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:

$$P(X_{ni} = x_{ni} \mid \theta_n) = P^*_{x_{ni}}(\theta_n) - P^*_{x_{ni}+1}(\theta_n) \qquad (1)$$

where:

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

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:

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

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

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

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

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

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

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

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

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

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

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 (β = .21, SE = .02, p <.001) and steeper declines (β = -.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	0.109
				0.046	0.046
Afterschool Caregiver	–	9	8	0.002	-0.002
				0.002	-0.000
Afterschool Caregiver	–	10	9	0.002	0.000
				0.004	0.004
Father	–	8	6	0.016	0.016
				0.033	0.033
Father	–	9	8	0.000	-0.000
				0.000	-0.000
Father	–	10	9	0.001	0.001
				0.000	-0.000
Father	–	11	10	0.001	-0.000
				0.001	-0.001
Father	–	15	11	0.001	-0.000
				0.001	-0.001
Mother	–	2	3	0.001	-0.001
				0.002	-0.001
Mother	–	3	4	0.001	0.001
				0.003	0.003
Mother	–	4	5	0.001	-0.001
				0.001	-0.001
Mother	–	5	6	0.033	-0.033
				0.057	-0.057
Mother	–	8	6	0.030	0.030
				0.053	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	-0.001
				0.002	-0.002
Other Caregiver	–	2	3	0.016	0.016
				0.090	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 0.014	-0.023 -0.014
Teacher	–	8	7	0.025 0.015	0.025 0.015
Teacher	–	9	8	0.003 0.002	0.003 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 0.063	-0.052 -0.063
Father	Mother	6	–	0.011 0.003	-0.005 -0.001
Afterschool Caregiver	Mother	6	–	0.092 0.033	-0.069 0.033
Teacher	Mother	6	–	0.022 0.010	-0.022 -0.010
Self-Report	Mother	15	–	0.011 0.024	0.010 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

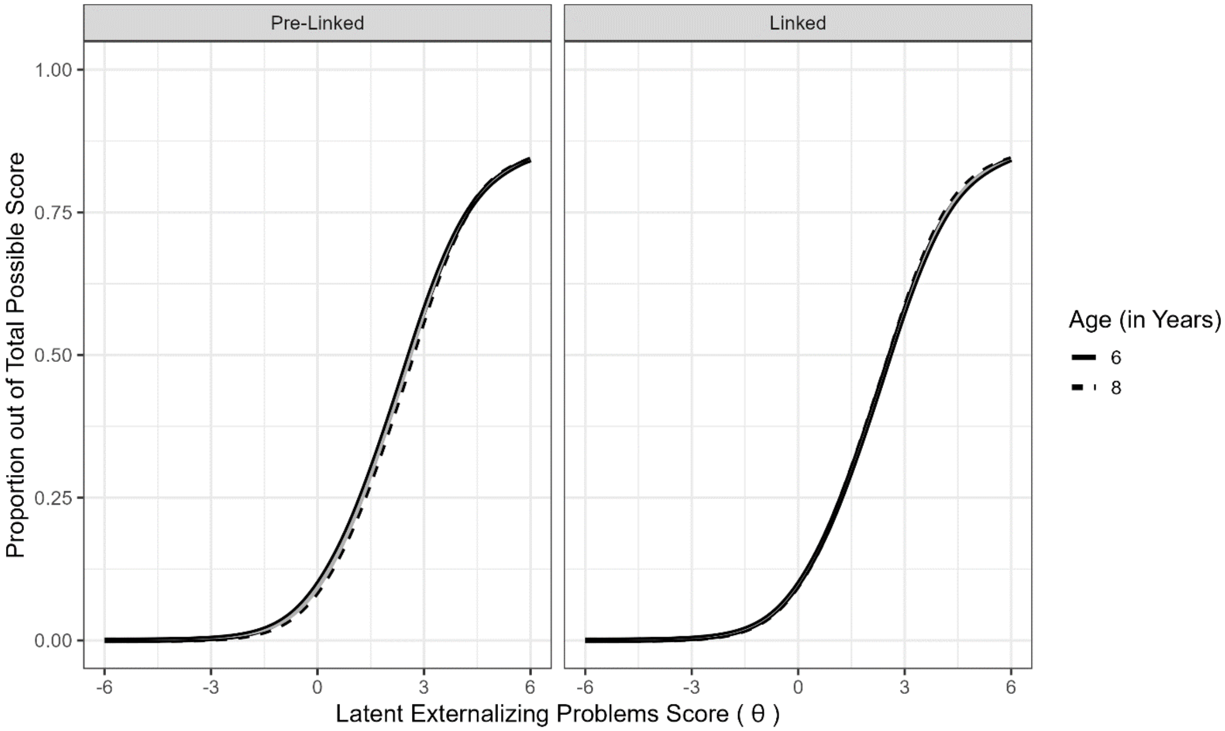
	General Factor				**Externalizing**				**Internalizing**			
Parameter	**B**	**β**	**SE**	***p***	**B**	**β**	**SE**	***p***	**B**	**β**	**SE**	***p***
Negative Affect	0.02	0.09	0.01	.021	0.02	0.08	0.01	.048	0.01	0.01	0.01	.857
Delay of Gratification	0.00	-0.02	0.00	.557	0.00	-0.03	0.00	.294	0.00	0.04	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.

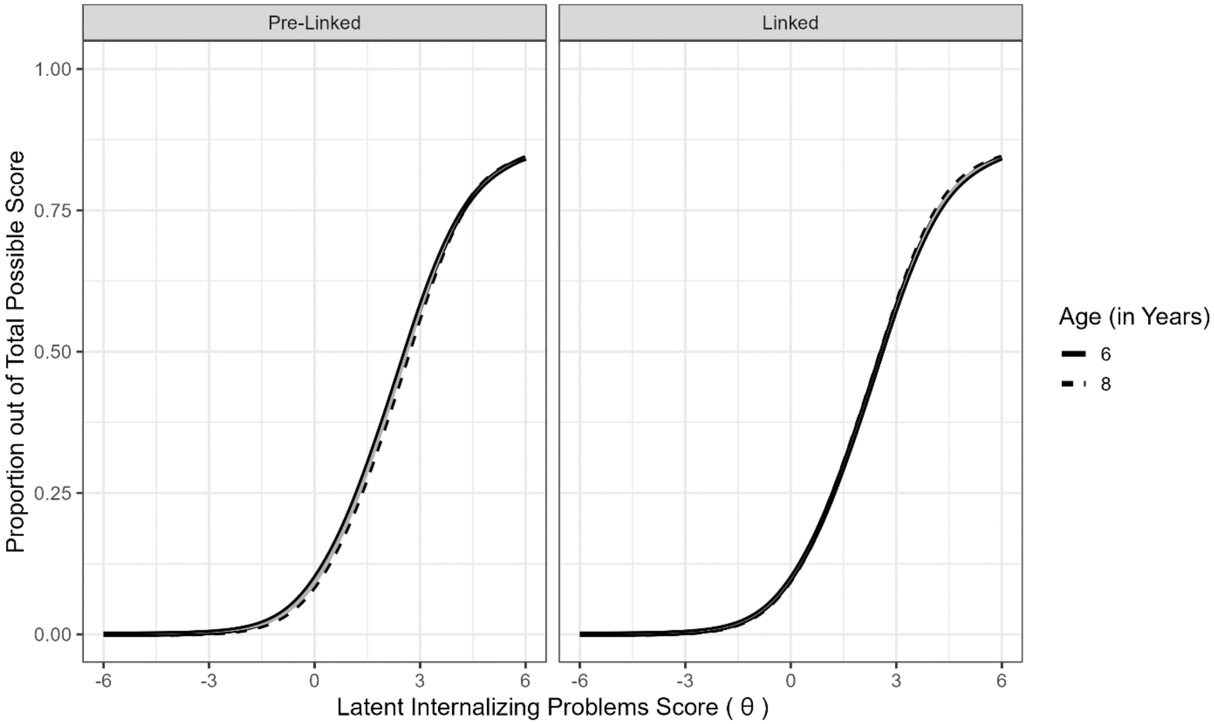
Supplementary Figure S1

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



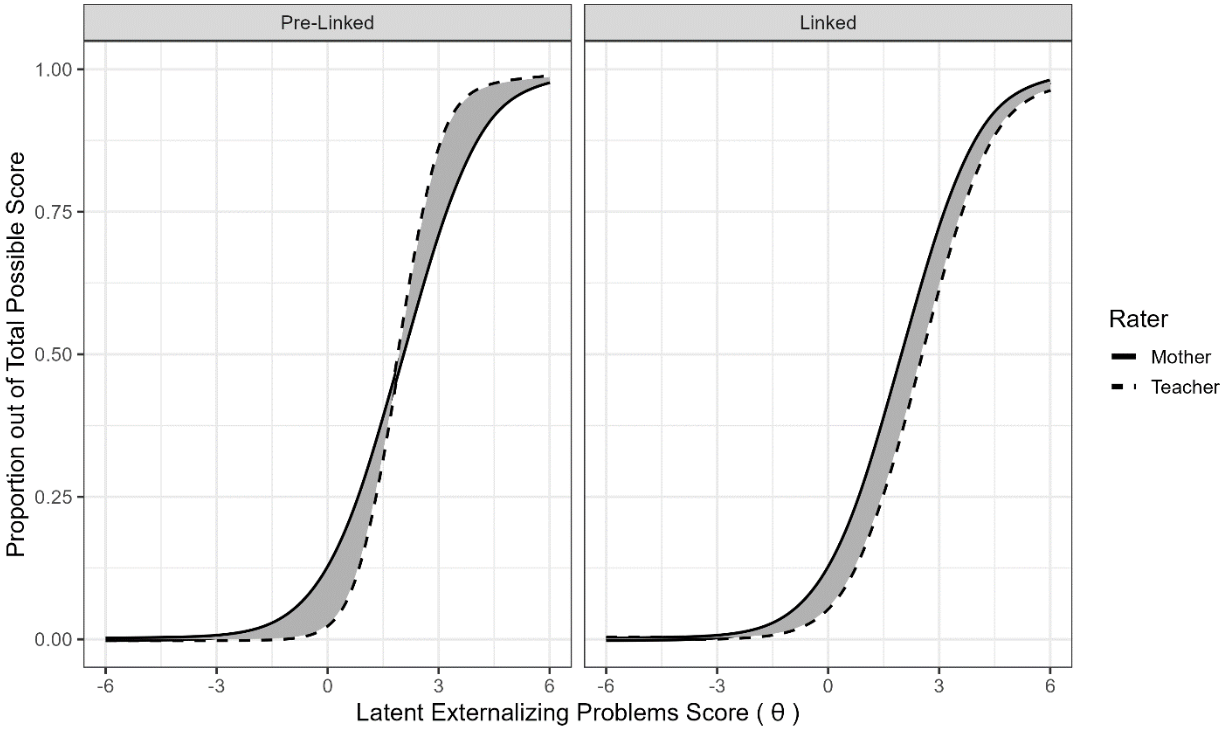
Supplementary Figure S2

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



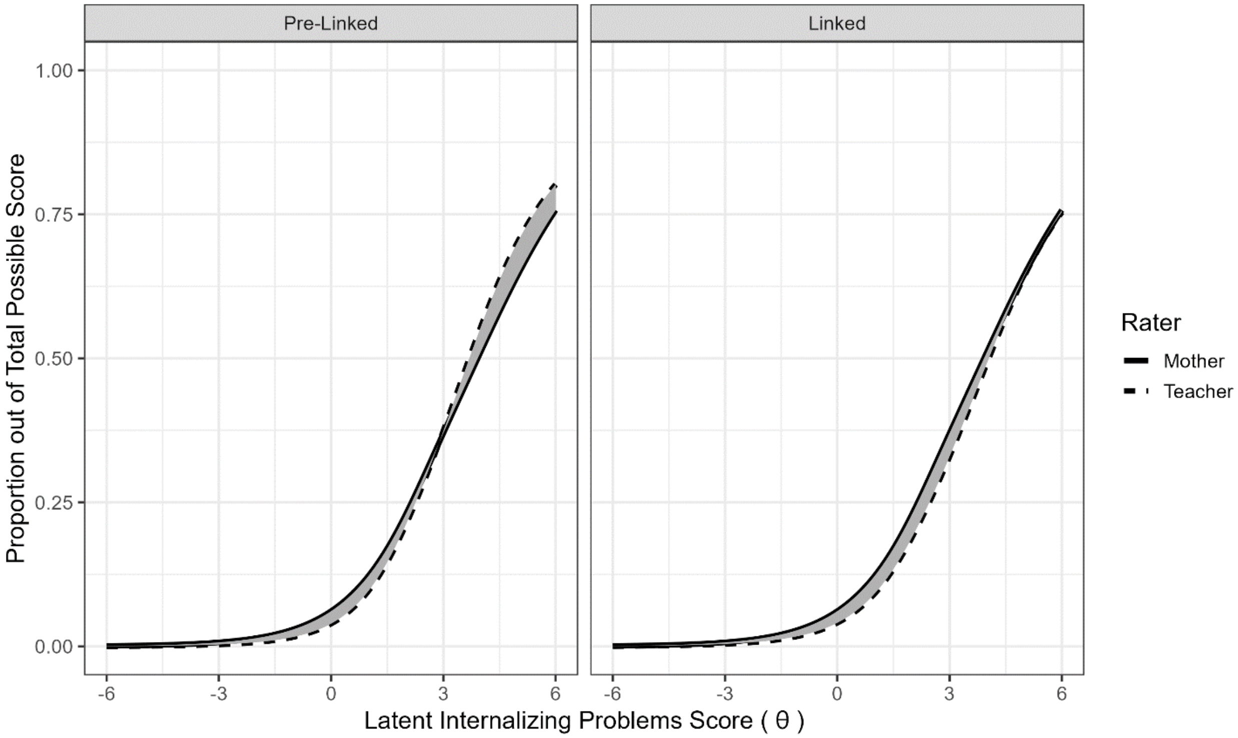
Supplementary Figure S3

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



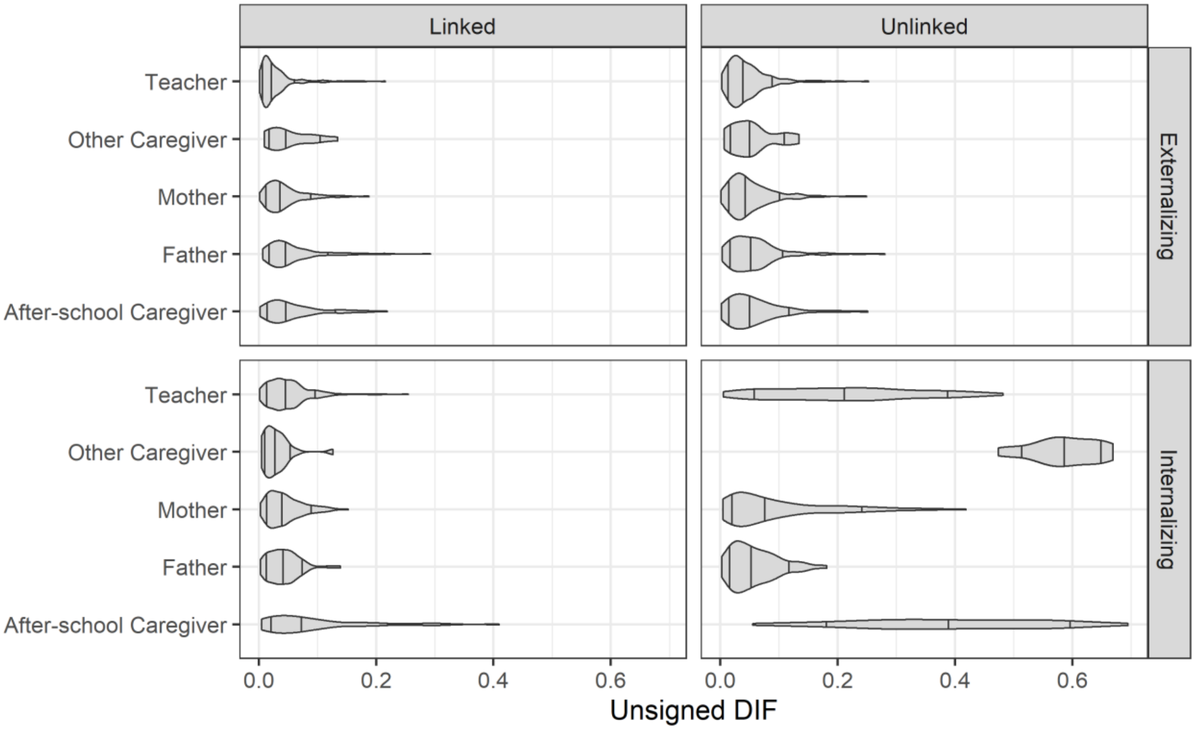
Supplementary Figure S4

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


Linked
Unlinked
Externalizing
Internalizing
Teacher
Other Caregiver
Mother
Father
After-school Caregiver
Teacher
Other Caregiver
Mother
Father
After-school Caregiver
Unsigned DIF