Supplementary Appendix S1. One-year stability coefficients of externalizing problem ratings within rater.

The mean one-year stability coefficient of externalizing problem ratings within rater was $r = .72$ for mothers' ratings (range: .56 to .80), $r = .76$ for fathers' ratings (range: .75 to .78), $r = .62$ for teachers' ratings (range: .53 to .68), $r = .60$ for afterschool caregivers' ratings (range: .57 to .63), and $r = .37$ for other caregivers' ratings (range: .35 to .39).

Supplementary Appendix S2. Details about the assessment of blood pressure, cortisol, and physical activity.

Blood Pressure

We included a measure of blood pressure as a potential physiological indicator of stress or arousal. Blood pressure has strong input from the sympathetic nervous system, which has shown hypoactivity in externalizing problems (for review, see Hastings et al., 2011). The participant's blood pressure was assessed during a lab visit at age 15 by certified personnel (for more information, see Sabol & Hoyt, 2017). Participants rested at least two minutes prior to getting their blood pressure taken. Blood pressure was taken from the nondominant arm via a blood pressure cuff while participants were seated. Five blood pressure readings were taken at 1-minute intervals. The last three available readings were used to calculate average blood pressure, consistent with prior work (Sabol & Hoyt, 2017). If fewer than three readings were taken, blood pressure was coded as missing. In the present study, blood pressure was operationalized as mean arterial pressure (mm Hg), which is the average blood pressure during a single heartbeat. Mean arterial pressure is an aggregate of systolic and diastolic blood pressure and is time-weighted to account for the fact that systole occupies ~⅓ and diastole occupies ~⅔ of a cardiac cycle. Mean arterial pressure is thus calculated as: $\frac{1}{3} \times \text{systolic blood pressure} + \frac{2}{3} \times \text{diastolic blood pressure}$ (Sesso et al., 2000). Mean arterial pressure is a strong predictor of cardiovascular disease (Sesso et al., 2000), and has been shown to be related to externalizing problems (Hastings et al., 2011).

Cortisol

We included a measure of cortisol as another physiological indicator of stress or arousal. Cortisol levels have been shown to be inversely related to externalizing problems, which may

reflect that externalizing problems are characterized by physiological hypoarousal and fearlessness (Shirtcliff et al., 2005). At age 15, the participant collected their saliva samples at home upon morning awakening for three consecutive school days using a salivette, which were used for later cortisol assay (for more information, see Roisman et al., 2009). Saliva samples were assayed using a highly sensitive enzyme immunoassay (Cat. No.1-0102/1-0112; Salimetrics, http://www.salimetrics.com). The accuracy metrics of the assay are reported in Roisman et al. (2009). Consistent with Roisman et al. (2009), cortisol values (mcg/dL) were averaged across up to three days of data collection. The mean number of days averaged into cortisol values was 2.92 (*SD* = 0.36).

Physical Activity

We included a measure of physical activity as assessed by accelerometer at age 15, given meta-analytics findings that interventions targeting increased physical activity in adolescence result in reductions in externalizing problems (Spruit et al., 2016). Spruit and colleagues hypothesized a number of various mechanisms for reasons why greater physical activity may lead to fewer behavior problems, including physiological effects, learning important social and moral skills through physical activities (e.g., sports), improved self-concept, and greater social inclusion. Participants wore a single-channel accelerometer (Computer Science and Applications, Inc.) for seven consecutive days during a typical school week (for more information, see Nader et al., 2008). Participants were asked to wear the accelerometer on a belt around the waist during waking hours for seven days, including two weekend days and five weekdays, excluding showering, bathing, water sports, or contact sports. On average, participants wore the accelerometer for 6.21 days (*SD* = 0.97), and 841.24 minutes per day (*SD* = 86.99). The accelerometer provided a continuous recording of minute-by-minute movement

counts. We operationalized physical activity as the participant's average percent of time per day spent in moderate to vigorous activity, based on metabolic equivalent tasks (moderate: ≥ 3; vigorous: ≥ 6; Nader et al., 2008).

Supplementary Appendix S3. Tests of uni-dimensionality of externalizing problem items.

One of the assumptions of the item response theory (IRT) models we used is that the externalizing problem items are uni-dimensional—that is, the items have one predominant dimension reflecting the underlying (latent) trait (i.e., externalizing problems). We tested the uni-dimensionality assumption by exploring (a) the ratio of the first to the second eigenvalue and (b) the proportion of variance the first eigenvalue accounts for by age and rater. The first eigenvalue for each age and rater combination ranged from 9.1 to 22.5 and the second eigenvalue ranged from 1.7 to 3.3. One criterion that has been suggested for uni-dimensionality is a ratio of first to second eigenvalues of ≥ 3.0 for an unrotated factor solution (Morizot et al., 2007). The ratio of the first to second eigenvalue ranged from 3.8 up to 12.3, with 30 of the 31 ratio statistics calculated being above 4. The eigenvalues suggested that the first factor accounted for considerably more variance than additional factors, consistent with uni-dimensionality. It has also been suggested that the first factor should account for at least 20% of the variance to meet the assumption of uni-dimensionality (Reckase, 1979). The proportion of variance the first eigenvalue accounted for ranged from .35 to .66, with the majority being between .37 and .5 (20 out of 31).

In sum, the assumption of uni-dimensionality was generally met. Although all of the second eigenvalues were above 1, a rule that is sometimes used for determining how many latent factors underlie the data structure, the ratio of the first to second eigenvalue and proportion of variance accounted for by the first eigenvalue provided evidence that the externalizing problem items were "uni-dimensional" enough for uni-dimensional IRT. We felt that the added complexity of modeling a second latent factor that adds between 5 and 8% of additional variance was not warranted because we were interested in the overall construct of externalizing problems.

Prior research has also suggested that IRT parameter estimates are robust to violations of uni-dimensionality (Harrison, 1986). Given evidence supporting that we approximately met the uni-dimensionality assumption of IRT, we proceeded with the IRT approach to vertical scaling.

Supplementary Appendix S4. Tests of differential item functioning of externalizing problem items.

Method

After fitting 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 linking procedure we used to link scores across measures, informants, and years minimizes scale-level latent construct differences rather than item-level differences (that would be minimized by the Haebara 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. Instead, we expected that the scale-level DIF would show improved performance on the DIF statistics 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 problems scale. The multiple differences are then averaged across the latent externalizing problems scale (for details, see Meade, 2010). 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 construct scores but has lower expected scores at higher latent construct 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.

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 construct-level (of the common items) instead of at the item-level. The expected scores at the latent construct-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. For example, 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 S2).

There is not strong guidance for interpreting effect sizes of DIF. We selected effect size cutoffs that would help us 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 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 711 common items from creating the developmental scales within a rater, 1 item showed evidence of DIF in terms of discrimination and 111 items 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 8% to 23% across raters, although the majority of these items showed only minor levels of DIF. Rates of moderate DIF ranged from 0% to 8% across raters. Teachers' ratings showing the highest rates of moderate DIF after linking, with about 8% of the 261 common items showing evidence of moderate DIF. Fathers' ratings showed the most evidence of minor DIF with about 16% of the 141 common items showing evidence of minor DIF and there was no evidence of any items continuously showing DIF across all ages. There were only two items that showed DIF across three pairs of ages: one item within the father developmental scale and another item in the teacher developmental scale. For these items, there was no evidence of systematic item-level DIF in the same direction. The severity shift was positive or negative with no apparent pattern. Supplementary Figure S1 shows the distribution of unsigned effect size statistics by rater both before and after linking. The figure illustrates that the majority of the 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 also had signed effect size statistics less than 0.05 in absolute value. Before linking, one item for mother, father, and teacher showed

evidence of non-uniform DIF across ages. After linking, only the father-rated item remained as showing evidence of non-uniform DIF across ages.

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. Of all four raters where a developmental scale was created and a total of 26 linkages examined, there was only one adjacent age linking that showed evidence of scale-level DIF after linking. This instance of DIF occurred for the teachers' ratings between ages 4 and 5, which reflected a change from the other caregivers' ratings on the Caregiver–Teacher Report Form (C–TRF) at age 4 to the teachers' ratings on the Teacher's Report Form (TRF) at age 5. This instance of DIF is classified as a DIF between ages within-rater because other caregivers and teachers were classified as the same rater role for purposes of linking (see Method section of the manuscript for more details).

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 13% to 83% across rater comparisons prior to linking and this percentage ranged from 10% to 58% across rater comparisons after linking. Even though some items showed some level of DIF, a majority of these were minor DIF with only six out of 108 items evaluated showing moderate DIF: three items differed between mothers' and teachers' ratings, and three items differed between mothers' and self-report. Of the items that showed DIF, only six of 108 items showed non-uniform DIF prior to linking, and no items showed non-uniform DIF after linking. Therefore, although there was evidence of item-level DIF, the linking improved the magnitude of DIF and also removed all non-uniform DIF.

Scale-level DIF. We also examined potential scale-level DIF between raters. There was

evidence of minor DIF for one of the scales prior to linking between mothers' and teachers' ratings at age 6; however, after linking there was no evidence of scale-level DIF and all DIF effect sizes were less than 0.01.

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 other caregivers' ratings on the C–TRF at age 4 to teachers' ratings on the TRF at age 5. In particular, items rated by other caregivers showed lower severity than items rated by teachers, which suggests that other caregivers endorsed higher rates of externalizing problems compared to teachers. The differences in severity between ratings by other caregivers' and teachers is not particularly surprising because it coincided with multiple simultaneous changes: (1) the age of the child (age 4 versus age 5) and the likely decreases with externalizing problems from ages 4 to 5, (2) the rater role (other caregiver versus teacher), (3) the likely context in which the child's behavior was observed (e.g., home/daycare/preschool versus school), and (4) the measure (C–TRF versus TRF). Thus, we exercise caution in interpreting the linking between other caregivers' ratings on the C–TRF at age 4 and teachers' ratings on the TRF at age 5. However, no other instances of DIF were observed across raters. 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.

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 measure 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 one item 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 (Knight & Zerr, 2010). Nevertheless, we examined the effect size of DIF and it was modest. Our vertical 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 S5. Details of vertical scaling (linking scores across informants, measures, and ages).

We fit a separate IRT model for each rater at each age, resulting in 31 IRT models (see Table 1 for the 31 rater-by-age instances). For example, we fit a separate IRT model for mothers' ratings at age 5 and mothers' ratings at age 6. Each IRT model estimates latent factor scores that represented a child's level of externalizing problems. We then linked externalizing problem scores across informants, measures, and ages to be on the same scale. See Figure 1 for a visualization of the measure to which each other measure was linked.

We used IRT to link the scores across informants, measures, and ages based on their common items. 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). The 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' scales so 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 2–4). We describe examples below.

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 years or informants from the same measure, all items were common items[1]. For example, we linked mothers' ratings at age 5

[1] However, any items that had a different number of response options endorsed across ages or rater roles were dropped from the linking. For example, if all mothers used only response options 0 or 1 for a given item at age 5, but the mothers used the 0, 1, and 2 response options for the same item at age 6, this item was not used in the linking.

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 Supplementary Table S2.

Our IRT approach to vertical scaling applied three steps to link scores from different measures to be on the same scale. First, we fit separate IRT models for each rater at each age (described above). Second, 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 proportion out of the total possible score) as a function of a child's latent level of externalizing problems. Third, 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 S7. We describe an example below.

See Figure 4 for an example of test characteristic curves of the common items of mother- and teacher-rated externalizing 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 minimizes 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 were: *A* (slope linking constant) = 1.74, and *B* (intercept linking constant) = -1.44 (see Supplementary Table S7). The left panel of the figure indicates that, prior to linking, mothers' ratings showed somewhat lower discrimination than teachers' ratings at age 6. 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 2–4).

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 mother' 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). We first linked scores within-rater (see Equation 5), and

then linked scores across raters to link scores to mothers' ratings (see Equation 6). 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 externalizing problems for a given rater and age in subsequent growth curve models.

Supplementary Appendix S6. Growth curve model formulas.

$$Y_{ij} = \beta_0 + b_{00i} + \epsilon_{ij}$$
$$Y_{ij} = \beta_0 + \beta_1(\text{age}_{ij} - 15) + \beta_2(\text{age}_{ij} - 15)^2 + b_{00i} + b_{10i}(\text{age}_{ij} - 15) + \epsilon_{ij}$$
$$Y_{ij} = \beta_0 + \beta_1(\text{age}_{ij} - 15) + \beta_2(\text{age}_{ij} - 15)^2 + \beta_3\text{rater}_{ij} + b_{00i} + b_{10i}(\text{age}_{ij} - 15) + \epsilon_{ij}$$
$$Y_{ij} = \beta_0 + \beta_1(\text{age}_{ij} - 15) + \beta_2(\text{age}_{ij} - 15)^2 + \beta_3\text{rater}_{ij} + b_{00i} + b_{10i}(\text{age}_{ij} - 15) + b_{20i}(\text{age}_{ij} - 15)^2 + \epsilon_{ij}$$
$$Y_{ij} = \beta_0 + \beta_1(\text{age}_{ij} - 15) + \beta_2(\text{age}_{ij} - 15)^2 + \beta_3\text{rater}_{ij} + \beta_4(\text{age}_{ij} - 15) \times \text{rater}_{ij} + b_{00i} + b_{10i}(\text{age}_{ij} - 15) + \epsilon_{ij}$$
$$Y_{ij} = \beta_0 + \beta_1(\text{age}_{ij} - 15) + \beta_2(\text{age}_{ij} - 15)^2 + \beta_3\text{rater}_{ij} + \beta_4(\text{age}_{ij} - 15) \times \text{rater}_{ij} + \beta_k\text{Demographics}_{ik} + b_{00i} + b_{10i}(\text{age}_{ij} - 15) + \epsilon_{ij}$$
$$Y_{ij} = \beta_0 + \beta_1(\text{age}_{ij} - 15) + \beta_2(\text{age}_{ij} - 15)^2 + \beta_3\text{rater}_{ij} + \beta_4(\text{age}_{ij} - 15) \times \text{rater}_{ij} + \beta_k\text{Demographics}_{ik} + \beta_5 EL_i + \beta_6 EL_i \times (\text{age}_{ij} - 15) + b_{00i} + b_{10i}(\text{age}_{ij} - 15) + \epsilon_{ij}$$
$$Y_{ij} = \beta_0 + \beta_1(\text{age}_{ij} - 15) + \beta_2(\text{age}_{ij} - 15)^2 + \beta_3\text{rater}_{ij} + \beta_4(\text{age}_{ij} - 15) \times \text{rater}_{ij} + \beta_k\text{Demographics}_{ik} + \beta_5 VC_i + \beta_6 VC_i \times (\text{age}_{ij} - 15) + b_{00i} + b_{10i}(\text{age}_{ij} - 15) + \epsilon_{ij}$$
$$Y_{ij} = \beta_0 + \beta_1(\text{age}_{ij} - 15) + \beta_2(\text{age}_{ij} - 15)^2 + \beta_3\text{rater}_{ij} + \beta_4(\text{age}_{ij} - 15) \times \text{rater}_{ij} + \beta_k\text{Demographics}_{ik} + \beta_5 EL_i + \beta_6 EL_i \times (\text{age}_{ij} - 15) + \beta_k\text{Biological}_{ik} + b_{00i} + b_{10i}(\text{age}_{ij} - 15) + \epsilon_{ij}$$
$$Y_{ij} = \beta_0 + \beta_1(\text{age}_{ij} - 15) + \beta_2(\text{age}_{ij} - 15)^2 + \beta_3\text{rater}_{ij} + \beta_4(\text{age}_{ij} - 15) \times \text{rater}_{ij} + \beta_k\text{Demographics}_{ik} + \beta_5 VC_i + \beta_6 VC_i \times (\text{age}_{ij} - 15) + \beta_k\text{Biological}_{ik} + b_{00i} + b_{10i}(\text{age}_{ij} - 15) + \epsilon_{ij}$$

Note: Y_{ij} is the externalizing 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 externalizing 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. Biological_{ik} represents a set of *k* bio-behavioral covariates used to examine

differences as a function of cortisol, blood pressure, and physical activity. The focal predictors of interest were β_5 and β_6 representing the association of expressive language and verbal comprehension with intercepts and slopes, respectively, of externalizing problems.

The data structure for a single rater would represent repeated measures nested within the participant. Because we included ratings from multiple informants of a given child, there are possibly multiple ratings for a given participant at a single time point. As such, the effect of rater role is considered cross classified rather than nested. That is, each rater does not provide a rating for each participant at every time point; rather, each rater provided a rating for a given participant at some time points based on the SECCYD data collection design. This more complicated cross-classified data structure was modeled by treating the data as repeated measures nested within the participant, by treating the effect of rater role as a fixed (rather than random) effect.

Treating rater role as a fixed effect has a few potential issues with the mixed models used. First, it is has been consistently shown that misspecifying the random effect structure does not lead to bias in the estimates of the fixed parameters which are of most interest in this study (Kwok et al., 2007; LeBeau, 2016; Murphy & Pituch, 2009). Thus, prior evidence provides support for the modeling approach we used in the current study. By contrast, misspecifying the random effect structure could lead to standard errors that are biased (Kwok et al., 2007; LeBeau, 2016; Murphy & Pituch, 2009). However, we corrected for the potential random effect misspecification by adding the rater role as fixed parameters, which should remove the variance associated with raters from the random effects, thus providing a correction factor for the standard errors. Treating raters as fixed parameters also impacts the types of inferences that can be made and who the inferences can be generalized to. With the rater role as a fixed effect, we made the assumption that these rater roles, i.e., mothers, fathers, teachers, afterschool caregivers, other caregivers, and self-report, would be the

most likely to provide ratings for externalizing problems in practice. The extent to which other rater roles are assessed, these study results may not generalize to those raters. We were also interested in exploring the extent to which the rater roles yielded different trajectories of participants' externalizing problems, which was more directly testable by treating the rater role as fixed instead of random.

Our modeling approach was also supported empirically. For example, the R^2 for the fixed effects was about 10% when modeling only the linear and quadratic trajectories with no other effects added to the model. The rater role fixed effect was added next which increased the R^2 for the fixed effects to 20% (an additional 10% of variance explained). The percent of variance explained by rater role was as large as the percent of variance explained by the trajectory terms. Furthermore, this explained variance by the rater role fixed effects resulted in a reduction in the residual variance component associated with the level 1 or repeated measurements in the model, from 0.727 to 0.595. Finally, as a sensitivity check, we fit the cross-classified model and the results were very similar in terms of R^2 explained, except instead of the explained variance being attributed to the fixed effects, it was included as part of the random component. For example, the cross-classified model that estimated a random effect for each rater role and included the linear and quadratic trajectory terms had nearly identical R^2 for random effects of 53% compared to the R^2 of 52% for the model that treated rater role as a fixed effect. These results suggest that we successfully adjusted for the effect of rater role with our approach, and suggest that not modeling this term would have resulted in the potential for significant bias in the standard errors and inferences made from the mixed model.

Supplementary Appendix S7. Tests of systematic missingness and how missing data were handled.

Tests of Systematic Missingness

We observed some systematic missingness of externalizing problem scores as a function of demographic and socioeconomic factors. The number of time points that a child had ratings of externalizing 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). Whites had more time points of ratings on average compared to African Americans (t[214.89] = 3.28, p = .001) but not compared to Hispanics (t[92.03] = 0.63, p = .532). 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 models.

How We Handled Missing Data

We modeled externalizing problem trajectories using a linear mixed model (LMM). Longitudinal LMM analyzes 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, LMM still uses the child's data for the first two measurement occasions. LMMs assume that the data are missing at random or completely at random. As a sensitivity test, we also examined findings after multiple imputation to account for missing data across ages and raters (as described below). Findings with multiple imputation were substantially similar, so we present results from the raw data.

Multiple Imputation

As a sensitivity test, we also examined findings after multiple imputation to account for missing data across ages and raters. To account for missingness across ages and raters, we expanded the data matrix to have rows for all possible raters at the ages those raters were intended to be assessed (i.e., mothers: ages 2, 3, 4, 5, 6, 8, 9, 10, 11, 15; fathers: ages 6, 8, 9, 10, 11, 15; teachers: ages 5, 6, 7, 8, 9, 10, 11; after-school caregivers: ages 6, 8, 9, 10; other caregivers: ages 2, 3, 4; self-report: age 15). We did not impute scores for raters at ages those raters were not intended to be assessed (e.g., self-report at age 2) because those columns would have had no observed data, which would have resulted in an overly sparse matrix for imputation.

We multiply imputed 100 data sets with the model variables using the mice package (van Buuren & Groothuis-Oudshoorn, 2011) in R. To account for longitudinal data in imputation, we used the 2l.pan function for imputing missing data at level 1 (i.e., time-varying externalizing problems), according to a mixed model, as described by van Buuren (2018). We included a quadratic term for age in the imputation model to allow externalizing problems to show non-linear change over time. We allowed the linear and quadratic terms for age to have random effects in the imputation of externalizing problems, to allow children to have different slopes. We included the time-invariant predictors as fixed effects. We used the 2lonly.pmm function to impute missing data at level 2 (i.e., time-invariant variables), which uses predictive mean matching (van Buuren, 2018). This multilevel imputation approach has proven successful with longitudinal data (Huque et al., 2018; Lüdtke et al., 2017; Vink et al., 2015).

Supplementary Appendix S8. Nested growth curve model comparisons.

We conducted several nested growth curve model comparisons to identify the best fitting form of change. First, we fit an unconditional means model (allowing each child to have different means) and an unconditional growth model (allowing each child to have different intercepts and slopes). Results from the unconditional means model are in Supplementary Table S8. Results from the unconditional growth model are in Supplementary Table S9. The unconditional growth model (AICc = 67,481.61) fit significantly better than the unconditional means model (AIC = 71,447.35; $\chi^2[4] = 3{,}973.70$, $p < .001$), indicating that children differed in their slopes.

We fit four models to develop the initial baseline trajectory prior to adding other predictors. Given the considerable trajectory differences as a function of rater role (a model that adjusted for rater role fit significantly better than the unconditional growth model; $\chi^2[3] = 2{,}623.90$, $p < .001$), we adjusted for rater role in each model. First, we fit a linear model trajectory with random intercepts and slopes (AICc = 64,863.68). Second, we fit a linear model that allowed for different linear trajectories for each rater role (AICc = 64,704.38), which fit significantly better than the previous model ($\chi^2[3] = 165.30$, $p < .001$). Third, we added a fixed quadratic term to the model (AICc = 63,758.19), which fit significantly better than the previous model ($\chi^2[1] = 948.20$, $p < .001$). Finally, we allowed the quadratic trajectories to vary based on rater role (AICc = 62,894.04), which fit significantly better than the previous model ($\chi^2[3] = 870.16$, $p < .001$). We also considered a model that included a random quadratic effect, but this model was not able to converge due to insufficient variance in the quadratic term. The model with random linear and fixed quadratic slopes that varied by rater role showed the best fit and had the smallest AICc, so it was used as the baseline model with which subsequent models were

compared. Results from the baseline growth model that accounts for the effects of rater role are in Supplementary Table S10.

References

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational and Behavioral Statistics*, *11*(2), 91–115. https://doi.org/10.3102/10769986011002091

Hastings, P. D., Shirtcliff, E. A., Klimes-Dougan, B., Allison, A. L., Derose, L., Kendziora, K. T., Usher, B. A., & Zahn-Waxler, C. (2011). Allostasis and the development of internalizing and externalizing problems: Changing relations with physiological systems across adolescence. *Development and Psychopathology*, *23*(4), 1149–1165. https://doi.org/10.1017/S0954579411000538

Huque, M. H., Carlin, J. B., Simpson, J. A., & Lee, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology*, *18*(1), 168. https://doi.org/10.1186/s12874-018-0615-6

Knight, G. P., & Zerr, A. A. (2010). Informed theory and measurement equivalence in child development research. *Child Development Perspectives*, *4*(1), 25–30. https://doi.org/10.1111/j.1750-8606.2009.00112.x

Kwok, O.-M., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A monte carlo study. *Multivariate Behavioral Research*, *42*(3), 557–592. https://doi.org/10.1080/00273170701540537

LeBeau, B. (2016). Impact of serial correlation misspecification with the linear mixed model. *Journal of Modern Applied Statistical Methods*, *15*(1), 389–416. https://doi.org/10.22237/jmasm/1462076400

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel

designs: A comparison of different strategies. *Psychological Methods*, *22*(1), 141–165. https://doi.org/10.1037/met0000096

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*(4), 728–743. https://doi.org/10.1037/a0018966

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407–421). Guilford Press. https://psycnet.apa.org/record/2007-11524-024

Murphy, D. L., & Pituch, K. A. (2009). The performance of multilevel growth curve models under an autoregressive moving average process. *The Journal of Experimental Education*, *77*(3), 255–284. https://doi.org/10.3200/JEXE.77.3.255-284

Nader, P. R., Bradley, R. H., Houts, R. M., McRitchie, S. L., & O'Brien, M. (2008). Moderate-to-vigorous physical activity from ages 9 to 15 years. *JAMA*, *300*(3), 295–305. https://doi.org/10.1001/jama.300.3.295

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. https://doi.org/10.1007/BF02294403

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207–230. https://doi.org/10.2307/1164671

Roisman, G. I., Susman, E., Barnett-Walker, K., Booth-LaForce, C., Owen, M. T., Belsky, J., Bradley, R. H., Houts, R., Steinberg, L., & The NICHD Early Child Care Research Network. (2009). Early family and child-care antecedents of awakening cortisol levels in

adolescence. *Child Development*, *80*(3), 907–920. https://doi.org/10.1111/j.1467-8624.2009.01305.x

Sabol, T. J., & Hoyt, L. T. (2017). The long arm of childhood: Preschool associations with adolescent health. *Developmental Psychology*, *53*(4), 752–763. https://doi.org/10.1037/dev0000287

Sesso, H. D., Stampfer, M. J., Rosner, B., Hennekens, C. H., Gaziano, J. M., Manson, J. E., & Glynn, R. J. (2000). Systolic and diastolic blood pressure, pulse pressure, and mean arterial pressure as predictors of cardiovascular disease risk in men. *Hypertension*, *36*(5), 801–807. https://doi.org/10.1161/01.HYP.36.5.801

Shirtcliff, E. A., Granger, D. A., Booth, A., & Johnson, D. (2005). Low salivary cortisol levels and externalizing behavior problems in youth. *Development and Psychopathology*, *17*(1), 167–184. https://doi.org/10.1017/S0954579405050091

Spruit, A., Assink, M., van Vugt, E., van der Put, C., & Stams, G. J. (2016). The effects of physical activity interventions on psychosocial outcomes in adolescents: A meta-analytic review. *Clinical Psychology Review*, *45*, 56–71. https://doi.org/10.1016/j.cpr.2016.03.006

van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC. https://stefvanbuuren.name/fimd/

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Vink, G., Lazendic, G., & van Buuren, S. (2015). Partioned predictive mean matching as a large data multilevel imputation technique. *Psychological Test and Assessment Modeling*, *57*(4), 577–594. https://dspace.library.uu.nl/handle/1874/325909

Supplementary Table S1. Correlation matrix of model variables.

Variable	Age	Sex	Income-to-Needs Ratio	African American	Hispanic	Externalizing Problems	Mean Arterial Pressure	Cortisol	Physical Activity	Vocabulary	Expressive Language
Age	–										
Sex	n/a	–									
Income-to-Needs Ratio	n/a	.01	–								
African American	n/a	.00	-.22***	–							
Hispanic	n/a	.00	-.06*	-.07*	–						
Externalizing Problems	-.20***	-.12***	-.11***	.10***	.01	–					
Mean Arterial Pressure	n/a	-.23***	-.05	.05	-.03	.05***	–				
Cortisol	n/a	.14***	.11***	-.08*	.02	-.06***	-.06	–			
Physical Activity	n/a	-.33***	-.03	.11*	.03	.06***	.01	-.05	–		
Vocabulary	n/a	.20***	.31***	-.35***	-.12***	-.18***	-.08*	.04	-.08†	–	
Expressive Language	n/a	.16***	.17***	-.19***	-.09***	-.11***	-.11***	.03	-.08†	.57***	–
% Missingness	0.00	0.00	6.67	0.00	0.00	39.79	37.24	36.36	55.87	15.10	17.16
M	7.55	0.48	2.86	0.13	0.06	-0.20	83.77	0.36	5.66	97.85	96.88
SD	3.49	0.50	2.61	0.34	0.24	1.11	5.89	0.18	3.50	15.85	14.53

Note: *** $p < .001$; * $p < .05$; † $p < .10$; all *p*s two-tailed. "n/a" indicates that the association of the variable with age is not applicable because the variable is treated as time-invariant.

Supplementary Table S2. The number of common items for each pair of measures.

Measure	CBCL 2–3	CBCL 4–18	C–TRF	TRF	YSR
CBCL 2–3	26				
CBCL 4–18	9	33			
C–TRF	18	14	40		
TRF	10	27	16	34	
YSR	8	30	14	27	30

Note. "CBCL" = Child Behavior Checklist, "C–TRF" = Caregiver–Teacher Report Form, "TRF" = Teacher's Report Form, "YSR" = Youth Self-Report. Numbers on the diagonal represent the total number of items in the Externalizing scale for that measure (e.g., the CBCL 4–18 has 33 items). Numbers below the diagonal represent, for that pair of measures, the number of items that are common to both of the measures. The number of unique items can be calculated by subtracting the number of common items from the total number of items. For instance, the CBCL 4–18 has 6 unique items when compared with the TRF (i.e., 33 total items minus 27 common items). Conversely, the TRF has 7 unique items when compared with the CBCL 4–18 (i.e., 34 total items minus 27 common items).

Supplementary Table S3. Cronbach's alpha estimates of internal consistency of externalizing problem scores by age and rater.

	Age (Years)										
	2	3	4	5	6	7	8	9	10	11	15
Mother	.88	.89	.88	.89	.89	–	.89	.89	.89	.89	.91
Father	–	–	–	–	.88	–	.88	.90	.91	.91	.91
Teacher	–	–	–	.94	.93	.94	.95	.95	.95	.95	–
After-School Caregiver	–	–	–	–	.92	–	.92	.92	.91	–	–
Other Caregiver	.91	.92	.95	–	–	–	–	–	–	–	–
Self-Report	–	–	–	–	–	–	–	–	–	–	.86

Note: "–" indicates not applicable because the particular rater did not provide ratings at the given time point.

Supplementary Table S4. Descriptive statistics of externalizing problems by age and rater.

	Age (Years)										
M	2	3	4	5	6	7	8	9	10	11	15
Mother	0.77	0.66	0.69	0.11	0.00	–	-0.13	-0.24	-0.30	-0.34	-0.49
Father	–	–	–	–	0.03	–	-0.18	-0.24	-0.38	-0.33	-0.41
Teacher	–	–	–	-0.91	-0.85	-0.86	-0.78	-0.84	-0.76	-0.50	–
After-School Caregiver	–	–	–	–	-0.21	–	-0.40	-0.44	-0.58	–	–
Other Caregiver	0.55	0.41	-0.47	–	–	–	–	–	–	–	–
Self-Report	–	–	–	–	–	–	–	–	–	–	0.41

	Age (Years)										
SD	2	3	4	5	6	7	8	9	10	11	15
Mother	0.79	0.80	0.80	0.91	0.93	–	0.94	0.94	0.98	0.96	0.98
Father	–	–	–	–	0.89	–	0.89	0.95	0.98	0.97	0.96
Teacher	–	–	–	1.05	1.07	1.08	1.16	1.09	1.11	1.17	–
After-School Caregiver	–	–	–	–	1.04	–	1.07	1.03	1.07	–	–
Other Caregiver	0.98	1.05	1.13	–	–	–	–	–	–	–	–
Self-Report	–	–	–	–	–	–	–	–	–	–	0.87

Note: "–" indicates not applicable because the particular rater did not provide ratings at the given time point.

Supplementary Table S5. Percentage of participants with externalizing problem scores at different numbers of time points.

	# of Time Points											
Rater	0	1	2	3	4	5	6	7	8	9	10	11
Mother	8.1	2.2	4.8	1.7	2.2	3.9	2.2	3.8	4.2	11.4	55.6	–
Father	26.0	9.0	5.9	7.0	8.8	13.6	29.8	–	–	–	–	–
Teacher	17.2	1.8	3.0	3.7	5.6	8.6	20.2	40.1	–	–	–	–
After-School Caregiver	67.2	15.2	8.1	6.2	3.4	–	–	–	–	–	–	–
Other Caregiver	27.3	25.1	20.7	26.9	–	–	–	–	–	–	–	–
Self-Report	29.8	70.2	–	–	–	–	–	–	–	–	–	–
Total	7.6	2.1	4.8	1.7	1.7	2.8	1.8	2.5	2.0	4.4	13.1	55.6

Note: "–" indicates not applicable because the particular rater did not provide ratings at the given number of time points. Percentages in a row may not sum exactly to 100.0% because of rounding error.

Supplementary Table S6. Correlation matrix of externalizing problem scores by rater.

Rater	Mother	Father	Teacher	After-School Caregiver	Other Caregiver	Self-Report
Mother	–					
Father	.56***	–				
Teacher	.32***	.32***	–			
After-School Caregiver	.39***	.41***	.44***	–		
Other Caregiver	.20***	n/a	n/a	n/a	–	
Self-Report	.32***	.33***	n/a	n/a	n/a	–

Note: *** $p < .001$; all *p*s two-tailed. "n/a" indicates not applicable because the two raters did not provide ratings at the same time point(s).

Supplementary Table S7. 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*
After-School Caregiver	–	8	6	1.136	-0.252
After-School Caregiver	–	9	8	0.984	-0.460
After-School Caregiver	–	10	9	1.129	-0.199
Father	–	8	6	1.029	-0.263
Father	–	9	8	1.123	-0.096
Father	–	10	9	1.114	-0.201
Father	–	11	10	0.963	0.059
Father	–	15	11	1.062	-0.121
Mother	–	2	3	0.985	0.158
Mother	–	3	4	1.010	-0.056
Mother	–	4	5	0.830	0.667
Mother	–	5	6	0.938	0.136
Mother	–	8	6	1.038	-0.159
Mother	–	9	8	1.037	-0.133
Mother	–	10	9	1.084	-0.970
Mother	–	11	10	0.999	-0.050
Mother	–	15	11	1.116	-0.220
Teacher (Other Caregiver)	–	2	3	0.906	0.145
Teacher (Other Caregiver)	–	3	4	0.750	0.782
Teacher (Other Caregiver)	– (Teacher)	4	5	0.806	0.507
Teacher	–	5	6	1.050	-0.106
Teacher	–	7	6	1.022	-0.026
Teacher	–	8	7	1.029	0.076
Teacher	–	9	8	0.994	-0.078
Teacher	–	10	9	0.970	0.088
Teacher	–	11	10	1.098	0.093
Father	Mother	6	–	0.935	0.041
After-School Caregiver	Mother	6	–	1.254	-0.318
Teacher	Mother	6	–	1.741	-1.439
Self-Report	Mother	15	–	0.856	0.489

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

Supplementary Table S8. Unconditional Means Model

	B	*SE*	*df*	*p*
Intercept	-0.14	0.02	1199.42	**< .001**
R^2 (fixed effects)	.000			
R^2 (fixed and random effects)	.300			

Note: *p*-values less than .05 in bold.

Supplementary Table S9. Unconditional Growth Model.

	B	β	*SE*	*df*	*p*
Intercept	-0.03	-0.17	0.03	1488.00	**.257**
Time (Linear)	0.19	0.68	0.01	15970.00	**< .001**
Time (Quadratic)	0.02	0.95	0.00	24440.00	**< .001**
R^2 (fixed effects)	.103				
R^2 (fixed and random effects)	.411				

Note: *p*-values less than .05 in bold. “Time” (in years) was centered to set the intercepts at the last time point (age 15). For example, time is coded such that age 2 = -13 and age 15 = 0.

Supplementary Table S10. Baseline Growth Model: Accounting for Effects of Rater Role.

	B	β	*SE*	*df*	*p*
Intercept	-0.46	-0.08	0.03	3044.00	**< .001**
Time (Linear)	0.01	0.88	0.01	23190.00	**.039**
Time (Quadratic)	0.01	1.08	0.00	23710.00	**< .001**
Father	0.14	0.01	0.04	23790.00	**< .001**
Teacher	2.04	-0.16	0.08	23630.00	**< .001**
After-School Caregiver	-0.74	-0.08	0.75	23270.00	.327
Self-Report	0.89	0.17	0.03	23220.00	**< .001**
Time (Linear) × Father	0.02	0.02	0.01	23490.00	.278
Time (Linear) × Teacher	0.68	1.17	0.02	23740.00	**< .001**
Time (Linear) × After-School Caregiver	-0.19	-0.12	0.22	23260.00	.393
Time (Quadratic) × Father	0.00	0.00	0.00	23370.00	.864
Time (Quadratic) × Teacher	0.04	0.92	0.00	23870.00	**< .001**
Time (Quadratic) × After-School Caregiver	-0.02	-0.15	0.02	23250.00	.291
R^2 (fixed effects)	.199				
R^2 (fixed and random effects)	.519				

Note: *p*-values less than .05 in bold. “Time” (in years) was centered to set the intercepts at the last time point (age 15). For example, time is coded such that age 2 = -13 and age 15 = 0. Mothers served as the reference rater to which fathers, teachers, after-school caregivers, and self-report were compared. Interaction terms with time reflect predictions of the linear or quadratic slopes. For instance, “Time (Linear) × Father” reflects differences in slopes of fathers’ ratings (compared to slopes of mothers’ ratings). Self-report was not allowed to predict the slopes because it was assessed at only one time point.

Supplementary Table S11. Growth Model with Demographic and Socioeconomic Factors.

	B	β	*SE*	*df*	*p*
Intercept	-0.28	-0.10	0.05	1478.00	**< .001**
Time (Linear)	0.02	0.88	0.01	11880.00	**.039**
Time (Quadratic)	0.01	1.09	0.00	22340.00	**< .001**
Father	0.15	0.01	0.04	22380.00	**< .001**
Teacher	2.03	-0.17	0.09	22270.00	**< .001**
After-School Caregiver	-0.52	-0.08	0.78	21950.00	.506
Self-Report	0.90	0.17	0.04	21880.00	**< .001**
Time (Linear) × Father	0.02	0.02	0.01	22130.00	.227
Time (Linear) × Teacher	0.68	1.18	0.02	22380.00	**< .001**
Time (Linear) × After-School Caregiver	-0.13	-0.08	0.23	21930.00	.571
Time (Quadratic) × Father	0.00	0.00	0.00	22020.00	.973
Time (Quadratic) × Teacher	0.04	0.93	0.00	22500.00	**< .001**
Time (Quadratic) × After-School Caregiver	-0.01	-0.12	0.02	21930.00	.424
Sex	-0.24	-0.11	0.05	1012.00	**< .001**
African American	0.30	0.08	0.08	1052.00	**.000**
Hispanic	0.16	0.02	0.11	1025.00	.139
Income-to-Needs Ratio	-0.04	-0.09	0.01	1031.00	**< .001**
Time (Linear) × Sex	0.00	-0.01	0.00	970.00	.413
Time (Linear) × African American	0.00	0.00	0.01	1068.00	.509
Time (Linear) × Hispanic	0.01	0.01	0.01	1005.00	.144
Time (Linear) × Income-to-Needs Ratio	0.00	-0.01	0.00	993.50	.215
R^2 (fixed effects)	.233				
R^2 (fixed and random effects)	.517				

Note: *p*-values less than .05 in bold. “Time” (in years) was centered to set the intercepts at the last time point (age 15). For example, time is coded such that age 2 = -13 and age 15 = 0. Mothers served as the reference rater to which fathers, teachers, after-school caregivers, and self-report were compared. Interaction terms with time reflect predictions of the linear or quadratic slopes. For instance, “Time (Linear) × Father” reflects differences in slopes of fathers’ ratings (compared to slopes of mothers’ ratings). Self-report was not allowed to predict the slopes because it was assessed at only one time point. Sex was coded such that male = 0 and female = 1.

In terms of ethnicity, Whites served as the reference group to which Blacks and Hispanics were compared.

Supplementary Table S12. Growth Model with Language Ability.

	Verbal Comprehension as Predictor					Expressive Language as Predictor				
	B	β	*SE*	*df*	*p*	*B*	β	*SE*	*df*	*p*
Intercept	0.37	-0.11	0.18	990.70	**.043**	-0.24	-0.12	0.19	977.80	.199
Time (Linear)	0.00	0.89	0.02	1476.00	.919	-0.03	0.88	0.02	1423.00	.116
Time (Quadratic)	0.01	1.09	0.00	21310.00	**< .001**	0.01	1.09	0.00	20820.00	**< .001**
Father	0.13	0.00	0.04	21400.00	**.001**	0.13	0.00	0.04	20900.00	**.001**
Teacher	2.03	-0.17	0.09	21290.00	**< .001**	2.05	-0.17	0.09	20790.00	**< .001**
After-School Caregiver	-0.56	-0.08	0.80	21010.00	.481	-0.89	-0.09	0.81	20520.00	.270
Self-Report	0.88	0.17	0.04	20940.00	**< .001**	0.89	0.17	0.04	20460.00	**< .001**
Time (Linear) × Father	0.02	0.02	0.01	21170.00	.306	0.02	0.02	0.02	20670.00	.241
Time (Linear) × Teacher	0.68	1.18	0.02	21380.00	**< .001**	0.69	1.19	0.02	20870.00	**< .001**
Time (Linear) × After-School Caregiver	-0.14	-0.09	0.23	20990.00	.537	-0.23	-0.15	0.23	20510.00	.323
Time (Quadratic) × Father	0.00	0.00	0.00	21070.00	.884	0.00	0.00	0.00	20590.00	.989
Time (Quadratic) × Teacher	0.04	0.94	0.00	21490.00	**< .001**	0.04	0.94	0.00	20980.00	**< .001**
Time (Quadratic) × After-School Caregiver	-0.01	-0.13	0.02	20990.00	.393	-0.02	-0.18	0.02	20500.00	.233
Sex	-0.19	-0.08	0.05	955.00	**< .001**	-0.22	-0.10	0.05	938.50	**< .001**
African American	0.22	0.04	0.09	985.00	**.016**	0.34	0.08	0.09	969.00	**< .001**
Hispanic	0.13	0.00	0.11	975.50	.238	0.19	0.02	0.11	956.10	.083
Income-to-Needs Ratio	-0.04	-0.07	0.01	972.90	**< .001**	-0.05	-0.09	0.01	955.80	**< .001**
Time (Linear) × Sex	0.00	0.00	0.00	924.40	.545	0.00	-0.01	0.00	900.40	.403
Time (Linear) × African American	0.01	0.01	0.01	984.40	.148	0.01	0.01	0.01	965.40	.087
Time (Linear) × Hispanic	0.02	0.01	0.01	956.90	.083	0.02	0.01	0.01	931.10	.057
Time (Linear) × Income-to-Needs Ratio	0.00	-0.01	0.00	938.60	.113	0.00	-0.01	0.00	917.10	.056
Verbal Comprehension	-0.01	-0.13	0.00	955.70	**< .001**	–	–	–	–	–
Expressive Language	–	–	–	–	–	0.00	-0.05	0.00	947.30	.790
Time (Linear) × Verbal Comprehension	0.00	0.01	0.00	927.90	.275	–	–	–	–	–
Time (Linear) × Expressive Language	–	–	–	–	–	0.00	0.02	0.00	912.00	**.003**

R^2 (fixed effects)	.249	.242
R^2 (fixed and random effects)	.520	.522

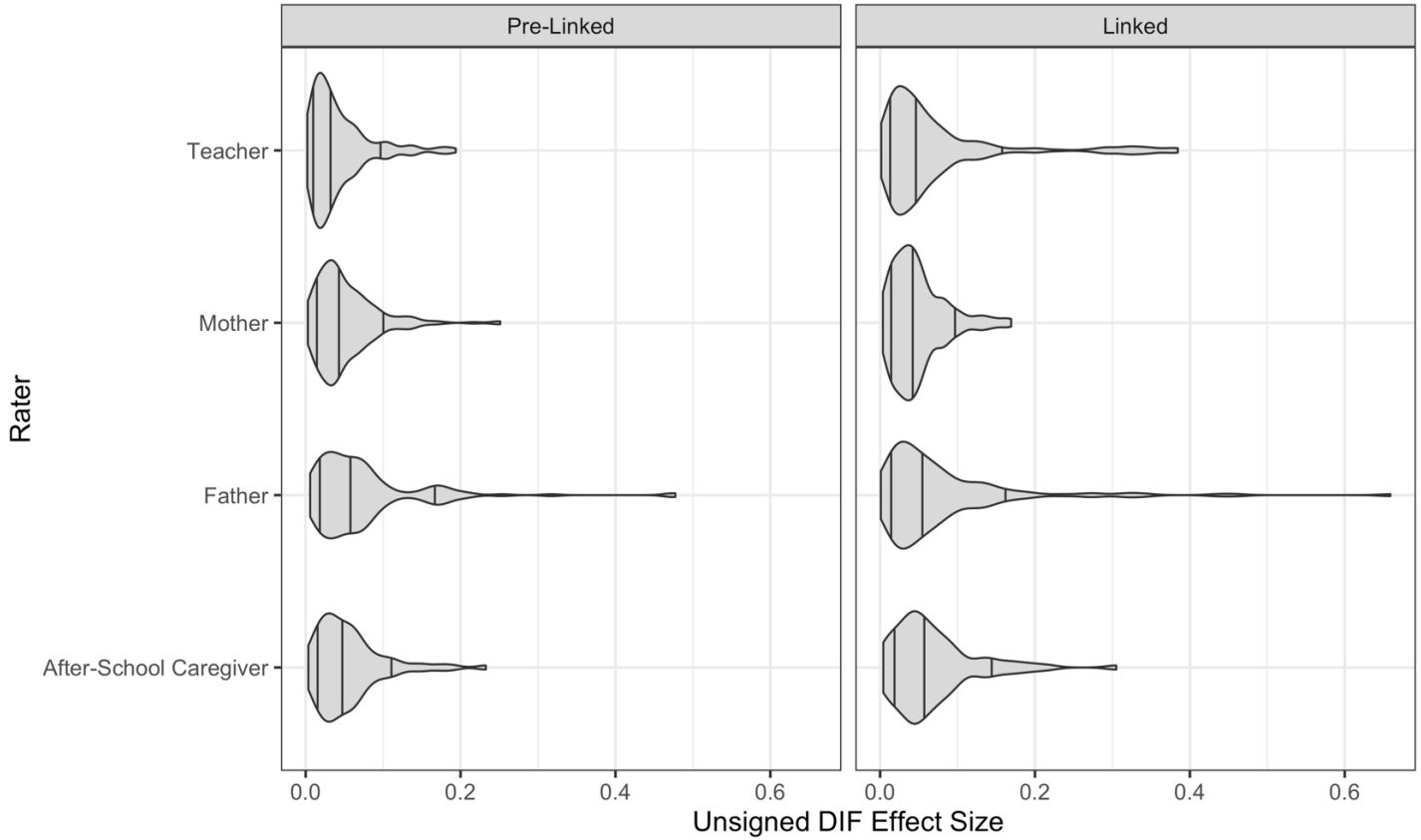
Note: Significant *p*-values in bold. "Time" (in years) was centered to set the intercepts at the last time point (age 15). For example, time is coded such that age 2 = -13 and age 15 = 0. Mothers served as the reference rater to which fathers, teachers, after-school caregivers, and self-report were compared. Interaction terms with time reflect predictions of the linear or quadratic slopes. For instance, "Time (Linear) × Father" reflects differences in slopes of fathers' ratings (compared to slopes of mothers' ratings). Self-report was not allowed to predict the slopes because it was assessed at only one time point. Sex was coded such that male = 0 and female = 1. In terms of ethnicity, Whites served as the reference group to which Blacks and Hispanics were compared. "–" indicates not applicable because the particular term was not estimated in that model.

Supplementary Table S13. Growth Model with Language Ability and Biological Covariates.

	Predicting Verbal Comprehension					Predicting Expressive Language				
	B	β	*SE*	*df*	*p*	*B*	β	*SE*	*df*	*p*
Intercept	0.07	-0.12	0.45	586.30	.871	-0.25	-0.12	0.47	566.30	.598
Time (Linear)	-0.03	0.85	0.02	793.00	.147	-0.06	0.86	0.02	746.30	**.008**
Time (Quadratic)	0.01	1.07	0.00	10610.00	**< .001**	0.01	1.08	0.00	10340.00	**< .001**
Father	0.13	0.01	0.05	10850.00	**.011**	0.13	0.01	0.05	10570.00	**.011**
Teacher	2.09	-0.17	0.12	10770.00	**< .001**	2.12	-0.17	0.12	10490.00	**< .001**
After-School Caregiver	-0.90	-0.10	1.09	10690.00	.410	-0.95	-0.10	1.10	10410.00	.391
Self-Report	0.86	0.16	0.05	10610.00	**< .001**	0.88	0.17	0.05	10330.00	**< .001**
Time (Linear) × Father	0.02	0.03	0.02	10720.00	.291	0.03	0.03	0.02	10440.00	.216
Time (Linear) × Teacher	0.70	1.20	0.03	10810.00	**< .001**	0.70	1.22	0.03	10530.00	**< .001**
Time (Linear) × After-School Caregiver	-0.25	-0.17	0.32	10680.00	.427	-0.26	-0.17	0.32	10400.00	.419
Time (Quadratic) × Father	0.00	0.01	0.00	10670.00	.708	0.00	0.02	0.00	10400.00	.577
Time (Quadratic) × Teacher	0.04	0.95	0.00	10870.00	**< .001**	0.04	0.96	0.00	10590.00	**< .001**
Time (Quadratic) × After-School Caregiver	-0.02	-0.21	0.02	10680.00	.310	-0.02	-0.21	0.02	10400.00	.309
Sex	-0.12	-0.07	0.07	508.70	.092	-0.16	-0.10	0.07	496.00	**.024**
African American	0.20	0.04	0.11	491.50	**.074**	0.32	0.07	0.11	480.10	**.005**
Hispanic	-0.02	-0.03	0.14	469.00	.865	0.02	-0.02	0.15	457.10	.912
Income-to-Needs Ratio	-0.03	-0.06	0.02	473.20	**.041**	-0.04	-0.10	0.02	457.90	**.004**
Time (Linear) × Sex	0.00	0.01	0.01	471.60	.576	0.00	0.01	0.01	458.10	.450
Time (Linear) × African American	0.01	0.01	0.01	499.60	.247	0.01	0.01	0.01	489.10	.241
Time (Linear) × Hispanic	0.01	0.01	0.01	464.20	.262	0.01	0.01	0.01	451.00	.197
Time (Linear) × Income-to-Needs Ratio	0.00	-0.01	0.00	465.00	.388	0.00	-0.01	0.00	447.80	.480
Mean Arterial Blood Pressure	0.00	0.03	0.00	476.30	.256	0.00	0.02	0.00	463.30	.398
Cortisol	-0.25	-0.04	0.14	479.80	.087	-0.24	-0.04	0.15	467.10	.105
Physical Activity	0.01	0.02	0.01	479.40	.413	0.00	0.02	0.01	466.40	.564
Verbal Comprehension	-0.01	-0.17	0.00	474.00	**< .001**	–	–	–	–	–

Expressive Language	–	–	–	–	–	0.00	-0.12	0.00	460.50	.125
Time (Linear) × Verbal Comprehension	0.00	0.02	0.00	478.20	.072	–	–	–	–	–
Time (Linear) × Expressive Language	–	–	–	–	–	0.00	0.03	0.00	458.00	**.002**
R^2 (fixed effects)			.261					.255		
R^2 (fixed and random effects)			.506					.511		

Note: Significant *p*-values in bold. "Time" (in years) was centered to set the intercepts at the last time point (age 15). For example, time is coded such that age 2 = -13 and age 15 = 0. Mothers served as the reference rater to which fathers, teachers, after-school caregivers, and self-report were compared. Interaction terms with time reflect predictions of the linear or quadratic slopes. For instance, "Time (Linear) × Father" reflects differences in slopes of fathers' ratings (compared to slopes of mothers' ratings). Self-report was not allowed to predict the slopes because it was assessed at only one time point. Sex was coded such that male = 0 and female = 1. In terms of ethnicity, Whites served as the reference group to which Blacks and Hispanics were compared. "–" indicates not applicable because the particular term was not estimated in that model.



Supplementary Figure S1. Violin plots of the distribution of unsigned effect size statistics of differential item functioning by rater both before and after linking. Vertical lines correspond to the 10^{th}, 50^{th}, and 90^{th} percentiles.