Supplementary Appendix

Appendix S1

Item Response Theory (IRT) assumptions

Method

We evaluated three IRT assumptions: (1) uni-dimensionality—the items have one predominant dimension reflecting the underlying (latent) trait (i.e., internalizing problems); (2) local independence—the items are uncorrelated when controlling for the latent dimension; and (3) monotonicity—the probability of endorsing a higher level on an item increases as the person's trait level of internalizing problems increases. Our criterion for uni-dimensionality was a ratio of first to second eigenvalues of ≥ 3.0 for an unrotated factor solution (Morizot, Ainsworth, & Reise, 2007). We evaluated the local independence of items by examining the X^2 local dependence (LD) statistic between each pair of items after controlling for the latent dimension (Chen & Thissen, 1997). We compared the LD statistics against a chi-square distribution using the mirt package (Chalmers, 2012) in R 3.3.2 to determine the proportion of pairwise items that showed greater dependency than would be expected by chance. We evaluated monotonicity by fitting non-parametric IRT models using Mokken scale analysis in the mokken package (van der Ark, 2007) in R. We examined the number of violations of monotonicity (i.e., decreases in the item step response function by rest score group) whose size was significantly greater than zero.

Results

In terms of dimensionality, the ratios of the first to second eigenvalues from unrotated factor solutions ranged from 3.22 to 3.90 across ages 14–24, suggesting that the data were "uni-dimensional enough" for IRT. In terms of local dependency, with an alpha level of .05 that was not corrected for multiple testing (4,650 pairwise associations), we observed that 4.7–8.8% of pairwise associations had statistically significant dependency, depending on the year. Thus, after accounting for the expected Type I error rate of 5%, there were between 0–3.8% of associations showing greater linear dependency than would be expected by chance, depending on the year. Thus, there was modest evidence of some local dependency at some ages. Nevertheless, IRT is robust to low and moderate violations of the local independence assumption (Fennessy, 1995). In terms of monotonicity, no items showed statistically significant non-monotonicity at any ages.

Discussion

Given evidence supporting that we approximately met the assumptions of IRT, we proceeded with the IRT approach to vertical scaling.

Appendix S2

Differential Item Functioning (DIF)

Method

After fitting IRT models, we examined whether there was differential item functioning (DIF) across age (comparable to tests of longitudinal measurement/factorial invariance). DIF examines whether the likelihood of endorsing a particular item differs between groups (in this case, between ages) for people with the same trait levels. DIF was explored using a multiple group framework in IRT in which item parameters were estimated simultaneously in the same model (i.e., concurrent calibration). In this framework, the baseline model was one in which item parameter estimates were allowed to vary *across* items, but item parameter estimates were constrained to be equal *within* item across time (allowing discrimination to differ from severity). To explore DIF, item parameters were iteratively allowed to vary across time. For example, the discrimination parameters were iteratively allowed to vary item by item to see if estimation of unique discrimination parameters at each age resulted in better model fit (based on nested model comparisons using chi-square difference tests). When exploring DIF for the discrimination parameters, a chi-square statistic with nine degrees of freedom was used to identify items with DIF. A similar procedure was used for the two severity terms, which resulted in a chi-square statistic with 18 degrees of freedom. To limit the impact of multiple testing (37 items × 2 parameters = 74 tests), we set an alpha level of .01 for identifying DIF, resulting in cutoffs of chi-square statistics greater than 21.66 and 34.80 for discrimination and severity, respectively. We also examined DIF by sex.

We then examined the effect size of DIF. Using a framework first defined by Raju (1990) and discussed by Meade (2010), the signed and unsigned differences between expected scores across ages were used to quantify magnitude of DIF. Both signed and unsigned differences were used to help explore whether DIF was non-uniform. With non-uniform DIF, the expected score curves across ages may cross and may cause the signed differences between expected scores to be zero, whereas the unsigned differences sum the absolute value of differences to approximate absolute expected score differences. Expected scores were estimated using a range of internalizing problem scores from eight standard deviations below the mean to eight standard deviations above the mean, and the average signed and unsigned difference were calculated for items showing DIF. Age 14 was used as the reference age for all calculations. The metric of these measures is in the raw score metric; for example, an effect size of 0.1 would indicate that scores at the focal age are 0.1 points larger compared to scores at age 14 (Meade, 2010).

Upon identifying an item as having parameters that differed across time, two additional IRT models were explored to assess the impact of DIF on the resulting internalizing problem scores. First, a partially constrained model was used that (a) constrained parameter estimates to equality within item across time for items showing no evidence of DIF and (b) freely estimated DIF parameters across time for items with evidence of DIF by age. Second, a model was explored that freely estimated all item parameters across ages. Both models were estimated using multiple group (concurrent calibration) IRT models with the mirt package in R. Model fit indices were used to identify which model was best fitting. Factor scores were generated to compare to the factor scores from IRT models fit separately by age, and were transformed to be on a common age 14 metric using vertical scaling (using calculations described in the Method section of the manuscript). Comparisons of model fit and factor scores were conducted as a sensitivity analysis to determine the sensitivity of the model (separate linking versus multiple group; impact of DIF) on the internalizing problems factor scores.

Results

We examined DIF by sex and by age. We are hesitant to interpret findings of DIF by sex because we had relatively small subgroups for fitting multiple group models by sex, which may have resulted in unreliable parameter estimates. We observed some instances of DIF by sex at some ages. Consistent with the interpretation that multiple group models by sex may have yielded unreliable parameter estimates, however, the items showing DIF were not consistent across ages, suggesting that items mostly did not reliably differ between males and females.

Initial exploration of DIF by age revealed that, out of 37 items, four items showed evidence of DIF in terms of discrimination and nine items showed evidence of DIF in terms of severity (all of which were common items). Three of these items showed evidence of DIF with respect to both discrimination and severity. No consistent trend was found with respect to the directionality of DIF. Some items showed increases in severity or discrimination with age, whereas other items showed decreases.

We examined the effect size of DIF. On average, DIF across time tended to have a small effect size, ranging from 0–0.16 raw score points. The signed and unsigned metrics were similar within an item across age, suggesting uniform DIF (i.e. the expected score curves did not cross), which is unsurprising because there was less evidence of DIF with respect to discrimination compared to severity.

The two multiple group models, one that allowed item parameters with evidence of DIF to be estimated freely across time and another that freely estimated all parameters, showed similar model fit. When using the chi-square model fit statistic, the model with all parameters freely estimated fit the data the best, which is unsurprising given the sample size and lengthy developmental span. However, when accounting for model complexity with a statistic such as the Akaike Information Criterion (AIC), the partially constrained model showed evidence of best model fit. Thus, model fit was not *considerably* worse fitting (accounting for model complexity) when constraining the non-DIF parameters across time to help anchor the latent variable to the same scale, suggesting that we were measuring the same construct in an equivalent way over time. At the same time, the model with all parameter estimates freely estimated was the best overall fitting model, which supports our decision to use separate IRT estimation and linking in the context of heterotypic continuity over a lengthy developmental span.

Discussion

We observed several items showing potential changes in discrimination or severity over time. Changes in severity are expected across a lengthy developmental span and are less likely than changes in discrimination to be serious threats to measuring the same construct. Compared to changes in severity, changes in discrimination are potentially more serious because they may reflect that an item does not tap the same construct at some ages. However, changes in discrimination may instead reflect meaningful developmental shifts in the construct (heterotypic continuity) when the items tap 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 items showing evidence of DIF showed changes in severity rather than discrimination, and effect sizes of DIF were small. Moreover, even for those items that changed in discrimination, they were still highly discriminating across time, 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. All instances of DIF had small effect sizes. Our vertical scaling approach accounted for DIF by estimating a separate IRT model at each age, thus allowing items' parameters to change over time, and using scaling parameters to link the scores across age and "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. Importantly, we found the same results with (a) a partially constrained model (within-item parameter constraints across time for non-DIF parameters), (b) a model excluding the items showing DIF, (c) separate IRT estimation and linking, and (d) Thurstone scaling, providing further confidence in our findings.

Appendix S3

Thurstone Scaling

When the different measures have common items over time, the Thurstone scaling approach to vertical scaling uses common items administered across ages to link the measures on the same scale by aligning their percentile scores based on a range of *z*-scores on the common items. This is based on the assumption that the two age groups to be linked have the same form of distribution (i.e., are normally distributed on the underlying trait within group), and that the groups' scores on the measures might differ in their mean and standard deviation.

Method

Vertical scaling involves placing two measures that assess a similar construct but differ in difficulty/severity on the same scale. Ideally, the two measures should have some items with the same contents to ensure scores on the measures can be linked (i.e., made comparable). In the present study, we used the Thurstone scaling approach to vertical scaling (as described in Kolen & Brennan, 2014) to transform scores on the YASR to the scale of the YSR (in addition to the IRT approach described in the manuscript). See Figure S1 for a depiction of the Thurstone scaling approach to vertical scaling.

The YSR and YASR have different but overlapping item content, so we needed to put them on the same scale. We applied Thurstone scaling that scales the scores across the different measures using the items that are in common across both measures (i.e., a common-item design, see Figure 1). Although the common items are used to determine the general form of change on the same scale, all developmentally relevant, construct-valid items are used to estimate each person's trait level on this scale. To link two measures using Thurstone scaling in a common-item design, *z*-scores are calculated from the percentile scores of the raw scores on the common items for each measure. A set of *z*-scores on the common items of each measure is selected for linking the two measures (ideally 10–20 z-scores on the common items of each measure that are not in the extremes of the distribution to prevent a distorted transformed scale). The same number of *z*-scores are selected from the common items of each measure to generate *z*-score pairs for linking the two measures (e.g., in the present study, we selected 17 *z*-scores from the common items of each measure resulting in 17 *z*-score pairs). The first assumption of Thurstone scaling in a common-item design is that the association between these *z*-score pairs (i.e., the selected set of *z*-scores for the common items) of the two measures to be linked is linear (Kolen & Brennan, 2014). The second assumption of Thurstone scaling is that the underlying trait is normally distributed. After examining the assumptions of Thurstone scaling, we linked YASR scores at age 20 to YSR scores at age 19 (i.e., the target scale that serves as the anchor). To do this, we applied the following steps, separately for males and females[1], of Thurstone scaling in a common-item design (Kolen & Brennan, 2014) to link YASR scores with YSR scores:

(1) To ensure a meaningful mean-level of change of internalizing problem scores across ages 14–24, we first examined scores on the 17 common items (i.e., the items that were common to both the YSR and YASR). The raw frequency distribution of scores on the common items is in Tables S1–S2). Participants’ mean scores on the common items are depicted in Panel A of Figure S2. The percentile ranks of raw scores on the common items are in Tables S3–S4.

(2) For both ages, we calculated *z*-scores of raw scores on the common items within age based on the percentile ranks from step 1, see Tables S5–S6.

[1] This was done because of the robust sex differences in levels of internalizing problems among adolescents and young adults, with females having higher levels than males (Hankin et al., 1998).

(3) For both ages, we calculated the mean and standard deviation of the unique *z*-scores (i.e., the *z*-score values in a given column of Tables S5–S6)[3] based on those unique raw scores whose associated *z*-scores were between -2 and +2 at the age of the target scale (i.e., age 19). We selected *z*-scores between -2 and +2 as recommended by Kolen and Brennan (2014) and because trimmed *z*-scores are more accurate in Thurstone scaling than using all *z*-scores. We calculated the *population* standard deviation of the *z*-scores, consistent with Kolen and Brennan (2014); all descriptive statistics of the sample used the *sample* standard deviation. The calculated mean and standard deviation of the *z*-scores between -2 and +2 at the age of the target scale are at the bottom of Tables S5–S6.

(4) The mean and standard deviation of the scaled scores were calculated by the following formulas (adapted from Kolen & Brennan, 2014):

$$\mu(\text{YASR}) = \sigma(\text{YSR})\left[\mu(z_{\text{YSR}}) - \frac{\sigma(z_{\text{YSR}})}{\sigma(z_{\text{YASR}})}\mu(z_{\text{YASR}})\right] + \mu(\text{YSR}) \tag{S1}$$

$$\sigma(\text{YASR}) = \frac{\sigma(z_{\text{YSR}})}{\sigma(z_{\text{YASR}})}\,\sigma(\text{YSR}) \tag{S2}$$

where YASR and YSR are vectors of raw scores on the YASR or YSR, respectively; z_{YASR} and z_{YSR} are the unique *z*-scores based on those raw scores whose associated *z*-scores were between -2 and +2 at age 19 (from step 3). The first component of Formula S1 scales the mean of the *z*-scores of the common items at age 20 to be on a *z*-score metric relative to the *z*-score metric of the common items at age 19 in order to retain changes in means and variances from age 19 to 20: $\left[\mu(z_{\text{YSR}}) - \frac{\sigma(z_{\text{YSR}})}{\sigma(z_{\text{YASR}})}\mu(z_{\text{YASR}})\right]$. The second component of Formula S1 then multiplies the scaled *z*-score metric by the standard deviation of the target scale and adds the mean of the target scale. This re-scales the age 20 *z*-score metric (on a scale that is relative to the age 19 *z*-score metric) to the metric of the total raw score at age 19 in order to make the re-scaled scores at age 20 comparable to the raw scores at age 19. Thus, the YASR scores at age 20 are re-scaled to be on the scale of the YSR at age 19, while still retaining changes in means and variances over time (based on the changes in means and variances of the common items).

Consistent with recommendations (Kolen & Brennan, 2014), we then applied a process of linking and chaining to link the remaining YASR scores to the YSR metric at age 19 based on the raw frequency distribution (Tables S7–S8), percentile ranks (Tables S9–S10), and *z*-scores of the total raw scores (Tables S11–S12). To do so, we repeated steps 1–4 by (a) linking the YASR scores at age 21 to the newly scaled YASR scores at age 20, (b) linking scores at age 22 to the newly scaled scores at age 21, (c) linking scores at age 23 to the newly scaled scores at age 22, and (d) linking scores at age 24 to the newly scaled scores at age 23. Linking and chaining allowed us to calculate a mean and standard deviation for the scaled YASR score at each age from ages 20 to 24. Based on the mean and standard deviation of the scaled scores for each year, we calculated a conversion table for converting YASR scores to YSR equivalents based on the scale of the YSR scores at age 19. We calculated a conversion table by multiplying the *z*-scores of the total raw scores (Tables S11–S12) by the standard deviation of the scaled score (from Equation S2; bottom of Tables S5–S6) and added the mean of the scaled score (from Equation S1; bottom of Tables S5–S6). The conversion table for converting YASR scores to YSR equivalents is in Table S13.

Results

Assumptions of Thurstone Scaling. We examined the two assumptions of Thurstone scaling in a common-item design: (1) the association between the selected *z*-score pairs from the common items of the two measures to be linked is linear, and (2) the underlying trait is normally distributed. Regarding assumption 1, we observed that the associations between the selected *z*-

score pairs from the common items of adjacent years were highly linear; no curvilinearity was observed. Regarding assumption 2, we observed that the raw total Internalizing scores were positively skewed (skew values ranged from 0.93 to 1.72 across years). Despite the skewed scores, it is plausible that the underlying trait (i.e., the internalizing spectrum) is normally distributed, especially given evidence that internalizing problems are dimensionally rather than categorically distributed (Markon, Chmielewski, & Miller, 2011). That is, the latent trait of internalizing is likely normally distributed even if the observed scores are not, which would be consistent with the assumption of Thurstone scaling. Thus, given theoretical and empirical evidence supporting that we approximately met these assumptions, we proceeded with vertical scaling using Thurstone scaling.

Linking the YASR Scores to the Scale of the YSR at Age 19. Next, we linked the YASR and YSR scores so that scores on the two measures were on the same scale and could be compared. To link the two measures, we re-scaled the YASR scores at age 20 to the scale of the YSR at age 19 (see steps 1–4 from the Statistical Analysis section of the Method section). The Thurstone scaling approach to vertical scaling is depicted in Figure S1. First, we examined scores at ages 19 and 20 on the 17 common items (i.e., the items that were common to both the YSR and YASR), and calculated percentile ranks, see Tables S3–S4. Second, we calculated *z*-scores of raw scores on the common items within age (Tables S5–S6) based on the percentile ranks from step 1. Third, we calculated the mean and standard deviation of the unique *z*-scores based on those raw scores (males: 0–15; females: 0–16) whose associated *z*-scores were between -2 and +2 at the age of the target scale (i.e., age 19).[2] The mean and standard deviation of these

[2] That is, we used the *unique* *z*-scores, not the vector of all *z*-scores from the raw scores (multiple participants may have the same raw score and therefore the same *z*-score).

z-scores are at the bottom of Tables S5–S6. Fourth, we calculated the mean and standard deviation of the scaled scores on the scale of the YSR at age 19 based on the mean and standard deviation of the raw scores and selected *z*-scores at each age.

The mean of the scaled score at age 20 for females was calculated as (equation S1): $\mu(\mathrm{YASR}_{20}) = 7.40\left[0.24 - \frac{0.96}{0.93}0.26\right] + 9.70 = 9.50.$[3] The standard deviation of the scaled score at age 20 for females was calculated as (equation S2): $\sigma(\mathrm{YASR}_{20}) = \frac{0.96}{0.93}\,7.40 = 7.63.$[4] We then applied linking and chaining to link the remaining YASR scores to the YSR metric at age 19. To do so, we repeated steps 1–4 by linking the YASR scores at age 21 to the newly scaled YASR scores at age 20, linking scores at age 22 to the newly scaled scores at age 21, etc.

Conversion Table for Converting YASR Scores to YSR Equivalents. Linking and chaining allowed us to calculate a mean and standard deviation for the scaled YASR score at each age from ages 20 to 24. Based on the mean and standard deviation of the scaled scores for each year, we converted YASR scores to YSR equivalents based on the scale of the YSR scores at age 19. To do this, we multiplied the *z*-scores of the total raw scores (Tables S11–S12) by the standard deviation of the scaled score (from Equation S2; bottom of Table S13) and added the mean of the scaled score (from Equation S1; bottom of Table S13).

The conversion table for converting YASR scores to YSR equivalents in our sample is in Table S13. Visual examination of the conversion table shows that many of the scores were

[3] These values came from the following sources: 7.40 (Table S7), 0.24 (Table S5), 0.96 (Table S5), 0.93 (Table S5), 0.26 (Table S5), 9.70 (Table S7), 9.50 (Table 2). Note that the above calculations are slightly different from the actual calculations due to rounding error.

[4] These values came from the following sources: 0.96 (Table S5), 0.93 (Table S5), 7.40 (Table S7), 7.63 (Table 2). Note that the above calculations are slightly different from the actual calculations due to rounding error.

highly similar before and after rescaling, while rescaling changed some of the scores by more than 2 points (particularly for females). The mean and standard deviation of the scaled scores are at the bottom of Table 13. Participants' mean internalizing problem scores, after rescaling the YASR scores to the metric of the YSR, are depicted in Panel B of Figure S2. Notably, the scores retained a highly similar pattern of mean-level change when examining the re-scaled total scores compared to when examining just the common items (see Panel A of Figure S2). Thus, the Thurstone Scaling approach successfully retained mean-level change when re-scaling the YASR scores to be on the same metric as the YSR while still using a more comparable scale.

Growth Curve Model. To examine growth curves, we first compared a linear growth curve model to a quadratic growth curve model in HLM to identify the best-fitting form of change for the rescaled internalizing problem scores. The model that allowed quadratic slopes to vary across individuals (i.e., random quadratic slopes) was not positive definite (i.e., not all variances in the variance-covariance matrix were non-zero and positive), likely because the variance in the quadratic term was close to zero ($\tau_{22} < .0001$). The small variance in the quadratic term suggested that individuals did not significantly differ in quadratic curvature. A model with random linear slopes and a quadratic term that was fixed across individuals (i.e., fixed quadratic slopes) fit better than a model with only random linear slopes ($\chi^2[1] = 24.46$, $p < .001$). A model with fixed cubic slopes fit better than the model with random linear slopes and fixed quadratic slopes ($\chi^2[1] = 5.63$, $p = .018$). A model with fixed quartic slopes fit better than the cubic model ($\chi^2[1] = 4.65$, $p = .031$), and was the best-fitting model (a model with fixed fifth-degree polynomial slopes did not significantly improve fit; $\chi^2[1] = 3.07$, $p = .080$). Individuals' quartic trajectories, and the average quartic trajectory for males and females are depicted in Figure S3. The average quartic trajectory showed slight decreases over time, primarily for females.

Overall, the growth curves showed little curvature, which would be consistent with evidence that likelihood ratio tests may be sensitive to small fit differences with larger sample sizes (Tomarken & Waller, 2003). Thus, the polynomial growth terms may have over-fit the data, especially given the lengthy developmental span. Moreover, there are difficulties in interpreting and replicating findings from polynomial growth models, and mapping polynomial growth terms onto developmental theory (Grimm, Ram, & Hamagami, 2011). For these reasons, for comparing the common items to the rescaled scores and for examining the predictors of change in internalizing problems, we examined the general form of change by examining the linear model for ease of interpretation.

In the linear growth curve model with no predictors of the intercepts or slopes, intercepts reflected an individual's estimated initial level of internalizing problems at age 14. Slopes reflected participants' linear change in internalizing problems over time. There was evidence of a negative slope (B = -0.08, t[3980] = -1.94, p = .053). In a similar growth curve model examining the trajectories of scores on the common items, however, the slope was not significant (B = -0.03, t[3980] = -1.08, p = .282).

Although the form of change for the age-relevant items versus common items was highly similar at the *group-level*, there were differences at the *individual-level*. Some participants showed *decreases* in internalizing problems over time when using the age-relevant items while they showed *increases* in internalizing problems when using the common items (or vice versa). The participants who showed decreases using the age-relevant items and increases using the common items presumably had higher levels of internalizing problems on the *non-common* items of the YSR (i.e., items that were on the Internalizing scale of the YSR but not the YASR) or lower levels on the *non-common* items of the YASR (compared to the other participants). Because the Somatic Complaints subscale was included in the Internalizing Scale of the YSR but not YASR, the majority (9 items, 60%) of the non-common Internalizing items of the YSR were items assessing somatic complaints. Therefore, we examined participants' levels of somatic complaints on the YSR. Consistent with expectations, participants who showed decreases in internalizing problems using the age-relevant items but increases using the common items showed higher mean levels of somatic complaints from ages 14–19 (*M* = 3.22) than participants who did not (*M* = 1.81; *t*[32.38] = -3.41, *p* = .002). The reverse was also true; participants who showed increases in internalizing problems using the age-relevant items but decreases using the common items, showed lower mean levels of somatic complaints from ages 14–19 (*M* = 0.93) than participants who did not (*M* = 1.94; *t*[30.76] = 4.49, *p* < .001).

We then examined sex and ethnicity as predictors of the intercepts and linear slopes of the rescaled internalizing problem scores (see Table S14). There were no significant linear slopes when controlling for the other model predictors. Females showed higher intercepts than males, but males and females did not significantly differ in their linear slopes. African Americans and those of "other" ethnicity did not significantly differ from European Americans in their intercepts or linear slopes. The model accounted for approximately three-fourths of the variance in internalizing problems over time.

Table S1. Raw score frequency distributions on common items for females.

	Age (years)									
score	14	15	16	17	19	20	21	22	23	24
0	6	10	8	5	12	17	14	15	20	21
1	12	8	8	18	13	18	25	13	23	18
2	19	19	21	18	30	27	20	17	12	18
3	20	18	23	23	20	21	15	19	19	19
4	18	17	15	16	20	20	23	21	17	14
5	13	12	23	14	12	14	15	15	22	21
6	18	14	15	20	11	19	27	25	15	19
7	16	15	16	17	28	18	12	17	20	7
8	13	12	11	8	10	18	11	15	12	13
9	11	14	10	10	15	14	18	18	18	19
10	11	9	10	6	20	9	14	9	14	20
11	9	13	15	7	6	10	10	8	9	6
12	11	4	7	10	5	7	6	9	8	6
13	6	6	9	11	1	5	8	7	5	11
14	7	9	5	5	6	4	2	2	10	12
15	7	2	7	6	3	6	5	6	5	5
16	5	6	2	2	1	2	2	2	6	3
17	3	7	4	7	4	3	3	5	3	2
18	5	2	5	2	5	2	1	4	2	2
19	1	5	4	2	1	4	3	3	4	5
20	1	1	2	1	4	3	1	2	1	1
21	0	2	1	2	1	0	1	0	0	2
22	0	4	0	2	1	0	2	2	0	1
23	0	0	3	1	1	0	2	1	1	1
24	0	1	1	1	0	1	0	0	3	1
25	0	1	0	0	0	2	0	0	0	2
26	0	0	0	0	0	0	1	0	0	0
27	0	0	0	0	0	0	1	1	0	0
28	1	0	1	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0
M	7.42	8.05	7.80	7.42	6.77	6.64	6.79	7.16	7.06	7.34
SD	5.03	5.77	5.58	5.45	5.09	5.25	5.37	5.25	5.35	5.62

Note: mean and standard deviation reflect the mean and standard deviation of the participants' scores on the common items (they do not reflect the mean and standard deviation of the values in the above column).

Table S2. Raw score frequency distributions on common items for males.

	Age (years)									
score	14	15	16	17	19	20	21	22	23	24
0	16	16	17	22	18	27	33	27	34	26
1	27	24	24	23	26	30	18	30	22	20
2	24	21	26	33	29	22	25	25	31	21
3	24	21	21	22	27	23	26	24	18	23
4	17	11	25	23	26	21	18	18	19	16
5	11	23	27	17	18	20	15	11	18	13
6	10	18	14	13	12	14	11	14	15	11
7	17	11	10	10	17	10	15	12	12	19
8	13	6	11	5	12	9	11	12	8	10
9	6	4	8	11	8	10	7	8	9	10
10	5	6	9	6	8	6	6	10	6	4
11	6	7	8	9	7	8	6	5	10	6
12	7	8	4	3	2	8	7	7	6	4
13	3	6	9	2	6	9	4	8	10	10
14	4	4	4	3	3	4	3	2	5	3
15	5	4	1	1	5	2	1	6	2	1
16	1	0	1	1	1	4	4	1	1	3
17	0	1	0	2	2	2	6	3	1	3
18	1	1	3	1	3	3	3	1	5	3
19	0	1	2	5	1	1	4	4	3	2
20	0	1	1	0	2	0	0	1	1	2
21	0	0	0	1	0	0	0	0	0	2
22	1	1	0	1	0	0	0	0	0	1
23	1	1	1	1	0	0	0	0	0	1
24	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	1	0	0	0	0
26	0	0	0	0	0	1	0	0	0	1
27	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	1	0	0	0	0	0
29	0	0	0	0	0	0	0	1	1	0
M	5.31	5.63	5.54	5.24	5.49	5.58	5.52	5.59	5.59	6.11
SD	4.52	4.76	4.54	4.89	4.76	5.03	5.04	5.15	5.17	5.53

Note: mean and standard deviation reflect the mean and standard deviation of the participants' scores on the common items (they do not reflect the mean and standard deviation of the values in the above column).

Table S3. Percentile ranks (divided by 100) on common items for females.

	Age (years)									
score	14	15	16	17	19	20	21	22	23	24
0	.01	.02	.02	.01	.03	.04	.03	.03	.04	.04
1	.06	.07	.05	.07	.08	.11	.11	.09	.13	.12
2	.13	.13	.12	.15	.17	.20	.20	.16	.20	.19
3	.22	.22	.22	.25	.28	.30	.28	.23	.26	.27
4	.31	.30	.30	.34	.37	.38	.36	.32	.33	.33
5	.38	.37	.38	.41	.44	.45	.43	.39	.41	.41
6	.46	.43	.47	.49	.49	.52	.52	.48	.49	.49
7	.54	.50	.54	.57	.57	.59	.60	.57	.55	.54
8	.61	.56	.60	.63	.66	.67	.65	.64	.62	.58
9	.66	.63	.64	.67	.71	.73	.71	.70	.68	.64
10	.71	.68	.69	.71	.79	.78	.77	.76	.74	.72
11	.76	.73	.74	.74	.84	.82	.82	.80	.79	.77
12	.81	.77	.79	.78	.87	.86	.86	.83	.82	.80
13	.85	.80	.83	.83	.88	.88	.88	.87	.85	.83
14	.88	.83	.86	.87	.90	.90	.90	.89	.88	.88
15	.91	.86	.88	.89	.92	.92	.92	.90	.91	.91
16	.94	.88	.90	.91	.93	.93	.93	.92	.93	.93
17	.96	.91	.92	.93	.93	.95	.95	.94	.95	.94
18	.98	.93	.94	.95	.96	.95	.95	.95	.96	.94
19	.99	.95	.96	.96	.97	.97	.96	.97	.97	.96
20	.99	.96	.97	.97	.98	.98	.97	.98	.98	.97
21	–	.97	.98	.98	.99	–	.98	–	–	.98
22	–	.98	–	.99	.99	–	.98	.99	–	.98
23	–	–	.99	.99	1.00	–	.99	.99	.99	.99
24	–	.99	.99	1.00	–	.99	–	–	1.00	.99
25	–	1.00	–	–	–	1.00	–	–	–	1.00
26	–	–	–	–	–	–	.99	–	–	–
27	–	–	–	–	–	–	1.00	1.00	–	–
28	1.00	–	1.00	–	–	–	–	–	–	–
29	–	–	–	–	–	–	–	–	–	–

Table S4. Percentile ranks (divided by 100) on common items for males.

	Age (years)									
score	14	15	16	17	19	20	21	22	23	24
0	.04	.04	.04	.05	.04	.06	.08	.06	.07	.06
1	.15	.14	.13	.16	.13	.18	.19	.18	.19	.17
2	.28	.26	.24	.29	.25	.29	.29	.30	.30	.27
3	.40	.37	.35	.41	.37	.39	.40	.41	.41	.37
4	.50	.45	.45	.52	.48	.48	.50	.50	.49	.46
5	.57	.54	.56	.61	.58	.57	.57	.57	.56	.53
6	.62	.64	.65	.68	.64	.64	.63	.62	.63	.58
7	.69	.71	.70	.73	.71	.69	.69	.67	.69	.65
8	.77	.76	.75	.77	.76	.73	.75	.73	.73	.72
9	.81	.78	.79	.81	.81	.77	.79	.77	.77	.76
10	.84	.81	.83	.85	.84	.80	.82	.81	.80	.80
11	.87	.84	.87	.88	.88	.83	.84	.84	.83	.82
12	.90	.88	.89	.91	.89	.87	.87	.87	.86	.84
13	.93	.91	.92	.92	.91	.91	.90	.90	.90	.87
14	.94	.94	.95	.93	.93	.93	.91	.92	.93	.91
15	.97	.96	.96	.94	.95	.94	.92	.94	.95	.92
16	.98	–	.97	.95	.96	.96	.93	.96	.95	.93
17	–	.97	–	.95	.97	.97	.96	.97	.96	.94
18	.99	.98	.98	.96	.98	.98	.98	.97	.97	.95
19	–	.98	.99	.98	.99	.99	1.00	.98	.99	.96
20	–	.99	.99	–	.99	–	–	.99	.99	.97
21	–	–	–	.99	–	–	–	–	–	.98
22	.99	.99	–	.99	–	–	–	–	–	.99
23	1.00	1.00	1.00	1.00	–	–	–	–	–	.99
24	–	–	–	–	–	–	–	–	–	–
25	–	–	–	–	–	.99	–	–	–	–
26	–	–	–	–	–	1.00	–	–	–	1.00
27	–	–	–	–	–	–	–	–	–	–
28	–	–	–	–	1.00	–	–	–	–	–
29	–	–	–	–	–	–	–	1.00	1.00	–

Table S5. *Z*-scores of common items for females.

	Age (years)									
score	14	15	16	17	19	20	21	22	23	24
0	-1.47	-1.40	-1.40	-1.36	-1.33	-1.26	-1.27	-1.36	-1.32	-1.31
1	-1.28	-1.22	-1.22	-1.18	-1.13	-1.07	-1.08	-1.17	-1.13	-1.13
2	-1.08	-1.05	-1.04	-0.99	-0.94	-0.88	-0.89	-0.98	-0.95	-0.95
3	-0.88	-0.88	-0.86	-0.81	-0.74	-0.69	-0.71	-0.79	-0.76	-0.77
4	-0.68	-0.70	-0.68	-0.63	-0.54	-0.50	-0.52	-0.60	-0.57	-0.59
5	-0.48	-0.53	-0.50	-0.44	-0.35	-0.31	-0.33	-0.41	-0.39	-0.42
6	-0.28	-0.36	-0.32	-0.26	-0.15	-0.12	-0.15	-0.22	-0.20	-0.24
7	-0.08	-0.18	-0.14	-0.08	0.05	0.07	0.04	-0.03	-0.01	-0.06
8	0.12	-0.01	0.04	0.11	0.24	0.26	0.22	0.16	0.17	0.12
9	0.31	0.16	0.21	0.29	0.44	0.45	0.41	0.35	0.36	0.30
10	0.51	0.34	0.39	0.47	0.63	0.64	0.60	0.54	0.55	0.47
11	0.71	0.51	0.57	0.66	0.83	0.83	0.78	0.73	0.74	0.65
12	0.91	0.68	0.75	0.84	1.03	1.02	0.97	0.92	0.92	0.83
13	1.11	0.86	0.93	1.02	1.22	1.21	1.16	1.11	1.11	1.01
14	1.31	1.03	1.11	1.21	1.42	1.40	1.34	1.30	1.30	1.18
15	1.51	1.20	1.29	1.39	1.62	1.59	1.53	1.49	1.48	1.36
16	1.71	1.38	1.47	1.57	1.81	1.78	1.72	1.68	1.67	1.54
17	1.91	1.55	1.65	1.76	2.01	1.98	1.90	1.87	1.86	1.72
18	2.10	1.72	1.83	1.94	2.21	2.17	2.09	2.07	2.04	1.90
19	2.30	1.90	2.01	2.12	2.40	2.36	2.28	2.26	2.23	2.07
20	2.50	2.07	2.18	2.31	2.60	2.55	2.46	2.45	2.42	2.25
21	–	2.24	2.36	2.49	2.80	–	2.65	–	–	2.43
22	–	2.42	–	2.67	2.99	–	2.83	2.83	–	2.61
23	–	–	2.72	2.86	3.19	–	3.02	3.02	2.98	2.79
24	–	2.76	2.90	3.04	–	3.31	–	–	3.16	2.96
25	–	2.94	–	–	–	3.50	–	–	–	3.14
26	–	–	–	–	–	–	3.58	–	–	–
27	–	–	–	–	–	–	3.77	3.78	–	–
28	4.09	–	3.62	–	–	–	–	–	–	–
29	–	–	–	–	–	–	–	–	–	–
M	0.12	-0.01	0.04	0.11	0.24	0.26	0.22	0.16	0.17	0.12
SD	0.97	0.85	0.88	0.90	0.96	0.93	0.91	0.93	0.92	0.87

Note: the dashed line reflects those raw scores at age 19 whose associated *z*-scores were between -2 and +2 (i.e., raw scores of 0 to 16). Mean and standard deviation reflect the mean and standard deviation of the *z*-scores whose associated raw scores ranged from 0 to 16 (i.e., the mean and standard deviation of the values above the dashed line).

Table S6. *Z*-scores of common items for males.

	Age (years)									
score	14	15	16	17	19	20	21	22	23	24
0	-1.18	-1.18	-1.22	-1.07	-1.15	-1.11	-1.10	-1.09	-1.08	-1.11
1	-0.95	-0.97	-1.00	-0.87	-0.94	-0.91	-0.90	-0.89	-0.89	-0.92
2	-0.73	-0.76	-0.78	-0.66	-0.73	-0.71	-0.70	-0.70	-0.69	-0.74
3	-0.51	-0.55	-0.56	-0.46	-0.52	-0.51	-0.50	-0.50	-0.50	-0.56
4	-0.29	-0.34	-0.34	-0.25	-0.31	-0.31	-0.30	-0.31	-0.31	-0.38
5	-0.07	-0.13	-0.12	-0.05	-0.10	-0.12	-0.10	-0.11	-0.11	-0.20
6	0.15	0.08	0.10	0.16	0.11	0.08	0.10	0.08	0.08	-0.02
7	0.38	0.29	0.32	0.36	0.32	0.28	0.29	0.27	0.27	0.16
8	0.60	0.50	0.54	0.57	0.53	0.48	0.49	0.47	0.47	0.34
9	0.82	0.71	0.76	0.77	0.74	0.68	0.69	0.66	0.66	0.52
10	1.04	0.92	0.98	0.97	0.95	0.88	0.89	0.86	0.85	0.70
11	1.26	1.13	1.20	1.18	1.16	1.08	1.09	1.05	1.05	0.88
12	1.48	1.34	1.42	1.38	1.37	1.28	1.29	1.25	1.24	1.06
13	1.70	1.55	1.64	1.59	1.58	1.47	1.49	1.44	1.43	1.25
14	1.93	1.76	1.86	1.79	1.79	1.67	1.68	1.63	1.63	1.43
15	2.15	1.97	2.08	2.00	2.00	1.87	1.88	1.83	1.82	1.61
16	2.37	–	2.30	2.20	2.21	2.07	2.08	2.02	2.01	1.79
17	–	2.39	–	2.41	2.42	2.27	2.28	2.22	2.21	1.97
18	2.81	2.60	2.74	2.61	2.63	2.47	2.48	2.41	2.40	2.15
19	–	2.81	2.96	2.82	2.84	2.67	2.68	2.61	2.59	2.33
20	–	3.02	3.18	–	3.05	–	–	2.80	2.79	2.51
21	–	–	–	3.23	–	–	–	–	–	2.69
22	3.70	3.44	–	3.43	–	–	–	–	–	2.87
23	3.92	3.65	3.84	3.64	–	–	–	–	–	3.05
24	–	–	–	–	–	–	–	–	–	–
25	–	–	–	–	–	3.86	–	–	–	–
26	–	–	–	–	–	4.06	–	–	–	3.60
27	–	–	–	–	–	–	–	–	–	–
28	–	–	–	–	4.72	–	–	–	–	–
29	–	–	–	–	–	–	–	4.55	4.53	–
M	0.49	0.39	0.43	0.46	0.42	0.38	0.39	0.37	0.37	0.25
SD	1.02	0.97	1.01	0.94	0.97	0.92	0.92	0.90	0.89	0.83

Note: the dashed line reflects those raw scores at age 19 whose associated *z*-scores were between -2 and +2 (i.e., raw scores of 0 to 15). Mean and standard deviation reflect the mean and standard deviation of the *z*-scores whose associated raw scores ranged from 0 to 15 (i.e., the mean and standard deviation of the values above the dashed line).

Table S7. Raw score frequency distributions on all items for females.

	Age (years)									
score	14	15	16	17	19	20	21	22	23	24
0	5	9	5	4	8	6	7	8	12	11
1	10	4	2	10	8	19	15	9	14	13
2	11	9	12	14	19	13	12	13	14	16
3	14	14	12	15	16	21	19	15	15	17
4	12	10	18	17	21	16	20	16	8	13
5	4	10	11	8	11	9	12	10	13	13
6	15	20	12	12	11	15	10	17	21	14
7	13	11	13	17	9	20	20	13	15	15
8	16	8	12	10	12	13	16	17	7	9
9	8	7	14	11	14	16	11	13	12	10
10	14	10	13	7	12	9	11	13	11	8
11	8	6	7	11	11	13	8	12	18	9
12	14	9	16	9	13	12	13	11	11	11
13	7	6	13	5	12	15	14	7	14	13
14	4	10	7	8	5	6	7	15	8	16
15	8	6	5	6	8	1	4	7	6	9
16	5	7	3	4	3	6	6	3	5	8
17	6	6	6	7	3	4	6	5	6	4
18	2	2	5	5	4	5	5	5	8	7
19	3	3	4	0	4	3	3	2	6	7
20	5	10	1	4	4	4	5	4	4	4
21	4	4	4	3	3	3	2	5	2	4
22	9	4	4	7	2	3	3	2	4	1
23	2	3	4	2	2	1	2	1	4	2
24	0	3	8	4	0	1	0	4	1	2
25	3	4	1	1	2	3	0	2	3	2
26	1	3	2	1	4	2	4	3	2	3
27	3	1	4	0	2	1	1	0	1	2
28	0	1	1	3	2	1	0	1	0	2
29	4	2	1	2	1	0	1	1	0	0
30	0	3	0	1	0	0	2	0	0	1
31	0	1	0	1	1	0	0	0	3	0
32	0	2	0	3	1	1	0	0	0	1
33	2	1	0	0	1	1	1	0	0	0
34	0	2	1	1	1	1	0	1	1	1
35	0	0	0	1	0	0	1	0	0	1
36	0	0	2	0	0	0	1	0	0	0
37	0	0	1	0	0	0	0	1	0	0
38	0	0	1	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0
41	0	0	1	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0
47	1	0	0	0	0	0	0	0	0	0
M	10.80	11.64	11.38	10.55	9.70	9.07	9.42	9.77	9.84	10.07
SD	7.85	8.38	8.09	8.03	7.40	6.83	7.13	6.90	7.10	7.42

Note: mean and standard deviation reflect the mean and standard deviation of the participants' scores on all items (they do not reflect the mean and standard deviation of the values in the above column).

Table S8. Raw score frequency distributions on all items for males.

	Age (years)									
score	14	15	16	17	19	20	21	22	23	24
0	9	11	13	17	13	13	14	9	15	15
1	14	16	18	18	17	16	18	18	24	10
2	20	18	15	24	18	24	20	29	17	17
3	21	14	19	19	26	28	20	20	14	18
4	15	14	19	19	18	8	16	16	29	16
5	5	21	16	15	19	14	14	16	14	19
6	12	11	16	10	16	24	14	11	10	12
7	17	9	13	14	13	14	15	9	14	11
8	9	11	16	11	15	7	9	13	9	10
9	12	6	8	7	11	13	11	14	12	14
10	9	9	7	6	10	10	8	6	11	7
11	11	8	10	5	4	6	8	11	6	11
12	6	7	9	7	7	5	6	5	5	3
13	7	3	9	4	11	5	9	5	6	3
14	4	6	4	3	5	5	4	5	7	7
15	4	3	3	7	4	5	4	6	6	3
16	2	3	5	1	1	7	6	5	4	3
17	2	5	4	4	4	10	3	2	6	4
18	4	4	7	6	2	3	1	6	8	5
19	0	2	2	3	1	3	1	10	4	6
20	3	2	4	1	5	2	5	1	2	3
21	3	1	0	0	2	2	1	5	1	0
22	1	3	1	2	3	1	4	0	2	2
23	4	1	1	2	2	2	4	0	2	4
24	0	1	1	2	3	2	3	2	1	1
25	0	2	1	1	0	2	2	2	5	1
26	0	1	2	1	1	1	2	1	0	2
27	0	0	0	1	0	1	0	0	0	2
28	2	1	1	3	0	0	1	2	2	1
29	0	0	0	0	1	0	0	0	0	2
30	2	1	1	1	1	0	0	0	0	1
31	0	1	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0
34	1	1	1	0	0	0	0	0	0	1
35	0	0	0	0	1	1	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0
37	0	0	0	1	0	1	0	0	0	1
38	0	0	0	0	0	0	0	0	1	0
39	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	1	0	0
M	7.98	8.02	7.89	7.52	7.65	7.94	7.95	8.10	8.04	8.74
SD	6.66	6.93	6.45	7.12	6.51	6.80	6.73	6.86	6.94	7.55

Note: mean and standard deviation reflect the mean and standard deviation of the participants' scores on all items (they do not reflect the mean and standard deviation of the values in the above column).

Table S9. Percentile ranks (divided by 100) on all items for females.

	Age (years)									
score	14	15	16	17	19	20	21	22	23	24
0	.01	.02	.01	.01	.02	.01	.02	.02	.02	.02
1	.05	.05	.03	.04	.05	.07	.06	.06	.08	.07
2	.10	.09	.06	.10	.11	.13	.12	.10	.13	.13
3	.15	.14	.11	.17	.19	.20	.18	.16	.19	.20
4	.22	.19	.18	.24	.27	.27	.26	.22	.24	.26
5	.25	.24	.24	.30	.34	.33	.33	.28	.28	.31
6	.30	.31	.29	.35	.39	.38	.37	.34	.35	.36
7	.37	.39	.35	.42	.43	.45	.43	.40	.42	.42
8	.43	.43	.40	.48	.47	.52	.51	.47	.47	.47
9	.49	.47	.46	.53	.53	.57	.57	.53	.50	.51
10	.54	.51	.52	.57	.59	.63	.61	.58	.55	.54
11	.59	.55	.57	.61	.64	.67	.65	.64	.61	.58
12	.64	.58	.62	.66	.69	.72	.69	.69	.67	.62
13	.69	.62	.68	.69	.74	.78	.75	.72	.71	.67
14	.72	.65	.73	.72	.78	.82	.79	.77	.76	.72
15	.75	.69	.75	.75	.81	.84	.81	.82	.79	.78
16	.78	.73	.77	.78	.83	.85	.83	.84	.81	.81
17	.80	.75	.79	.80	.85	.87	.86	.86	.83	.83
18	.82	.77	.81	.83	.86	.89	.88	.88	.86	.86
19	.84	.79	.83	–	.88	.91	.90	.89	.89	.88
20	.85	.82	.85	.85	.90	.92	.92	.90	.91	.90
21	.87	.85	.85	.87	.91	.93	.93	.92	.92	.92
22	.91	.87	.87	.89	.92	.95	.94	.94	.93	.93
23	.93	.89	.89	.91	.93	.95	.95	.94	.95	.94
24	–	.90	.92	.93	–	.96	–	.95	.96	.94
25	.94	.91	.94	.94	.94	.97	–	.97	.97	.95
26	.95	.93	.94	.94	.95	.98	.96	.98	.98	.96
27	.96	.94	.96	–	.97	.98	.98	–	.98	.97
28	–	.95	.97	.95	.97	.99	–	.99	–	.98
29	.98	.95	.97	.96	.98	–	.98	.99	–	–
30	–	.97	–	.97	–	–	.98	–	–	.99
31	–	.98	–	.98	.99	–	–	–	.99	–
32	–	.98	–	.99	.99	.99	–	–	–	.99
33	.99	.99	–	–	.99	.99	.99	–	–	–
34	–	1.00	.98	.99	1.00	1.00	–	.99	1.00	.99
35	–	–	–	1.00	–	–	.99	–	–	1.00
36	–	–	.98	–	–	–	1.00	–	–	–
37	–	–	.99	–	–	–	–	1.00	–	–
38	–	–	.99	–	–	–	–	–	–	–
39	–	–	–	–	–	–	–	–	–	–
40	–	–	–	–	–	–	–	–	–	–
41	–	–	1.00	–	–	–	–	–	–	–
42	–	–	–	–	–	–	–	–	–	–
43	–	–	–	–	–	–	–	–	–	–
44	–	–	–	–	–	–	–	–	–	–
45	–	–	–	–	–	–	–	–	–	–
46	–	–	–	–	–	–	–	–	–	–
47	1.00	–	–	–	–	–	–	–	–	–

Table S10. Percentile ranks (divided by 100) on all items for males.

	Age (years)									
score	14	15	16	17	19	20	21	22	23	24
0	.03	.03	.03	.04	.03	.03	.03	.02	.03	.04
1	.08	.10	.10	.12	.09	.09	.10	.08	.11	.09
2	.17	.18	.17	.22	.17	.17	.19	.18	.20	.16
3	.27	.27	.25	.32	.26	.29	.28	.29	.27	.24
4	.36	.34	.33	.41	.35	.36	.36	.37	.36	.32
5	.41	.43	.41	.49	.44	.41	.43	.43	.45	.40
6	.45	.51	.48	.54	.51	.49	.49	.50	.50	.47
7	.53	.56	.54	.60	.57	.57	.56	.54	.55	.53
8	.59	.61	.61	.66	.63	.62	.61	.59	.60	.57
9	.64	.65	.66	.70	.69	.66	.65	.64	.64	.63
10	.70	.69	.69	.73	.73	.71	.70	.69	.69	.68
11	.75	.73	.73	.76	.76	.74	.73	.73	.73	.72
12	.79	.78	.77	.79	.79	.77	.76	.76	.75	.75
13	.82	.80	.81	.81	.82	.79	.80	.78	.77	.77
14	.85	.82	.84	.83	.86	.81	.83	.80	.80	.79
15	.87	.85	.86	.85	.88	.83	.84	.83	.83	.81
16	.88	.86	.88	.87	.89	.86	.87	.85	.85	.83
17	.89	.88	.89	.88	.90	.89	.89	.87	.87	.84
18	.91	.90	.92	.90	.91	.92	.90	.88	.90	.87
19	–	.92	.94	.93	.92	.93	.90	.92	.92	.89
20	.93	.93	.95	.93	.93	.94	.91	.94	.94	.91
21	.94	.94	–	–	.94	.95	.93	.96	.95	–
22	.95	.95	.96	.94	.96	.96	.94	–	.95	.92
23	.96	.96	.97	.95	.97	.96	.96	–	.96	.93
24	–	.96	.97	.96	.98	.97	.97	.97	.97	.95
25	–	.97	.98	.97	–	.98	.98	.98	.98	.95
26	–	.98	.98	.97	.99	.99	.99	.99	–	.96
27	–	–	–	.98	–	.99	–	–	–	.97
28	.98	.98	.99	.99	–	–	1.00	.99	.99	.98
29	–	–	–	–	.99	–	–	–	–	.98
30	.99	.99	.99	.99	.99	–	–	–	–	.99
31	–	.99	–	–	–	–	–	–	–	–
32	–	–	–	–	–	–	–	–	–	–
33	–	–	–	–	–	–	–	–	–	–
34	1.00	1.00	1.00	–	–	–	–	–	–	.99
35	–	–	–	–	1.00	.99	–	–	–	–
36	–	–	–	–	–	–	–	–	–	–
37	–	–	–	1.00	–	1.00	–	–	–	1.00
38	–	–	–	–	–	–	–	–	1.00	–
39	–	–	–	–	–	–	–	–	–	–
40	–	–	–	–	–	–	–	1.00	–	–

Table S11. *Z*-scores of all items for females.

	Age (years)									
score	14	15	16	17	19	20	21	22	23	24
0	-1.38	-1.39	-1.41	-1.31	-1.31	-1.33	-1.32	-1.41	-1.39	-1.36
1	-1.25	-1.27	-1.28	-1.19	-1.17	-1.18	-1.18	-1.27	-1.25	-1.22
2	-1.12	-1.15	-1.16	-1.07	-1.04	-1.04	-1.04	-1.12	-1.10	-1.09
3	-0.99	-1.03	-1.03	-0.94	-0.90	-0.89	-0.90	-0.98	-0.96	-0.95
4	-0.87	-0.91	-0.91	-0.82	-0.77	-0.74	-0.76	-0.84	-0.82	-0.82
5	-0.74	-0.79	-0.79	-0.69	-0.63	-0.60	-0.62	-0.69	-0.68	-0.68
6	-0.61	-0.67	-0.66	-0.57	-0.50	-0.45	-0.48	-0.55	-0.54	-0.55
7	-0.48	-0.55	-0.54	-0.44	-0.36	-0.30	-0.34	-0.40	-0.40	-0.41
8	-0.36	-0.43	-0.42	-0.32	-0.23	-0.16	-0.20	-0.26	-0.26	-0.28
9	-0.23	-0.32	-0.29	-0.19	-0.09	-0.01	-0.06	-0.11	-0.12	-0.14
10	-0.10	-0.20	-0.17	-0.07	0.04	0.14	0.08	0.03	0.02	-0.01
11	0.03	-0.08	-0.05	0.06	0.18	0.28	0.22	0.18	0.16	0.13
12	0.15	0.04	0.08	0.18	0.31	0.43	0.36	0.32	0.30	0.26
13	0.28	0.16	0.20	0.31	0.45	0.58	0.50	0.47	0.45	0.40
14	0.41	0.28	0.32	0.43	0.58	0.72	0.64	0.61	0.59	0.53
15	0.54	0.40	0.45	0.55	0.72	0.87	0.78	0.76	0.73	0.66
16	0.66	0.52	0.57	0.68	0.85	1.01	0.92	0.90	0.87	0.80
17	0.79	0.64	0.69	0.80	0.99	1.16	1.06	1.05	1.01	0.93
18	0.92	0.76	0.82	0.93	1.12	1.31	1.20	1.19	1.15	1.07
19	1.05	0.88	0.94	–	1.26	1.45	1.34	1.34	1.29	1.20
20	1.17	1.00	1.07	1.18	1.39	1.60	1.48	1.48	1.43	1.34
21	1.30	1.12	1.19	1.30	1.53	1.75	1.62	1.63	1.57	1.47
22	1.43	1.24	1.31	1.43	1.66	1.89	1.76	1.77	1.71	1.61
23	1.56	1.35	1.44	1.55	1.80	2.04	1.90	1.92	1.85	1.74
24	–	1.47	1.56	1.68	–	2.19	–	2.06	2.00	1.88
25	1.81	1.59	1.68	1.80	2.07	2.33	–	2.21	2.14	2.01
26	1.94	1.71	1.81	1.92	2.20	2.48	2.32	2.35	2.28	2.15
27	2.07	1.83	1.93	–	2.34	2.63	2.47	–	2.42	2.28
28	–	1.95	2.05	2.17	2.47	2.77	–	2.64	–	2.42
29	2.32	2.07	2.18	2.30	2.61	–	2.75	2.79	–	–
30	–	2.19	–	2.42	–	–	2.89	–	–	2.69
31	–	2.31	–	2.55	2.88	–	–	–	2.98	–
32	–	2.43	–	2.67	3.01	3.36	–	–	–	2.96
33	2.83	2.55	–	–	3.15	3.51	3.31	–	–	–
34	–	2.67	2.80	2.92	3.28	3.65	–	3.51	3.41	3.23
35	–	–	–	3.05	–	–	3.59	–	–	3.36
36	–	–	3.04	–	–	–	3.73	–	–	–
37	–	–	3.17	–	–	–	–	3.94	–	–
38	–	–	3.29	–	–	–	–	–	–	–
39	–	–	–	–	–	–	–	–	–	–
40	–	–	–	–	–	–	–	–	–	–
41	–	–	3.66	–	–	–	–	–	–	–
42	–	–	–	–	–	–	–	–	–	–
43	–	–	–	–	–	–	–	–	–	–
44	–	–	–	–	–	–	–	–	–	–
45	–	–	–	–	–	–	–	–	–	–
46	–	–	–	–	–	–	–	–	–	–
47	4.61	–	–	–	–	–	–	–	–	–

Table S12. *Z*-scores of all items for males.

	Age (years)									
score	14	15	16	17	19	20	21	22	23	24
0	-1.20	-1.16	-1.22	-1.06	-1.18	-1.17	-1.18	-1.18	-1.16	-1.16
1	-1.05	-1.01	-1.07	-0.92	-1.02	-1.02	-1.03	-1.03	-1.02	-1.02
2	-0.90	-0.87	-0.91	-0.78	-0.87	-0.87	-0.88	-0.89	-0.87	-0.89
3	-0.75	-0.72	-0.76	-0.64	-0.71	-0.73	-0.74	-0.74	-0.73	-0.76
4	-0.60	-0.58	-0.60	-0.49	-0.56	-0.58	-0.59	-0.60	-0.58	-0.63
5	-0.45	-0.44	-0.45	-0.35	-0.41	-0.43	-0.44	-0.45	-0.44	-0.50
6	-0.30	-0.29	-0.29	-0.21	-0.25	-0.29	-0.29	-0.31	-0.29	-0.36
7	-0.15	-0.15	-0.14	-0.07	-0.10	-0.14	-0.14	-0.16	-0.15	-0.23
8	0.00	0.00	0.02	0.07	0.05	0.01	0.01	-0.01	-0.01	-0.10
9	0.15	0.14	0.17	0.21	0.21	0.16	0.16	0.13	0.14	0.03
10	0.30	0.29	0.33	0.35	0.36	0.30	0.31	0.28	0.28	0.17
11	0.45	0.43	0.48	0.49	0.51	0.45	0.45	0.42	0.43	0.30
12	0.60	0.57	0.64	0.63	0.67	0.60	0.60	0.57	0.57	0.43
13	0.75	0.72	0.79	0.77	0.82	0.74	0.75	0.71	0.71	0.56
14	0.90	0.86	0.95	0.91	0.98	0.89	0.90	0.86	0.86	0.70
15	1.05	1.01	1.10	1.05	1.13	1.04	1.05	1.01	1.00	0.83
16	1.20	1.15	1.26	1.19	1.28	1.18	1.20	1.15	1.15	0.96
17	1.35	1.30	1.41	1.33	1.44	1.33	1.35	1.30	1.29	1.09
18	1.50	1.44	1.57	1.47	1.59	1.48	1.49	1.44	1.44	1.23
19	–	1.58	1.72	1.61	1.74	1.63	1.64	1.59	1.58	1.36
20	1.80	1.73	1.88	1.75	1.90	1.77	1.79	1.73	1.72	1.49
21	1.95	1.87	–	–	2.05	1.92	1.94	1.88	1.87	–
22	2.10	2.02	2.19	2.03	2.20	2.07	2.09	–	2.01	1.76
23	2.26	2.16	2.34	2.17	2.36	2.21	2.24	–	2.16	1.89
24	–	2.30	2.50	2.32	2.51	2.36	2.39	2.32	2.30	2.02
25	–	2.45	2.65	2.46	–	2.51	2.54	2.46	2.44	2.15
26	–	2.59	2.81	2.60	2.82	2.66	2.68	2.61	–	2.29
27	–	–	–	2.74	–	2.80	–	–	–	2.42
28	3.01	2.88	3.12	2.88	–	–	2.98	2.90	2.88	2.55
29	–	–	–	–	3.28	–	–	–	–	2.68
30	3.31	3.17	3.42	3.16	3.43	–	–	–	–	2.81
31	–	3.31	–	–	–	–	–	–	–	–
32	–	–	–	–	–	–	–	–	–	–
33	–	–	–	–	–	–	–	–	–	–
34	3.91	3.75	4.04	–	–	–	–	–	–	3.34
35	–	–	–	–	4.20	3.98	–	–	–	–
36	–	–	–	–	–	–	–	–	–	–
37	–	–	–	4.14	–	4.27	–	–	–	3.74
38	–	–	–	–	–	–	–	–	4.32	–
39	–	–	–	–	–	–	–	–	–	–
40	–	–	–	–	–	–	–	4.65	–	–

Table S13. Thurstone-scaled conversion table of YASR to YSR equivalents.

	Males					Females				
	Age (years)					Age (years)				
YASR score	20	21	22	23	24	20	21	22	23	24
0	-0.3	-0.4	-0.5	-0.4	-0.2	-0.6	-0.6	-0.5	-0.7	-0.6
1	0.8	0.6	0.5	0.6	0.8	0.5	0.5	0.6	0.4	0.5
2	1.8	1.6	1.5	1.6	1.8	1.6	1.6	1.7	1.5	1.6
3	2.8	2.6	2.6	2.6	2.8	2.7	2.7	2.8	2.6	2.7
4	3.8	3.6	3.6	3.7	3.8	3.8	3.8	3.9	3.7	3.8
5	4.8	4.7	4.6	4.7	4.8	4.9	4.9	5.0	4.8	4.9
6	5.8	5.7	5.6	5.7	5.8	6.1	6.0	6.1	5.9	6.0
7	6.8	6.7	6.7	6.7	6.8	7.2	7.1	7.2	7.0	7.1
8	7.8	7.7	7.7	7.7	7.8	8.3	8.2	8.3	8.1	8.2
9	8.8	8.8	8.7	8.8	8.8	9.4	9.3	9.4	9.2	9.3
10	9.9	9.8	9.7	9.8	9.8	10.5	10.4	10.5	10.3	10.5
11	10.9	10.8	10.8	10.8	10.8	11.7	11.5	11.6	11.4	11.6
12	11.9	11.8	11.8	11.8	11.8	12.8	12.6	12.7	12.5	12.7
13	12.9	12.9	12.8	12.8	12.8	13.9	13.6	13.8	13.6	13.8
14	13.9	13.9	13.8	13.9	13.8	15.0	14.7	14.9	14.7	14.9
15	14.9	14.9	14.9	14.9	14.8	16.1	15.8	16.0	15.8	16.0
16	15.9	15.9	15.9	15.9	15.8	17.2	16.9	17.2	16.9	17.1
17	16.9	16.9	16.9	16.9	16.8	18.4	18.0	18.3	18.0	18.2
18	17.9	18.0	17.9	17.9	17.8	19.5	19.1	19.4	19.1	19.3
19	19.0	19.0	19.0	18.9	18.8	20.6	20.2	20.5	20.2	20.4
20	20.0	20.0	20.0	20.0	19.8	21.7	21.3	21.6	21.3	21.5
21	21.0	21.0	21.0	21.0	–	22.8	22.4	22.7	22.4	22.6
22	22.0	22.1	–	22.0	21.8	24.0	23.5	23.8	23.5	23.7
23	23.0	23.1	–	23.0	22.8	25.1	24.6	24.9	24.6	24.8
24	24.0	24.1	24.1	24.0	23.8	26.2	–	26.0	25.7	25.9
25	25.0	25.1	25.1	25.1	24.8	27.3	–	27.1	26.8	27.0
26	26.0	26.2	26.1	–	25.8	28.4	27.9	28.2	27.9	28.1
27	27.0	–	–	–	26.8	29.5	29.0	–	29.0	29.2
28	–	28.2	28.2	28.1	27.8	30.7	–	30.4	–	30.3
29	–	–	–	–	28.8	–	31.2	31.5	–	–
30	–	–	–	–	29.8	–	32.3	–	–	32.5
31	–	–	–	–	–	–	–	–	33.3	–
32	–	–	–	–	–	35.1	–	–	–	34.7
33	–	–	–	–	–	36.2	35.5	–	–	–
34	–	–	–	–	33.8	37.4	–	37.1	36.6	36.9
35	35.1	–	–	–	–	–	37.7	–	–	38.0
36	–	–	–	–	–	–	38.8	–	–	–
37	37.2	–	–	–	36.8	–	–	40.4	–	–
38	–	–	–	38.3	–	–	–	–	–	–
39	–	–	–	–	–	–	–	–	–	–
40	–	–	40.5	–	–	–	–	–	–	–
M	7.77	7.68	7.79	7.78	8.50	9.50	9.73	10.26	10.12	10.53
SD	6.88	6.88	7.03	7.07	7.55	7.63	7.80	7.64	7.79	8.17

Note: values reflect the YASR scores on the scale of the YSR at age 19. Mean and standard deviation reflect the mean and standard deviation of the participants' re-scaled YASR scores on the YSR scale at age 19 (they do not reflect the mean and standard deviation of the values in the above column).

Table S14. Linear growth curve model of Thurstone-scaled internalizing problems.

Variable	*B*	β	*SE*	*DF*	*p*
intercept	8.225	0.015	0.445	3977	< .001
time	-0.024	-0.033	0.057	3977	.674
Predictors of the intercepts					
female	3.267	0.181	0.602	539	< .001
African American	-1.331	-0.065	0.829	539	.109
Other Ethnicity	-2.610	-0.028	2.540	539	.305
Predictors of the slopes					
female	-0.102	-0.022	0.077	3977	.187
African American	-0.011	-0.002	0.110	3977	.920
Other Ethnicity	0.162	0.009	0.322	3977	.614
Variance components	*SD*				
intercept	6.19				
time	0.74				
residual	4.26				
Correlation between intercept and slope			*r* = .47		
Model Pseudo-R^2	.747				

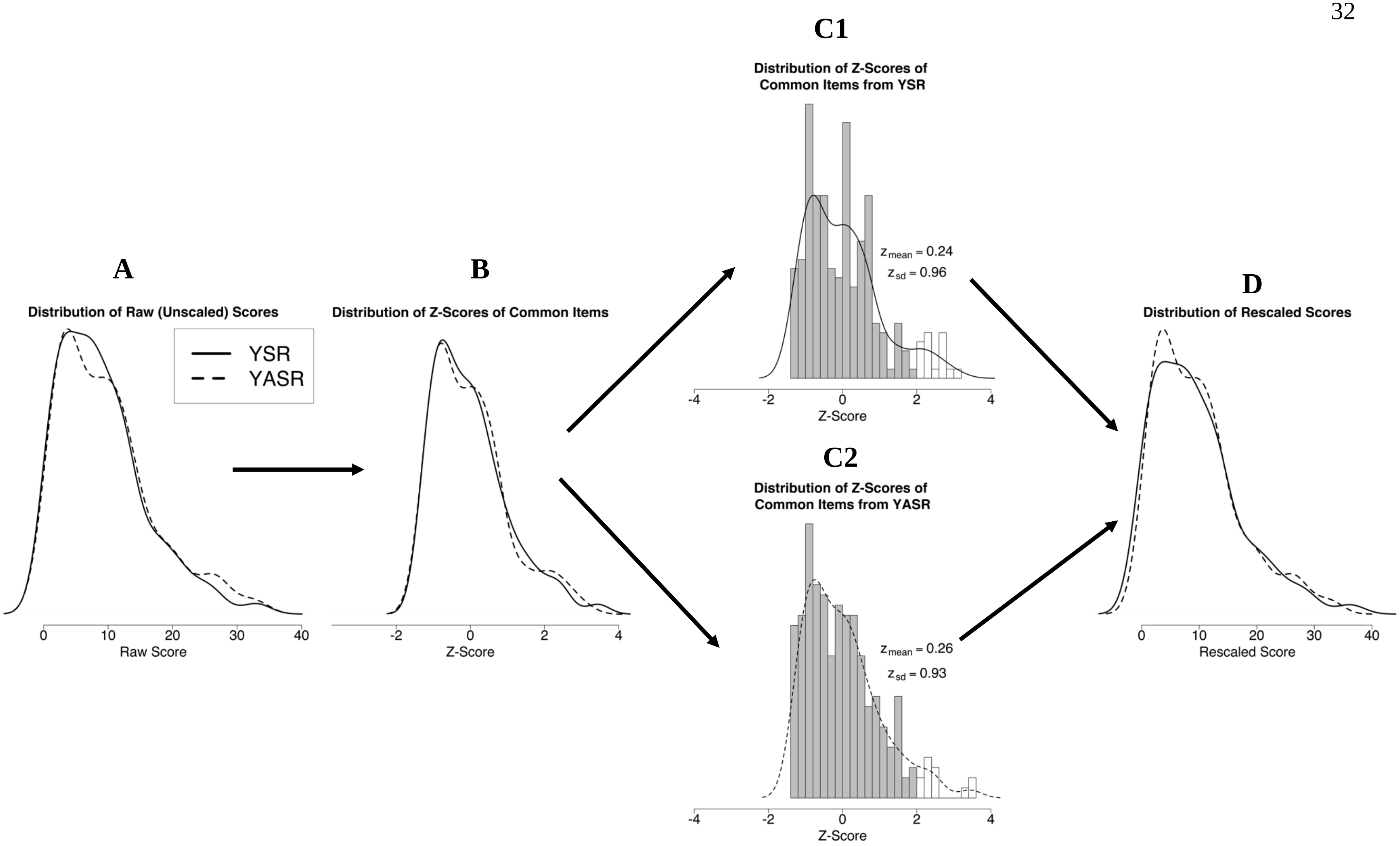

Figure S1. Depiction of steps in vertical scaling using Thurstone scaling with a common-item design. YSR = Youth Self-Report at age 19 (target scale). YASR = Young Adult Self-Report at age 20. Panel A depicts the raw score distributions of the two measures (distributions are depicted with kernel density estimation). Panel B depicts the distribution of *z*-scores of the items that are common to both measures (i.e., the common items). Panel C depicts the distribution of *z*-scores of the common items for each measure (Panel C1 = YSR, Panel C2 = YASR), along with the calculations of the mean and standard deviation of *z*-scores within the target range of -2 to +2. Each histogram bar reflects the frequency of a given *z*-score (corresponding to a given raw score) on the measure. Gray histogram bars reflect *z*-scores within the target range of -2 to +2 that were used for calculating the mean and standard deviation. Note that the *z*-score for each unique raw score i.e., gray histogram bar, is used in the calculation (rather than all observed *z*-scores), so the mean and standard deviation do not necessarily equal 0 and 1, respectively. The measures are rescaled to be on the same scale by using the mean and the standard deviation of the *z*-scores of the common items to align their percentile scores. Panel D depicts the rescaled scores (i.e., scores from the YASR on the scale of the YSR). The mean and standard deviation of the rescaled scores were calculated using Equations S1 and S2, respectively. We calculated a conversion table by multiplying the *z*-scores of the total raw scores by the standard deviation of the scaled score and added the mean of the scaled score (see Table 2). The figure shows that, in comparison to the YSR, the unscaled YASR scores were over-represented at lower levels of the scale and under-represented at upper levels of the scale (presumably because of fewer items in the YASR; see Panel A). Rescaling the scores made the scales more comparable. Note that, by design, the distributions of rescaled scores for the two measures do not perfectly overlap. Vertical scaling does not create the same distribution (mean and standard deviation) for each measure because it retains differences in means and variances across the two measures (based on the means and variances of the common items). Nevertheless, the scores are on a more comparable scale. Although the common items are used to determine the general form of change on the same scale, all developmentally relevant, construct-valid items are used to estimate each person's trait level on this scale.

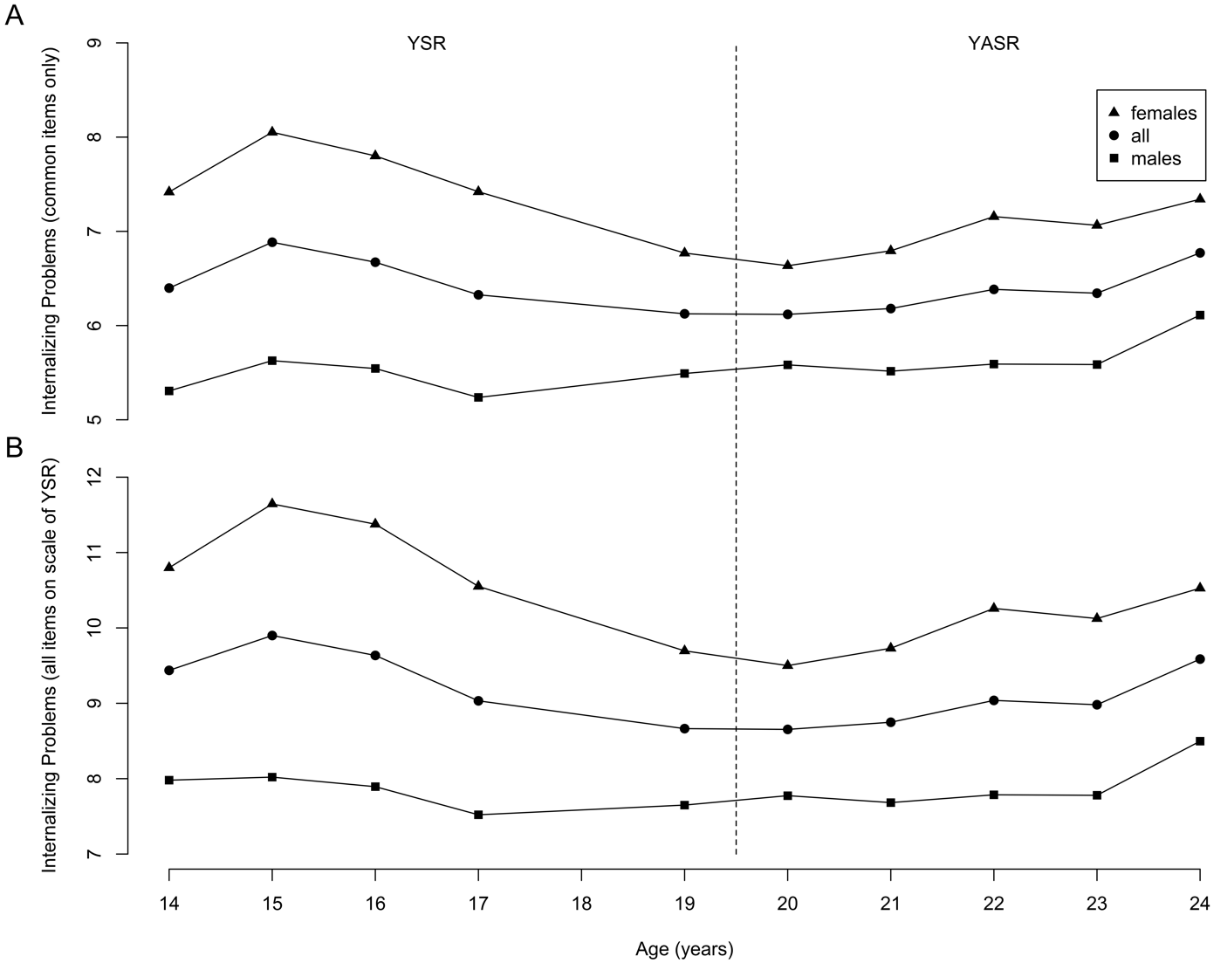

Figure S2. Panel A depicts participants' mean scores on the *common* items (i.e., the items that were common to the Internalizing scale of the Youth Self-Report, YSR, and Young Adult Self-Report, YASR). Panel B depicts participants' mean internalizing problem scores on *all* age-relevant items of the Internalizing scale, after rescaling the YASR scores to the metric of the YSR (based on the scale of the YSR at age 19) using Thurstone scaling. Internalizing problems to the left of the dashed line (i.e., ages 14–19) were rated on the YSR. Internalizing problems to the right of the dashed line (i.e., ages 20–24) were rated on the YASR. Internalizing problem reports were not collected at age 18.

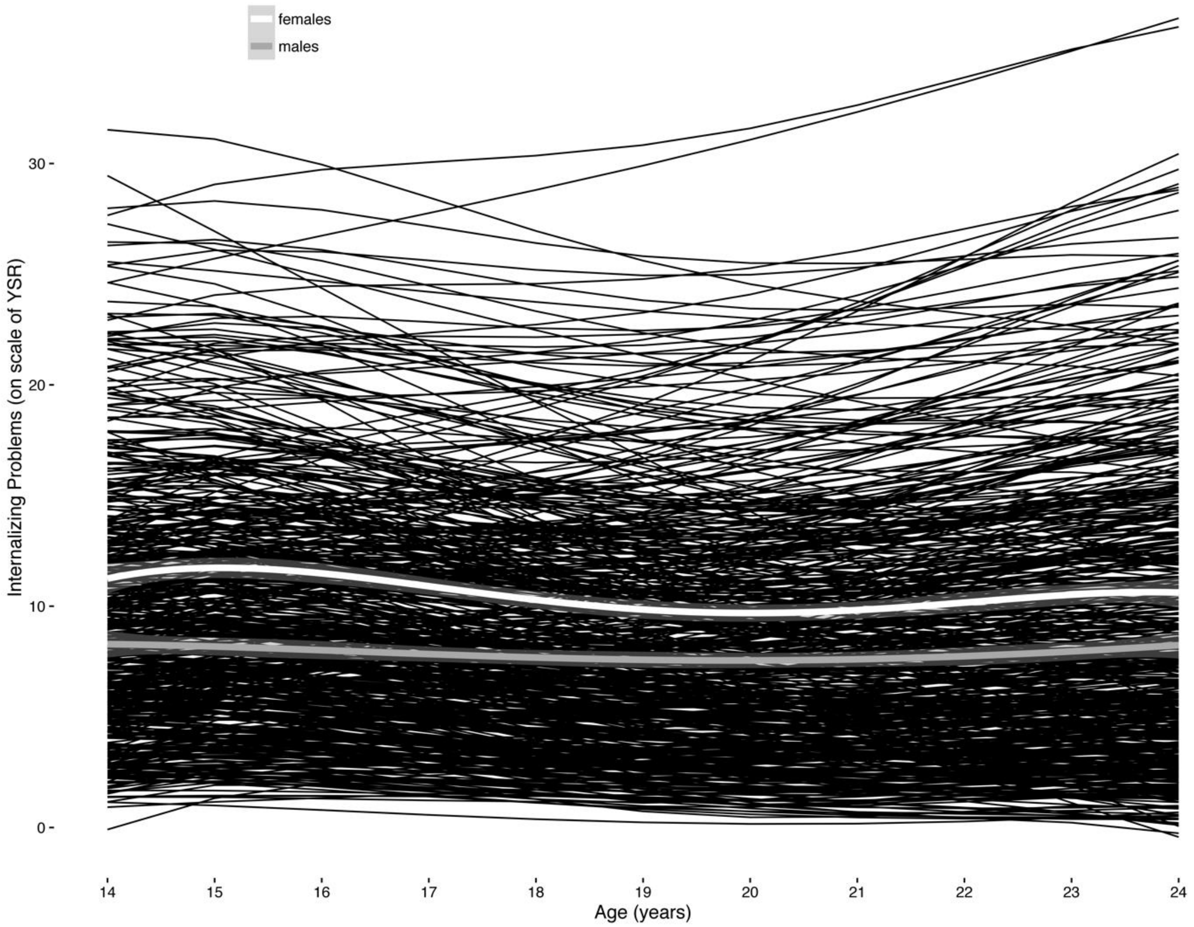

Figure S3. Individuals' fitted quartic trajectories of Thurstone-scaled internalizing problems in black. Average quartic trajectory for females in white. Average quartic trajectory for males in gray.

References

Chalmers, R. P. (2012). mirt: a multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1-29.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265-289. doi:10.3102/10769986022003265

Fennessy, L. M. (1995). *The impact of local dependencies on various IRT outcomes.* (Doctoral dissertation), Available from ProQuest Dissertations & Theses database. (UMI No. 9524701)

Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear growth curves in developmental research. *Child Development, 82*, 1357-1371. doi:10.1111/j.1467-8624.2011.01630.x

Hankin, B. L., Abramson, L. Y., Moffitt, T. E., Silva, P. A., McGee, R., & Angell, K. E. (1998). Development of depression from preadolescence to young adulthood: Emerging gender differences in a 10-year longitudinal study. *Journal of Abnormal Psychology, 107*, 128-140. doi:10.1037/0021-843x.107.1.128

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

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY, US: Springer.

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

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95,* 728-743. doi: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). New York, NY, US: Guilford Press.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207. doi:10.1177/014662169001400208

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with 'well fitting' models. *Journal of Abnormal Psychology, 112,* 578-598. doi:10.1037/0021-843X.112.4.578

van der Ark, L. A. (2007). Mokken scale analysis in R. *2007, 20*, 19. doi:10.18637/jss.v020.i11