linked to	UDIF	SDIF
Afterschool Caregiver	–	8	6	0.109 0.046	0.109 0.046
Afterschool Caregiver	–	9	8	0.002 0.002	-0.002 -0.000
Afterschool Caregiver	–	10	9	0.002 0.004	0.000 0.004
Father	–	8	6	0.016 0.033	0.016 0.033
Father	–	9	8	0.000 0.000	-0.000 -0.000
Father	–	10	9	0.001 0.000	0.001 -0.000
Father	–	11	10	0.001 0.001	-0.000 -0.001
Father	–	15	11	0.001 0.001	-0.000 -0.001
Mother	–	2	3	0.001 0.002	-0.001 -0.001
Mother	–	3	4	0.001 0.003	0.001 0.003
Mother	–	4	5	0.001 0.001	-0.001 -0.001
Mother	–	5	6	0.033 0.057	-0.033 -0.057
Mother	–	8	6	0.030 0.053	0.030 0.053
Mother	–	9	8	0.002 0.002	0.002 0.000
Mother	–	10	9	0.002 0.000	0.000 -0.000
Mother	–	11	10	0.001 0.001	-0.001 -0.000
Mother	–	15	11	0.001 0.002	-0.001 -0.002
Other Caregiver	–	2	3	0.016 0.090	0.016 0.090
Other Caregiver	–	3	4	0.002	-0.001

				0.001	-0.001
Teacher	–	5	6	0.002	0.001
				0.001	0.001
Teacher	–	7	6	0.023	-0.023
				0.014	-0.014
Teacher	–	8	7	0.025	0.025
				0.015	0.015
Teacher	–	9	8	0.003	0.003
				0.002	0.001
Teacher	–	10	9	0.002	-0.002
				0.002	-0.001
Teacher	–	11	10	0.002	-0.001
				0.001	-0.001
Other Caregiver	Mother	2	–	0.052	-0.052
				0.063	-0.063
Father	Mother	6	–	0.011	-0.005
				0.003	-0.001
Afterschool Caregiver	Mother	6	–	0.092	-0.069
				0.033	0.033
Teacher	Mother	6	–	0.022	-0.022
				0.010	-0.010
Self-Report	Mother	15	–	0.011	0.010
				0.024	0.024

Note. “UDIF” = Unsigned DIF effect size statistic; “SDIF” = signed DIF effect size statistic. Externalizing DIF statistics are presented in the top row of each cell; internalizing problems DIF statistics are presented in the bottom row of each cell.

Supplementary Table S6

Regression Coefficients from Growth Curve Models

	Externalizing					Internalizing				
Parameter	**B**	**β**	**SE**	***df***	***p***	**B**	**β**	**SE**	***df***	***p***
Intercept	-0.04	-1.37	0.02	1462.88	.030	0.07	-0.82	0.02	1640.20	< .001
Age (centered at 15)	-72.56	-1.55	1.51	11195.14	< .001	0.07	-0.77	2.04	2399.46	.974
Afterschool Caregiver	-0.25	-4.08	0.16	23280.28	.114	-1.41	-1.91	0.17	22532.43	< .001
Other Caregiver	-14.72	-0.08	1.34	23591.82	< .001	-7.50	-0.41	1.46	22817.52	< .001
Father	-0.07	-0.03	0.02	23287.65	.002	-0.14	-0.05	0.02	22474.78	< .001
Self-Report	0.68	-0.10	0.03	23220.39	< .001	0.75	-0.33	0.04	22360.33	< .001
Teacher	-0.26	0.11	0.03	23095.90	< .001	-0.93	0.11	0.03	22231.39	< .001
Age (Quadratic)	27.42	-0.48	1.88	2790.90	< .001	-4.76	-0.37	1.87	3504.15	.011
Age x Afterschool Caregiver	20.68	-4.03	17.36	23286.63	.234	-38.35	-2.31	18.80	22648.15	.041
Age x Other Caregiver	-2989.10	0.03	292.52	23603.68	< .001	-1859.49	-0.05	318.58	22828.32	< .001
Age x Father	16.58	0.03	5.53	23204.13	.003	-11.47	-0.02	5.99	22474.52	.056
Age x Teacher	68.42	0.13	4.15	23253.98	< .001	18.86	0.03	4.49	22471.20	< .001
Afterschool Caregiver x Age (Quadratic)	-45.98	-1.51	48.23	23327.43	.340	-13.44	-0.89	52.08	22565.05	.796
Other Caregiver x Age (Quadratic)	-1121.08	-0.08	118.01	23631.92	< .001	-720.29	-0.02	128.53	22850.01	< .001
Father x Age (Quadratic)	-1.52	0.00	4.54	23702.35	.738	4.91	0.01	4.91	22922.37	.318
Teacher x Age (Quadratic)	-72.60	-0.14	8.88	23142.49	< .001	-58.28	-0.10	9.60	22282.12	< .001

Note. β = standardized factor loadings; Standard error (SE) and *p* of unstandardized factor loadings; Interactions are signified by an x;

"*df*" = degrees of freedom.

Supplementary Table S7

Demographic Characteristics as Predictors of Growth Curves

	Externalizing					Internalizing				
Parameter	**B**	**β**	**SE**	***df***	***p***	**B**	**β**	**SE**	***df***	***p***
Intercept	0.08	-1.43	0.03	1340.40	.012	0.08	-0.90	0.03	1253.95	.009
Age	-70.60	-1.60	2.99	1357.31	< .001	-5.68	-0.85	3.51	1363.46	.106
Afterschool Caregiver	-0.25	-4.24	0.16	21119.50	.116	-1.35	-2.16	0.18	21260.21	< .001
Other Caregiver	-15.33	-0.08	1.36	21644.71	< .001	-8.46	-0.39	1.51	21533.88	< .001
Father	-0.07	-0.03	0.02	21111.43	.001	-0.13	-0.04	0.02	21190.31	< .001
Self-Report	0.69	-0.11	0.03	21066.62	< .001	0.76	-0.33	0.04	21088.19	< .001
Teacher	-0.29	0.11	0.03	20940.48	< .001	-0.92	0.12	0.03	20976.98	< .001
Age (Quadratic)	27.32	-0.50	1.89	2610.09	< .001	-5.71	-0.40	1.92	3316.41	.003
Female	-0.14	-0.07	0.03	1115.79	< .001	0.07	0.03	0.03	1109.29	.039
African American	0.16	0.05	0.06	1159.15	.005	0.07	0.02	0.05	1145.20	.185
Hispanic	0.10	0.02	0.07	1115.14	.158	0.05	0.01	0.07	1098.54	.482
INR	-0.03	-0.08	0.01	1146.26	< .001	-0.02	-0.05	0.01	1127.03	< .001
Age x Afterschool Caregiver	27.36	-4.20	17.72	21244.32	.123	-34.01	-2.57	19.55	21366.62	.082
Age x Other Caregiver	-3120.75	0.04	296.55	21655.50	< .001	-2070.49	-0.05	328.19	21543.73	< .001
Age x Father	16.53	0.03	5.53	21135.56	.003	-13.29	-0.02	6.10	21188.33	.029
Age x Teacher	68.24	0.13	4.19	21152.41	< .001	16.85	0.03	4.63	21189.73	< .001
Afterschool Caregiver x Age (Quadratic)	-41.78	-1.58	49.22	21137.37	.396	3.49	-1.00	54.27	21291.66	.949

(Continued)

Supplementary Table S7 Continued

Other Caregiver x Age (Quadratic)	-1175.80	-0.06	119.64	21674.51	< .001	-804.94	0.00	132.41	21562.78	< .001
Father x Age (Quadratic)	-0.82	0.00	4.53	21525.51	.857	6.36	0.01	5.00	21601.36	.203
Teacher x Age (Quadratic)	-74.76	-0.14	8.95	20975.52	< .001	-55.19	-0.10	9.89	21021.65	< .001
Age x Female	-0.25	0.00	2.85	990.58	.930	15.14	0.03	3.37	1021.65	< .001
Age x African American	1.52	0.00	4.84	1092.18	.753	-15.45	-0.02	5.72	1117.80	.007
Age x Hispanic	10.29	0.01	6.19	1026.20	.096	-1.12	0.00	7.33	1057.61	.879
Age x INR	-0.77	-0.01	0.56	1021.25	.172	0.07	0.00	0.67	1058.61	.921

Note. β = standardized factor loadings; Standard error (SE) and *p* of unstandardized factor loadings; Interactions are signified by an x; "*df*" = degrees of freedom; "INR" = income-to-needs-ratio.

Supplementary Table S8

Regression Coefficients of Predictors in the Growth Curve Models

	Externalizing					Internalizing				
Parameter	**B**	**β**	**SE**	***df***	***p***	**B**	**β**	**SE**	***df***	***p***
Negative Emotionality	0.27	0.17	0.03	893.56	< .001	0.22	0.13	0.02	904.71	< .001
Delay of Gratification	-0.02	-0.06	0.01	883.96	< .001	-0.01	-0.03	0.01	896.41	.028
Intercept	-0.94	-1.46	0.12	906.23	< .001	-0.78	-0.90	0.11	912.95	< .001
Age	-31.14	-1.62	10.67	847.49	.004	19.76	-0.85	9.38	2300.44	.035
Afterschool Caregiver	-0.21	-4.29	0.18	18052.91	.225	-1.37	-2.13	0.20	18923.19	< .001
Other Caregiver	-15.51	-0.07	1.49	18399.48	< .001	-8.33	-0.40	1.69	18900.02	< .001
Father	-0.08	-0.03	0.02	18013.75	< .001	-0.13	-0.04	0.03	18789.89	< .001
Self-Report	0.68	-0.12	0.04	17975.62	< .001	0.74	-0.33	0.04	18692.61	< .001
Teacher	-0.31	0.11	0.03	17870.04	< .001	-0.92	0.11	0.03	18636.08	< .001
Age (Quadratic)	27.89	-0.49	2.03	2262.69	< .001	-5.26	-0.40	2.11	2973.71	.013
Female	-0.14	-0.06	0.03	875.91	< .001	0.08	0.03	0.03	889.37	.013
African American	0.12	0.04	0.06	898.72	.049	0.05	0.02	0.06	913.57	.341
Hispanic	0.07	0.02	0.08	878.06	.395	-0.01	0.00	0.08	884.87	.907
INR	-0.02	-0.06	0.01	892.03	< .001	-0.01	-0.03	0.01	893.98	.055
Age x Afterschool Caregiver	33.17	-4.25	19.54	18159.23	.090	-37.77	-2.54	22.16	18955.70	.088
Age x Other Caregiver	-3155.25	0.05	324.22	18407.36	< .001	-2041.40	-0.05	367.64	18907.92	< .001
Age x Father	15.85	0.03	5.87	18024.67	.007	-15.20	-0.03	6.67	18729.26	.023
Age x Teacher	68.56	0.13	4.51	18017.71	< .001	15.63	0.03	5.13	18750.57	.002
Afterschool Caregiver x Age (Quadratic)	-27.17	-1.60	53.97	18062.51	.615	-3.32	-0.99	61.29	18966.12	.957

(Continued)

Supplementary Table S8 Continued

Other Caregiver x Age (Quadratic)	-1186.87	-0.04	130.99	18422.91	< .001	-793.33	-0.01	148.52	18930.17	< .001
Father x Age (Quadratic)	0.86	0.00	4.85	18339.78	.859	8.58	0.01	5.50	19100.28	.119
Teacher x Age (Quadratic)	-75.60	-0.14	9.62	17888.49	< .001	-58.54	-0.10	10.95	18669.79	< .001
Age x Female	-0.14	0.00	3.06	817.60	.964	16.40	0.04	2.68	2231.92	< .001
Age x African American	8.58	0.01	5.43	876.53	.115	-11.16	-0.01	4.81	2452.38	.020
Age x Hispanic	14.18	0.02	7.06	825.26	.045	-4.88	0.00	6.17	2228.66	.429
Age x INR	-1.53	-0.02	0.62	848.46	.013	-0.54	0.00	0.54	2273.04	.318
Age x Delay of Gratification	0.32	0.00	0.54	828.54	.557	0.14	0.00	0.47	2259.39	.770
Age x Negative Emotionality	-9.86	-0.03	2.36	830.62	< .001	-6.41	-0.02	2.07	2226.79	.002

Note. β = standardized factor loadings; Standard error (SE) and *p* of unstandardized factor loadings; "*df*" = degrees of freedom; "INR" = income-to-needs-ratio.

Supplementary Table S9

Standardized Factor Loadings from Bifactor Model

	General		EXT		INT	
Item	**β**	***SE***	**β**	***SE***	**β**	***SE***
CBCL 6–18 Item 3	.50	.08	.34	.11		
CBCL 6–18 Item 7	.35	.09	.28	.12		
CBCL 6–18 Item 12	.39	.08			.42	.08
CBCL 6–18 Item 14	.31	.07			.40	.09
CBCL 6–18 Item 16	.51	.07	.16	.08		
CBCL 6–18 Item 19	.43	.08	.46	.16		
CBCL 6–18 Item 20	.47	.07				
CBCL 6–18 Item 21	.45	.11				
CBCL 6–18 Item 22	.61	.10	.28	.12		
CBCL 6–18 Item 23	.61	.09				
CBCL 6–18 Item 26	.53	.09	.04	.12		
CBCL 6–18 Item 27	.44	.09	.33	.11		
CBCL 6–18 Item 31	.38	.09			.31	.09
CBCL 6–18 Item 32	.13	.10			.35	.12
CBCL 6–18 Item 33	.43	.07			.26	.08
CBCL 6–18 Item 34	.46	.06			.22	.08
CBCL 6–18 Item 35	.35	.07			.44	.09
CBCL 6–18 Item 37	.52	.08				
CBCL 6–18 Item 39	.60	.08	.01	.13		
CBCL 6–18 Item 42	.22	.07			.27	.11
CBCL 6–18 Item 43	.54	.09				
CBCL 6–18 Item 45	.42	.08			.43	.10
CBCL 6–18 Item 50	.34	.07			.44	.09
CBCL 6–18 Item 51	.33	.07			.43	.09
CBCL 6–18 Item 52	.26	.06			.46	.08
CBCL 6–18 Item 54	.34	.08			.37	.09
CBCL 6–18 Item 56A	.32	.07			.23	.10
CBCL 6–18 Item 56B	.33	.09			.24	.12
CBCL 6–18 Item 56C	.32	.08			.35	.08
CBCL 6–18 Item 56D	.25	.08			.16	.08
CBCL 6–18 Item 56E	.17	.07			.18	.09
CBCL 6–18 Item 56F	.28	.11			.27	.11
CBCL 6–18 Item 56G	.20	.05			.19	.07
CBCL 6–18 Item 57	.46	.08				
CBCL 6–18 Item 63	.41	.10	.17	.14		
CBCL 6–18 Item 65	.43	.07			.16	.08

(Continued)

Supplementary Table S9 Continued

CBCL 6–18 Item 67	.35	.11				
CBCL 6–18 Item 68	.48	.09	.13	.08		
CBCL 6–18 Item 69	.52	.09			.28	.11
CBCL 6–18 Item 71	.25	.09			.46	.11
CBCL 6–18 Item 72	.33	.06				
CBCL 6–18 Item 74	.39	.09	.23	.12		
CBCL 6–18 Item 75	.10	.08			.40	.12
CBCL 6–18 Item 80	.35	.08			.30	.11
CBCL 6–18 Item 81	.40	.10				
CBCL 6–18 Item 82	.47	.10				
CBCL 6–18 Item 86	.50	.11	.44	.14		
CBCL 6–18 Item 87	.55	.10	.29	.14		
CBCL 6–18 Item 88	.62	.09			.35	.12
CBCL 6–18 Item 89	.54	.08			.18	.09
CBCL 6–18 Item 90	.58	.08	.25	.11		
CBCL 6–18 Item 93	.40	.14	.46	.13		
CBCL 6–18 Item 94	.43	.08	.18	.09		
CBCL 6–18 Item 95	.61	.09	.33	.12		
CBCL 6–18 Item 96	.47	.10				
CBCL 6–18 Item 97	.51	.09				
CBCL 6–18 Item 101	.39	.08				
CBCL 6–18 Item 102	.38	.07			.37	.09
CBCL 6–18 Item 103	.45	.07			.48	.09
CBCL 6–18 Item 104	.46	.09	.39	.11		
CBCL 6–18 Item 105	.37	.11				
CBCL 6–18 Item 106	.38	.10				
CBCL 6–18 Item 111	.35	.06			.26	.09
CBCL 6–18 Item 112	.32	.10			.59	.11
YSR Item 18	.24	.06				
YSR Item 91	.31	.08			.15	.06
ECV	.686					
ECVs - EXT	.092					
ECVs - INT	.222					

Note. Items derived from the Child Behavior Checklist (CBCL) 6–18 & Youth Self Report (YSR); β = standardized factor loadings Standard error (SE) derived from unstandardized factor loadings; ECV = explained common variance; ECVs = explained common variance of specific factor.

Supplementary Table S10

Regression Coefficients of the Predictors in the Bifactor Model

	General Factor				Externalizing				Internalizing			
Parameter	**B**	**β**	**SE**	***p***	**B**	**β**	**SE**	***p***	**B**	**β**	**SE**	***p***
Negative Affect	0.03	0.11	0.01	.002	0.02	0.10	0.01	.016	0.00	-0.01	0.01	.853
Delay of Gratification	-0.01	-0.11	0.00	.001	0.00	0.04	0.00	.358	0.01	0.10	0.00	.007

Note. β = standardized factor loadings; Standard error (SE) and *p* of unstandardized factor loadings.

Supplementary Table S11

Regression Coefficients of the Predictors and Covariates in the Bifactor Model

	General Factor				Externalizing				Internalizing			
Parameter	**B**	**β**	**SE**	***p***	**B**	**β**	**SE**	***p***	**B**	**β**	**SE**	***p***
Negative Affect	0.02	0.09	0.01	.015	0.02	0.08	0.01	.059	0.01	0.01	0.01	.864
Delay of Gratification	0.00	-0.04	0.00	.243	0.00	-0.04	0.00	.183	0.00	0.02	0.00	.449
Father	0.01	0.04	0.01	.053	-0.01	-0.04	0.01	.057	-0.01	-0.03	0.01	.094
Self-Report	0.06	0.16	0.01	< .001	0.22	0.67	0.02	< .001	0.11	0.34	0.01	< .001
Female	-0.03	-0.09	0.01	.011	0.06	0.21	0.01	< .001	0.09	0.29	0.01	< .001
African American	0.03	0.06	0.02	.061	-0.04	-0.08	0.02	.009	-0.04	-0.07	0.02	.019
Hispanic	0.04	0.05	0.02	.082	0.02	0.02	0.03	.582	-0.02	-0.03	0.02	.319
INR	-0.01	-0.10	0.00	.007	0.00	0.01	0.00	.765	0.01	0.08	0.00	.051

Note. β = standardized factor loadings; Standard error (SE) and *p* of unstandardized factor loadings; "INR" = income-to-needs-ratio.

Supplementary Table S12

Regression Coefficients of the Predictors and Covariates, including Early Cognitive Ability, in the Bifactor Model

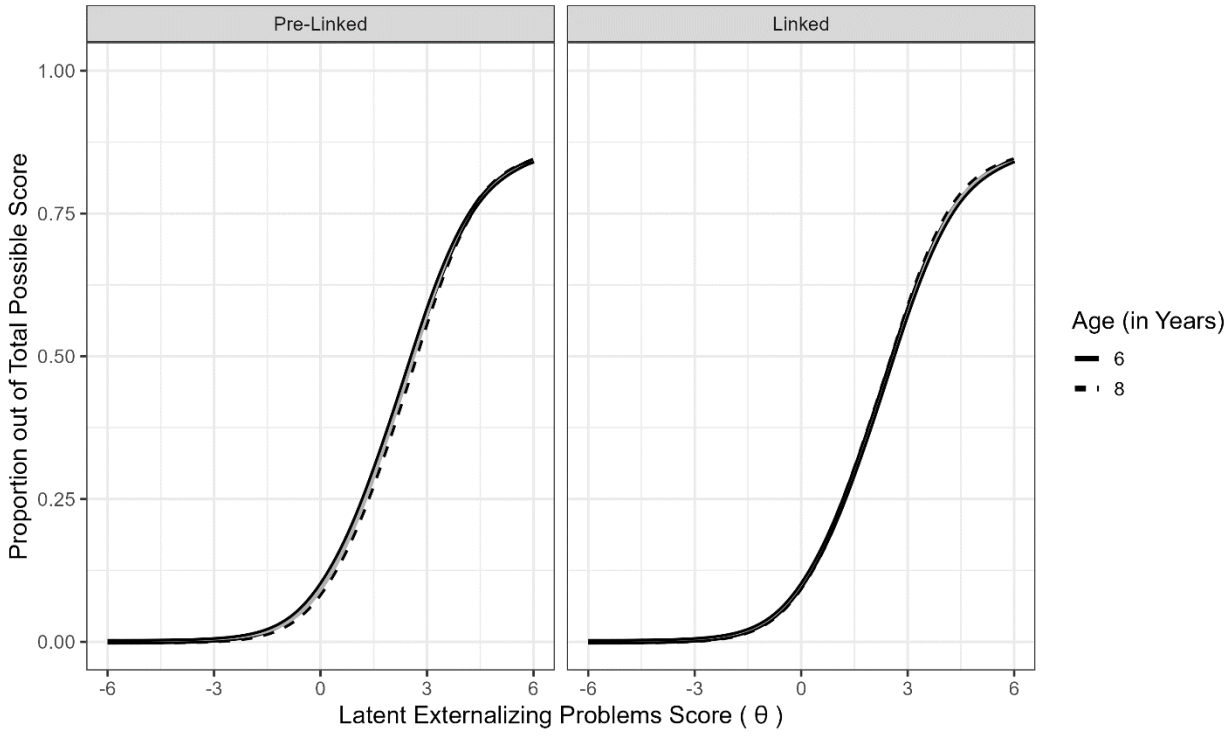
	General Factor				Externalizing				Internalizing			
Parameter	**B**	**β**	**SE**	***p***	**B**	**β**	**SE**	***p***	**B**	**β**	**SE**	***p***
Negative Affect	0.02	0.09	0.01	.021	0.02	0.08	0.01	.048	0.01	0.01	0.01	.857
Delay of Gratification	0.00	-0.02	0.00	.557	0.00	-0.03	0.00	.294	0.00	0.04	0.00	.269
Father	0.01	0.03	0.01	.089	-0.01	-0.04	0.01	.040	-0.01	-0.03	0.01	.066
Self-Report	0.06	0.17	0.01	< .001	0.21	0.66	0.02	< .001	0.11	0.34	0.01	< .001
Female	-0.02	-0.07	0.01	.056	0.06	0.21	0.01	< .001	0.09	0.30	0.01	< .001
African American	0.03	0.05	0.02	.138	-0.04	-0.07	0.02	.019	-0.04	-0.07	0.02	.022
Hispanic	0.03	0.05	0.02	.116	0.02	0.02	0.03	.539	-0.02	-0.03	0.02	.301
INR	-0.01	-0.10	0.00	.020	0.00	0.01	0.00	.668	0.01	0.08	0.00	.039
Early Cognitive Ability	0.00	-.07	0.00	.073	0.00	-0.02	0.00	.557	0.00	-0.04	0.00	.245

Note. β = standardized factor loadings; Standard error (SE) and *p* of unstandardized factor loadings; "INR" = income-to-needs-ratio.

Bayley = Bayley Scales of Infant Development.

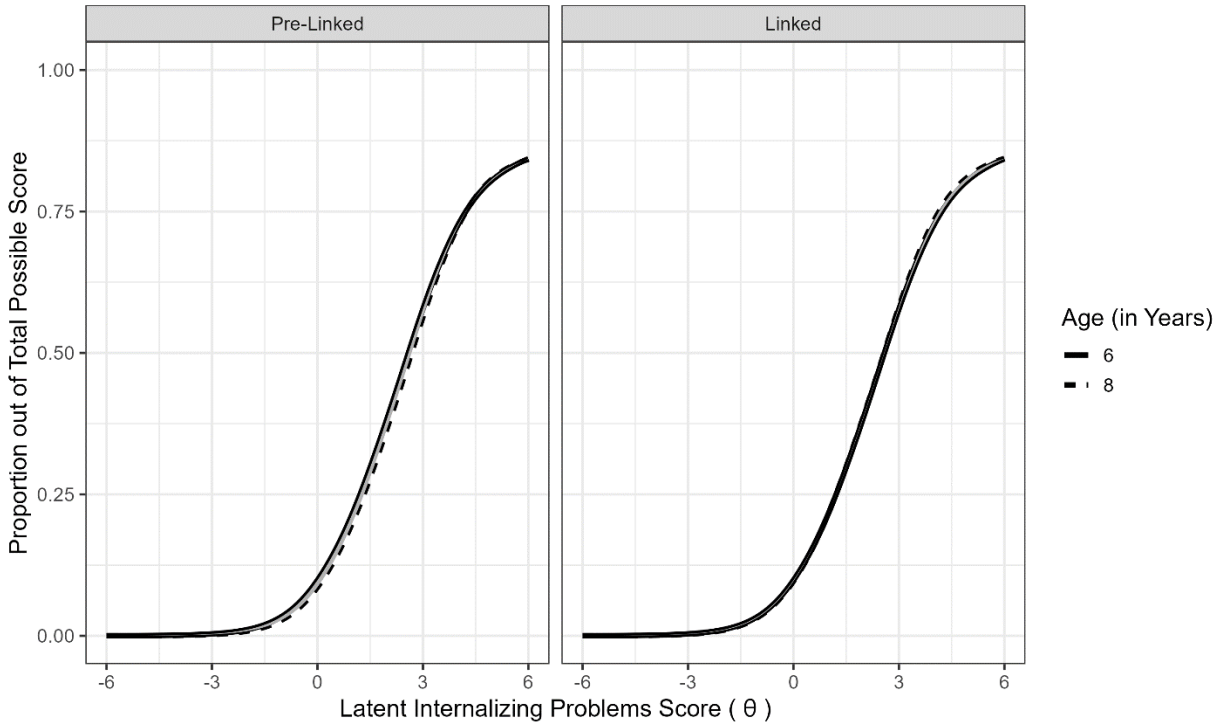
Supplementary Figure S1

Test Characteristic Curves of Pre-linked and Linked Externalizing Problem Scores for Mothers Across Two Ages



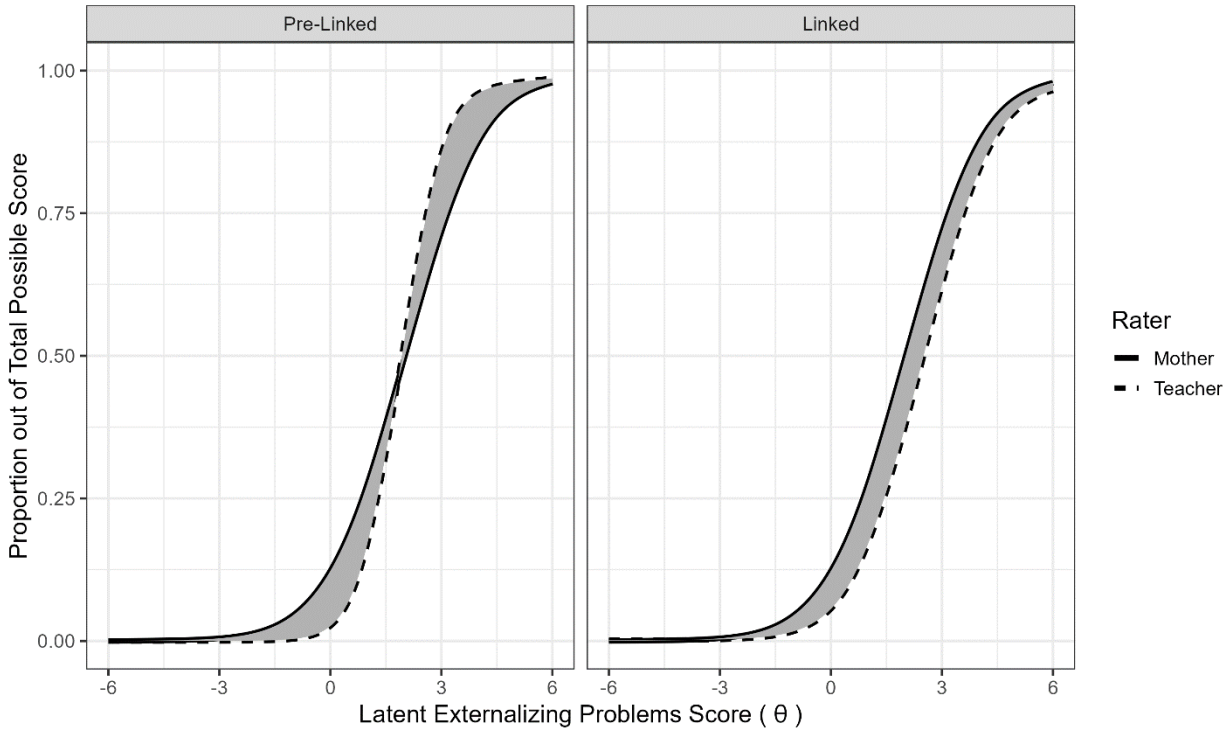
Supplementary Figure S2

Test Characteristic Curves of Pre-linked and Linked Internalizing Problem Scores for Mothers Across Two Ages



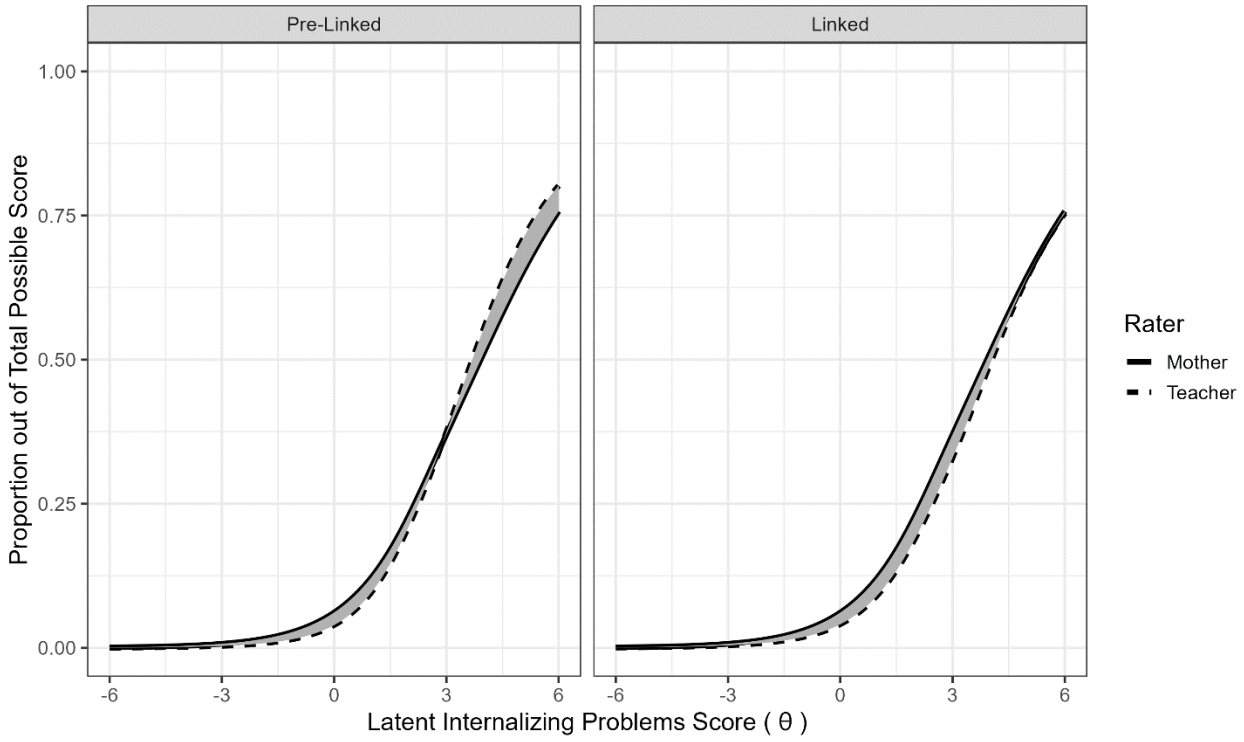
Supplementary Figure S3

Test Characteristic Curves of Pre-linked and Linked Externalizing Problem Scores Between Mothers and Teachers



Supplementary Figure S4

Test Characteristic Curves of Pre-linked and Linked Internalizing Problem Scores Between Mothers and Teachers



Supplementary Figure S5

Distribution of Item-Level Differential Item Functioning (DIF) Effect Size Statistics Between Ages by Rater Type