Published in final edited form as:
Dev Psychol. 2018 March ; 54(3): 586–599. doi:10.1037/dev0000449.

Development of Internalizing Problems from Adolescence to Emerging Adulthood: Accounting for Heterotypic Continuity with Vertical Scaling

Isaac T. Petersen[a], **Oliver Lindhiem**[b], **Brandon LeBeau**[a], **John E. Bates**[c], **Gregory S. Pettit**[d], **Jennifer E. Lansford**[e], and **Kenneth A. Dodge**[e]

[a]University of Iowa

[b]University of Pittsburgh

[c]Indiana University

[d]Auburn University

[e]Duke University

Abstract

Manifestations of internalizing problems, such as specific symptoms of anxiety and depression, can change across development, even if individuals show strong continuity in rank-order levels of internalizing problems. This illustrates the concept of heterotypic continuity, and raises the question of whether common measures might be construct-valid for one age but not another. This study examines mean-level changes in internalizing problems across a long span of development at the same time as accounting for heterotypic continuity by using age-appropriate, changing measures. Internalizing problems from age 14–24 were studied longitudinally in a community sample (N = 585), using Achenbach's Youth Self-Report (YSR) and Young Adult Self-Report (YASR). Heterotypic continuity was evaluated with an item response theory (IRT) approach to vertical scaling, linking different measures over time to be on the same scale, as well as with a Thurstone scaling approach. With vertical scaling, internalizing problems peaked in mid-to-late adolescence and showed a group-level decrease from adolescence to early adulthood, a change that would not have been seen with the approach of using only age-common items. Individuals' trajectories were sometimes different than would have been seen with the common-items approach. Findings support the importance of considering heterotypic continuity when examining development and vertical scaling to account for heterotypic continuity with changing measures.

Keywords

internalizing problems; heterotypic continuity; vertical scaling; changing measures; developmental trajectories; longitudinal

Correspondence concerning this article should be addressed to Isaac T. Petersen, Department of Psychological and Brain Sciences, University of Iowa, 309 Stuit Hall, Iowa City, IA 52242. isaac-t-petersen@uiowa.edu.

Internalizing problems, including depression and anxiety, are among the most common and burdensome problems that adolescents and adults experience. The broadband, dimensional concept of internalizing problems, which represents multiple, specific symptoms, was discovered through factor-analytic test development work (Achenbach & Edelbrock, 1978). Key findings with this concept include the following: (1) The internalizing problems dimension captures individuals' meaningful clinical and sub-clinical difficulties, including risk for anxiety or mood disorders, relationship conflicts, ineffective parenting, and poor health (e.g., Eaton et al., 2013). (2) Internalizing problems have well-established norms, and (3) they show both considerable stability in individual differences over time and some rank-order change, along with (4) some normative (mean-level) fluctuations across development. These findings suggest further questions toward understanding how internalizing problems develop and toward charting both normative patterns and individual trajectories of internalizing problems.

According to considerable prior research, rates of depression and other internalizing problems have been shown to increase in adolescence, peak in mid-to-late adolescence, and decrease into adulthood (Adkins, Wang, Dupre, van den Oord, & Elder, 2009). As an example of robust sex differences in the development of internalizing problems, females show higher levels of depression than males, with sex differences emerging around the onset of puberty and the greatest sex differences appearing in mid-to-late adolescence (Hankin et al., 1998). The marked developmental changes and sex differences in depression during this developmental era make the transition from adolescence to adulthood particularly important to study.

Heterotypic Continuity

Among the many studies of internalizing problems, we have found none that examined trajectories of internalizing problems using measures adjusted to maintain construct validity and consider mean-level change over a lengthy developmental span. This is important because it appears that internalizing problems change in their manifestation over time (Avenevoli & Steinberg, 2002) and different measures may be needed at different ages to accurately understand how internalizing problems develop. Internalizing problems may manifest differently in adolescents compared to adults. It has been shown that somatic complaints (e.g., headaches, stomachaches, heart pounding) are more strongly associated with and more common in those with internalizing problems earlier than later in development (Achenbach, 1991, 1997; Ryan et al., 1987). This is an example of *heterotypic continuity*, which refers to persistence of an underlying construct or process with manifestations that change over the course of development (Petersen, Hoyniak, McQuillan, Bates, & Staples, 2016). Heterotypic continuity occurs when the *same* psychological reasons underlie *different* behaviors at different ages. Heterotypic continuity is analogous to the transformation of water to ice or steam, or of a caterpillar to a butterfly—the underlying core is preserved but the manifestation changes. Using measures that change with development to maintain construct validity of internalizing problems could be important for better understanding of (a) the normative trajectory of internalizing problems across ages, (b) individual differences in trajectories of internalizing problems, and (c) how risk and protective factors influence individuals' development of internalizing problems. This would build construct validity and advance understanding of how internalizing problems develop. Developmental psychology seeks to understand processes of continuity and change across the lifespan and not just limited windows of time or stages of life. However, studying heterotypic continuity over long spans of development poses methodological and theoretical challenges and opportunities.

The Challenge of Heterotypic Continuity when Examining Development

Measuring the development of internalizing problems

Because of the heterotypic continuity of internalizing problems, measures have been designed to accommodate changes in the manifestation of internalizing problems. Most notably, the Internalizing scale of the Youth Self-Report (YSR; Achenbach, 1991) designed for 11- to 18-year-olds includes items reflecting anxiety, depression, and somatic complaints. The Internalizing scale on the Young Adult Self-Report (YASR; Achenbach, 1997) designed for 18- to 30-year-olds includes items reflecting anxiety and depression but not somatic complaints. To chart internalizing problems from adolescence to adulthood using the YSR and YASR, then, it is a challenge to measure participants' actual change in internalizing problems, despite changing measures.

Heterotypic continuity as the focus of study

Heterotypic continuity is a developmental phenomenon examined by many researchers using structural equation modeling (SEM) or item response theory (IRT). Examining how strongly items relate to a latent trait (i.e., item factor loadings in SEM or item discrimination in IRT) can help determine which behaviors most strongly reflect a construct at a given point in development (i.e., continuity of the factor structure at the behavior/item level). Other researchers have examined the continuity of constructs at the latent/syndrome level (e.g., Snyder, Young, & Hankin, 2017). Given how much developmental research demonstrates how constructs change in their expression over time, surprisingly little research in developmental psychology has explored the best ways to account for heterotypic continuity, that is, how to examine individuals' developmental trajectories in constructs that change in their manifestation over time.

Accounting for the heterotypic continuity of internalizing problems

When examining continuity and change of individuals' trajectories on a construct, heterotypic continuity can be useful in advancing developmental theory and practice. It can also become a confound that needs to be accounted for, rather than the focus of study. When examining change, especially over a lengthy developmental span, it is important to consider and, if necessary, account for heterotypic continuity. If heterotypic continuity is not properly accounted for, the same measure may not reflect the same construct across time and, therefore, scores on the measure may not be comparable across time. To account for heterotypic continuity, changes in measurement should accommodate changes in the manifestation of the construct to retain construct validity invariance (Knight & Zerr, 2010). For example, for developmental reasons, the measurement of internalizing problems should assess somatic problems to a greater degree earlier in development. Thus, the consequence of heterotypic continuity is that *different* items over time may be necessary to assess the *same* construct over time. There are three primary approaches to measuring a construct over time, each with its own advantages and limitations.

Approaches to Measuring a Construct over Time

The three approaches to measuring a construct over time include administering (1) all possible items across all ages, (2) only the common items across all ages, and (3) the construct-valid items at each age. Traditionally, developmental psychologists have used all possible items (approach 1) or only the common items (approach 2) across all ages when measuring a construct over time. Below, we discuss the approaches and the importance of using the construct-valid items at each age instead (approach 3), which is depicted in Figure 1.

(1) All possible items across all ages

The first approach to measuring a construct over time uses all possible items across all ages. One advantage is that it is a comprehensive approach to assessment that allows examining change in each item across the developmental span. The approach has key disadvantages, however. First, it is inefficient. It requires extra time to assess all items across all ages. Second, it could assess items that are developmentally inappropriate at a given age (because of changes in item difficulty or severity, i.e., how infrequently an item is correct or endorsed). For example, in a test of math ability, it would be developmentally inappropriate to ask a 7-year-old an advanced calculus question. Third, the aggregation of scores on all possible items could result in a score that lacks construct validity invariance and therefore becomes incomparable over time if the construct changes in its manifestation (because of changes in item discrimination, i.e., how strongly the item relates to the trait). For example, a measure including somatic complaints to assess internalizing problems in adulthood may not reflect the same construct as assessed by the same measure of internalizing problems in adolescence. Thus, the same measure may not reflect the same construct at different ages. Thus, aggregating scores on all possible items across all ages could produce problems for interpretation when heterotypic continuity is likely.

(2) Only the common items across all ages

The second approach to measuring a construct over time is to use only the common items across all ages. Using only the common items across all ages has the advantage that it is efficient, but also has key disadvantages. First, using only the common items results in a loss of information because there are fewer items assessing the construct, which may make the measure less sensitive to developmental change. Excluding from all ages items that are developmentally inappropriate at some ages and developmentally appropriate at other ages could result in the systematic loss of information on the full scope of internalizing problems, which is crucial for assessing individual differences, especially at low and clinical levels of problems. For instance, in a hypothetical study of internalizing problems from 2 to 18 years of age, suicidality would not likely be used as a common item across all ages because it would be developmentally inappropriate to ask parents of a 2-year-old whether the child is suicidal. Omitting suicidality across all ages, however, would result in a loss of information regarding an important internalizing problem with high severity. Second, the measure may lack content validity because it is not measuring the construct as a whole, in particular, the age-specific manifestations.

(3) Construct-valid items at each age

The third approach to measuring a construct over time is to use the construct-valid items at each age. In the context of heterotypic continuity, this would mean using different items across time—those items at a given age that are valid for the target construct. Using the construct-valid items at each age has several advantages. First, it retains content validity and construct validity invariance. Second, it is more efficient than using all possible items across all ages. And where there is heterotypic continuity, it is the best way to maintain construct validity invariance.

There are still important issues in using the construct-valid items at each age when the items differ across time. For one, there is the question of how to measure individuals' change in a construct when a different measure is used at each age. Different scores on the measures' different items over time could reflect either (a) a person's change in the trait, or (b) an artifactual change resulting from the different measures/items at each age having different meaning. Assuming the measures reflect the same construct over time (i.e., construct validity invariance), the next consideration for determining whether different scores for an individual over time reflect actual change is the issue of statistical equivalence. Are the measures' scores on the same metric or scale so they can be meaningfully compared? First, the measures should have the same range of possible scores. Second, in order to measure absolute change (rather than solely an individual's change relative to others), a score on the measure at T1 should reflect the same trait level on the construct as the same score on the measure at T2. There are several possible solutions to ensuring statistical equivalence of different measures over time, including: (a) age-norming, (b) average/percentage scores, and (c) vertical scaling.

Age-norming

Age-norming (e.g., standard scores and percentiles) is commonly used to compare scores on different measures because age-normed scores have a similar mathematical metric. Standard scores (e.g., *t*- or *z*-scores) have a fixed mean and standard deviation. Percentiles have a fixed range (0–100). Age-norming can be useful for examining individuals' *relative* change (i.e., change relative to other individuals in the sample or relative to a norm-referenced sample). However, because age-normed scores have a fixed scale, they cannot detect *absolute* change (i.e., change in an individual's trait level or the group mean or variability over time). Standardizing scores with a fixed range or mean and standard deviation does not ensure the scores are on the same metric, so age-norming is generally inadvisable when examining development (Moeller, 2015).

Average or percentage scores

Another approach to comparing scores across different measures is an average score or percentage score that accounts for the different number of items in each measure. A major assumption of average and percentage scores is that the items on the different measures do not differ in discrimination or severity (defined in the next paragraph). However, it is unlikely that measures with different items will have the same severity, especially when the item content differs across the two measures. Thus, average or percentage scores are not advisable in most contexts dealing with different measures over time. To compare scores on different measures over time, researchers recommend vertical scaling (Kolen & Brennan, 2014).

Vertical scaling

In vertical scaling, measures that assess a similar construct but differ in difficulty or severity are placed on the same scale. Vertical scaling is widely used in educational testing because the same test items tend to become easier relative to a given level of ability as children get older. Multiple approaches exist for vertical scaling. For the present study, we used the IRT approach to vertical scaling (Kolen & Brennan, 2014). We fit IRT models that estimate two properties of each item: (1) discrimination and (2) difficulty (severity). An item's discrimination parameter describes how well the item distinguishes between low and high levels of the trait. For example, an item asking how often a person feels depressed will have a higher discrimination for internalizing problems compared to an item asking how often the person reads. An item's difficulty parameter describes the trait level (level on the latent criterion) at which the probability of endorsing the item is 50%. In the context of psychopathology, a higher difficulty parameter reflects a higher, more severe level of internalizing problems, so henceforth we refer to the difficulty parameter as severity. For example, an item asking how often a person thinks about suicide will have a higher severity compared to an item asking how often the person feels sad. Based on the items' parameters and participants' responses on the items, IRT estimates each person's latent trait level of internalizing problems (i.e., ability score or theta).

When the different measures have common items over time, the IRT approach to vertical scaling uses common items administered across ages to link the measures on the same scale by finding scaling parameters that put the trait level scores on the same metric. The scaling parameters are determined as the linear transformation (i.e., the intercept and slope parameter) that, when applied to the second measure, minimizes the differences between the probability of an individual endorsing the common items across the two measures. 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.

A number of studies have used vertical scaling in the fields of education and cognitive testing to measure growth with changing measures over time. As one prime example, McArdle and colleagues (2009) examined the development of cognitive ability from 2 to 72 years of age.

Limitations of Previous Research

Despite the numerous studies using vertical scaling in education and cognitive testing, to our knowledge, no studies have examined the development of psychopathology or social development more generally using vertical scaling. Moreover, despite researchers acknowledging the importance of examining the heterotypic continuity of internalizing problems (Sterba, Prinstein, & Cox, 2007), to our knowledge, no studies have examined trajectories of internalizing problems with changing measures to account for heterotypic continuity, maintain construct validity, and examine mean-level change over a lengthy span of development.

To our knowledge, the only studies examining trajectories of *broadband* internalizing problems with changing measures come from the Australian Temperament Project, which examined trajectory classes from ages 3–15 (and anxiety and depression from ages 11–27; Betts et al., 2016; Letcher, Sanson, Smart, & Toumbourou, 2012; Letcher, Smart, Sanson, & Toumbourou, 2009; Toumbourou, Williams, Letcher, Sanson, & Smart, 2011). The studies did not link the different measures or account for changes in the measures' scales, however, so they did not allow interpretation of mean-level change across measurement changes.

The challenge of heterotypic continuity has led researchers to frequently grapple with the issue of developmental equivalence or to avoid examining the development of internalizing problems across lengthy spans. Many studies have used all possible items or only the common items to maintain the same measure over time. Regarding all possible items, we have seen many studies of internalizing problems that have used measures outside the ages they were originally designed to assess (Adkins et al., 2009; Broeren, Muris, Diamantopoulou, & Baker, 2013; Côté et al., 2009; Crocetti, Klimstra, Keijsers, Hale, & Meeus, 2009; Hale III, Raaijmakers, Muris, van Hoof, & Meeus, 2008; Leadbeater, Thompson, & Gruppuso, 2012; Mathiesen, Sanson, Stoolmiller, & Karevold, 2009; Meadows, Brown, & Elder, 2006; Miers, Blöte, de Rooij, Bokhorst, & Westenberg, 2013; Morin et al., 2011). We have also seen studies of internalizing problems that used only common items over time (Fanti & Henrich, 2010; Gilliom & Shaw, 2004; Sterba et al., 2007).

Thus, studies frequently deal with the issue of developmental equivalence and, in many cases, resort to using a measure at an age outside the age range of validation or to discarding relevant items. An important advance for the field is learning how to handle changes in measurement, because ignoring the heterotypic continuity of internalizing problems over lengthy spans of development likely results in measures that violate construct validity (if summing all possible items) or content validity (if using only common items). Moreover, it also allows measuring internalizing problems in age-appropriate ways, for the sake of understanding development across important developmental transitions.

To ignore heterotypic continuity by using only common items and discarding items reflecting the age-specific manifestations of internalizing problems (e.g., somatic complaints) results in a measure that captures the development of specific problems (i.e., the common items) without capturing the development of the *construct* of internalizing, and can result in inaccurate trajectories. Chen and Jaffee (2015) found that the common items failed to detect the adolescent-onset of externalizing problems observed in a subgroup when using age-relevant items. The challenge of heterotypic continuity may account for why we have not seen studies examining trajectories of broadband internalizing problems across the transition from adolescence into adulthood. We approached this problem by comparing different approaches to vertical scaling. Vertical scaling approaches are widely used in other fields to examine change with different measures over time (Kolen & Brennan, 2014), so it seemed plausible that vertical scaling would be a useful approach to account for heterotypic continuity in developmental psychology.

The Present Study

We examined the development of internalizing problems over a decade of life, and used vertical scaling with different measures over time because internalizing problems demonstrate heterotypic continuity. After re-scaling the different measures of internalizing problems to be on the same scale to account for heterotypic continuity, we examined growth curves of internalizing problems and whether the trajectories differed by sex or ethnicity.

Method

Participants

Children (*N* = 585) were recruited for the Child Development Project (Dodge, Bates, & Pettit, 1990) from two cohorts in 1987 and 1988 from schools at three sites: Nashville, TN; Knoxville, TN; and Bloomington, IN. The schools and the sample represented families with a broad range of socioeconomic status (SES), representative of the populations at the respective sites. The Hollingshead index of SES (*M* = 39.53, *SD* = 14.01, range: 8 to 66, stratum 1: 17% of the sample, 2: 33%, 3: 25%, 4: 16%, 5: 9%) reflected a broad range for the original sample, which was 52% male, 81% European American, 17% African American, and 2% of “other” ethnicity. Over the course of the project, the Child Development Project protocols have been approved by Institutional Review Boards (IRBs) at Indiana University, Vanderbilt University, the University of Tennessee, Auburn University, and Duke University. The current protocol, “How Chronic Conduct Problems Develop” (protocol number 40) is approved by the Duke University IRB.

Children were followed annually with parents', teachers', peers', and/or self-report ratings of the children's internalizing problems. The present study focuses on self-report ratings of adolescents' and young adults' internalizing problems from 14 to 24 years of age. We focused on self-reports because of the accuracy of adolescents' reports of their own internalizing problems—adolescents are in a unique position to report on their subjective experiences of internalizing problems (De Los Reyes & Kazdin, 2005).

Measures

Internalizing problems—Adolescents rated their level of internalizing problems annually on the Internalizing Scale of the YSR (Achenbach, 1991) from ages 14 to 19 (except age 18). From ages 20 to 24, they rated their internalizing problems annually on the Internalizing Scale of the YASR (Achenbach, 1997). Adolescents rated internalizing problems on the YSR and YASR as “not true,” “somewhat or sometimes true,” or “very true or often true,” scored 0, 1, and 2, respectively (although vertical scaling does not require scores on the same response scale). Scores on the Internalizing scale were summed across items. Internal consistency of items ranged from α = .89 to .91, depending on the year. The Achenbach scales have strong validity, including content, construct, and criterion-related validity (Sattler & Hoge, 2006).

Items on the Internalizing scale differed somewhat between the YSR (31 items) and YASR (23 items) in ways that reflected the heterotypic continuity of internalizing problems. For instance, somatic complaints were included in the measure of adolescents' internalizing problems on the YSR, but they were not included in the measure of young adults' internalizing problems on the YASR. The YSR Internalizing scale included Withdrawn, Somatic Complaints, and Anxious/Depressed subscales, whereas the YASR Internalizing scale included Withdrawn and Anxious/Depressed subscales. The Internalizing scale on the YSR and YASR shared 17 common items, while 14 of the items on the Internalizing scale were unique to the YSR and 6 were unique to the YASR.[1] Descriptive statistics and a Pearson correlation matrix of the raw Internalizing sum scores at each age are in Table 1. Possible scores ranged from 0–62 on the YSR, 0–46 on the YASR, and 0–34 on the 17 common items of the YASR and YSR, with higher scores reflecting higher levels of internalizing problems.

We also examined the association of scores on the Internalizing scale with scores on the Internalizing and Externalizing scales one year later. Internalizing problems showed strong convergent and discriminant validity. The average correlation of internalizing problems with later internalizing problems was $r = .72$ (95% CI: .67–.76). The average correlation of internalizing problems with later externalizing problems was $r = .41$ (95% CI: .33–.49).

Statistical Analysis

In the present study, we used the IRT approach to vertical scaling (as described in Kolen & Brennan, 2014) to transform scores on the YASR to the scale of the YSR. In the context of vertical scaling, IRT estimates people's latent trait scores of internalizing problems over time (i.e., a latent variable or "true score" approach to vertical scaling). As further validation of the findings from the IRT approach to vertical scaling, we also conducted an alternative approach to vertical scaling, known as Thurstone scaling (as described in Kolen & Brennan, 2014). Unlike IRT, the Thurstone scaling approach to vertical scaling retains the raw metric (i.e., an observed score or "raw score" approach to vertical scaling). Our findings from the Thurstone scaling approach are available in Supplementary Appendix S3.

IRT models—Internalizing problems were analyzed with graded response models in IRT using the mirt package (Chalmers, 2012) in R. The mirt package uses an expectation-maximization algorithm known as marginal maximum likelihood, which uses all available data and provides valid inferences when data are missing at random or completely at random. Graded response models allow polytomous variables with more than two response categories (e.g., 0–2 Likert scale in the present study). The models estimated three parameters for each item: (1) discrimination (a), (2) severity for the threshold from 0–1 (b_1), and severity for the threshold from 1–2 (b_2). We examined model fit with RMSEA and CFI. We fit a separate IRT model at each age for the purposes of linking the measures across time, rather than fitting all items in the same model (i.e., concurrent calibration). Although concurrent calibration procedures tend to have greater precision of item parameter estimates,

[1] The YASR item reflecting whether the adult was concerned about his or her looks was not administered at each age the YASR was administered, so it was not included in our calculations of the Internalizing scale.

separate estimation is considered safer over lengthy developmental spans because the uni-dimensionality assumption of IRT is more likely to be violated in concurrent calibration (Kolen & Brennan, 2014).

Vertical scaling—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 IRT approach to vertical scaling to transform scores on the YASR to the scale of the YSR. The YSR and YASR have different but overlapping item content, so we needed to put them on the same scale. We applied vertical scaling that scales the scores across the different measures using the items that are in common across both measures (i.e., a common-item non-equivalent group or anchor instrument design, see Figure 1). We applied the following steps for vertical 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). Participants' mean scores on the common items are in Table 2 and are depicted in Panel A of Figure 2.

2. As described earlier, we fit separate IRT models at each age.

3. We used vertical scaling procedures to calculate scaling parameters that linked the IRT factor scores (trait level scores or theta) from the YSR and YASR at different ages on the same scale. Vertical scaling uses common items administered across ages to link the measures on the same scale by finding scaling parameters that put the trait level scores on the same metric. We used the plink package (Weeks, 2010) in R to calculate Stocking-Lord scaling parameters. To link the measures, scaling parameters were calculated using an iterative algorithm that minimizes the sum of squared differences between the expected aggregate scores for the common items for each measure. Thus, the scaling parameters minimize the differences between the probability of an individual endorsing the common items across the two measures (or ages).

To calculate scaling parameters, we first set the target scale to be the YSR at age 14, and calculated scaling parameters at age 15 to link the YSR scores at age 15 to be on the same scale as the YSR at age 14. We then applied a process of linking and chaining (Kolen & Brennan, 2014) to calculate scaling parameters to link the remaining scores to the YSR metric at age 14. To do so, we repeated steps 1–3 by (a) linking the scores at age 16 to the newly scaled scores at age 15, (b) linking scores at age 17 to the newly scaled scores at age 16, and (c) etc., until scores at all ages, including the YASR scores, had been linked to the target YSR scale at age 14. The scaling parameters include an intercept parameter, B, and a slope parameter, A, that link the trait level scores at one age to the trait level scores at the prior age, by linking the discrimination and severity parameters at the two ages using the following formulas (Kolen & Brennan, 2014):

$$a(\text{age}_i)=\frac{a(\text{age}_j)}{A} \quad (1)$$

$$b(\text{age}_i)=A \times b(\text{age}_j)+B \quad (2)$$

where $a(\text{age}_i)$ and $a(\text{age}_j)$ represent the discrimination parameter for the common items at age i and age j, respectively; $b(\text{age}_i)$ and $b(\text{age}_j)$ represent the severity parameter for the common items at age i and age j, respectively; A represents the slope scaling parameter, and B represents the intercept scaling parameter.

1. We then used the scaling parameters to calculate the trait level scores at each age on the same scale. We used expected a posteriori (EAP) factor scores as individuals' trait level scores of internalizing problems. The A and B transformation constants rescale the standard deviation and mean, respectively, of the trait level scores to put the measures on a comparable scale, while still retaining changes in means and variances over time (based on the changes in means and variances of the common items). The vertically scaled scores were calculated by the following formula (Kolen & Brennan, 2014):

$$\theta(\text{age}_{14})=A \times \theta(\text{age}_j)+B \quad (3)$$

where $\theta(\text{age}_{14})$ represents the vector of trait level scores (i.e., factor scores) on the metric of the YSR at age 14, and $\theta(\text{age}_j)$ represents the vector of trait level scores on the YSR or YASR at the remaining ages. When linking and chaining were completed, all scores were placed on the YSR age 14 metric.

Growth curve model—After vertically scaling the scores of internalizing problems to be on the same scale, we then examined individuals' trajectories of internalizing problems. To examine individuals' growth curves of vertically scaled internalizing problems, we used the lme function of the nlme package (Pinheiro, Bates, DebRoy, & Sarkar, 2009) in R for hierarchical linear modeling (HLM).[2] HLM can handle missingness and unbalanced data (Singer & Willett, 2003). We compared linear and curvilinear (polynomial) forms of growth using nested model comparisons with likelihood ratio tests. After settling on a form of growth, we examined sex and ethnicity as predictors of individuals' trajectories of internalizing problems.

[2] Although we considered multiple imputation approaches to handle missingness, to fairly compare the approach of using the common items to using the rescaled scores, we used only the observed data.

Results

IRT Models

After determining that we approximately met IRT assumptions (Appendix S1) and observed only modest differential item functioning (DIF; Appendix S2), we fit a separate IRT graded response model at each age using the self-reported questionnaire items of internalizing problems. RMSEA estimates ranged from .058 to .072, depending on the year. CFI estimates ranged from .92 to .97, depending on the year. Thus, our model fit was adequate to good.

Linking the YASR (and YSR) Scores to the Scale of the YSR at Age 14

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 scores at all ages to the scale of the YSR at age 14 (see steps 1–4 from the Vertical Scaling section of the Statistical Analysis section of the Method section). First, we examined scores on the 17 common items (i.e., the items that were common to both the YSR and YASR), see Table 2 and Panel A of Figure 2. Second, we fit separate IRT models at each age (see previous section).

Third, we used vertical scaling to put the IRT scores on the same scale. We calculated linear scaling parameters (slope: A; intercept: B) that linked the IRT scores at each age to the scores at the preceding age. The linear scaling parameters are in Table 2.[3]

Fourth, we used the scaling parameters to calculate individuals' internalizing problem scores on the same scale as the YSR at age 14. Age 15 scores were rescaled to the target scale of the age 14 scores by multiplying the age 15 scores by 1.091 and adding 0.017. We then applied linking and chaining to link the remaining scores to the YSR metric at age 14. To do so, we repeated steps 1-4 by linking the scores at age 16 to the newly scaled scores at age 15, linking scores at age 17 to the newly scaled scores at age 16, etc. For instance, age 16 scores were rescaled to the target scale at age 14 by the scaling parameters at age 16 (to transfer the scores to the age 15 metric) and then by the scaling parameters at age 15 (to transfer the scores to the age 14 metric). We applied this process of linking and chaining until all scores, including YASR scores, had been rescaled to the metric of the YSR at age 14.[4]

The mean and standard deviation of the vertically scaled scores are in Table 2. Participants' mean internalizing problem scores, after rescaling the YASR scores to be on the same metric as the YSR, are depicted in Panel B of Figure 2. Notably, the scores retained a highly similar pattern of mean scores by age when examining the re-scaled total scores compared to when examining just the common items (see Panel A of Figure 2). Thus, the IRT approach to vertical scaling successfully retained mean-level change when re-scaling the YASR scores to

[3] Scaling parameters where A equals 1 and B equals 0 would represent no adjustment, so greater deviations from those values reflect greater adjustment to put the scores on the same scale. Notably, all of the scaling parameters for scores at adjacent ages were relatively close to these values (A ≈ 1 and B ≈ 0), indicating that only small adjustments were necessary to link the scores at adjacent ages.

[4] IRT factor scores have a mean of 0 and a standard deviation of 1, so the IRT-based internalizing problem scores at age 14 have a relatively normal distribution with a mean of 0 and standard deviation of 1. Scores at subsequent ages were linked to the target scale at age 14, so deviations from a mean of 0 and a standard deviation of 1 reflect changes in means and variances over time. For an example calculation of linking and chaining, see the note of Table 2.

be on the same metric as the YSR while still using a more comparable scale. Moreover, vertically scaled scores from IRT were highly correlated with vertically scaled scores from Thurstone scaling ($r = .95-.97$, depending on the year).

Growth Curve Model

To examine growth curves, we first compared a linear growth curve model to polynomial forms of change in HLM to identify the best-fitting form of change for the rescaled internalizing problem scores. A model with random linear slopes fit better than a model with a linear slope component that was fixed across individuals (i.e., fixed linear slopes; $\chi^2[2] = 486.49$, $p < .001$). A model with a random linear slope component and a fixed quadratic component fit better than a model with only random linear slopes ($\chi^2[1] = 24.25$, $p < .001$). A model with a random linear slope component and a random quadratic component fit better than a model with a random linear slope component and a fixed quadratic component ($\chi^2[3] = 64.37$, $p < .001$), and was the best fitting model (model fit did not significantly improve when adding a fixed cubic component: $\chi^2[1] = 2.25$, $p = .133$). Thus, a quadratic form of change was the best-fitting form of change for the rescaled internalizing problem scores. Individuals' quadratic trajectories, and the average quadratic trajectory for males and females are depicted in Figure 3. The average quadratic 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 a significant negative mean of the slopes ($B = -0.01$, $\beta = -0.04$, $t[3980] = -2.21$, $p = .027$). In a similar growth curve model examining the trajectories of scores on the common items, however, the mean of the slopes was not significant ($B = -0.03$, $\beta = -0.02$, $t[3980] = -1.08$, $p = .282$).

We conducted sensitivity analyses to determine the sensitivity of our findings to other vertical scaling approaches. The factor scores from a partially constrained multiple group (concurrent calibration) IRT model (within-item parameter constraints across time for non-DIF parameters) showed similar evidence of a negative mean of the slopes ($B = -0.009$, $\beta = -0.03$, $t[3980] = -2.00$, $p = .045$). There was also a negative mean of the slopes using factor scores from a comparable IRT model that excluded the items showing DIF ($B = -0.01$, $\beta = -0.04$, $t[3980] = -2.16$, $p = .031$). The Thurstone approach to vertical scaling also showed evidence of a negative mean of the slopes (*B* = -0.08, β = -0.03, *t*[3980] = -1.94, *p* = .053), at a trend-level.

In addition to differences in the form of change for the age-relevant items versus common items at the *group-level*, there were also 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* = 3.36) than participants who did not (*M* = 1.81; *t*[29.07] = -3.69, *p* < .001). 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* = 0.66) than participants who did not (*M* = 1.95; *t*[31.93] = 6.11, *p* < .001).

We then examined sex and ethnicity as predictors of the intercepts and linear slopes of the rescaled internalizing problem scores (see Table 3). The mean of the linear slopes was not significant when controlling for the other model predictors. Females showed higher intercepts than males, and showed a trend toward greater decreases over time compared to males. Although African Americans showed a trend toward lower intercepts, African Americans and those of "other" ethnicity did not significantly differ from European Americans in their intercepts or linear slopes.

Discussion

Heterotypic continuity, the change in the manifestation of a construct or process over time, presents challenges to studying individuals over lengthy spans of development, and may necessitate using different measures over time. We examined self-reports of internalizing problems on the YSR from ages 14 to 19 and on the YASR from ages 20 to 24. The YSR Internalizing scale includes items reflecting anxiety, depression, and somatic complaints, whereas its YASR counterpart includes items reflecting anxiety and depression but not somatic complaints. The challenge is measuring actual change rather than change in the meaning of the measures. We applied a vertical scaling technique to account for the heterotypic continuity of internalizing problems and the change in measurement.

Applying vertical scaling with age-relevant items to account for the heterotypic continuity of internalizing problems, we observed a pattern of means by age at the group-level that was similar to what we would have observed had we used the common items (see Figure 2), but with important differences. The age-relevant items showed a *group-level* pattern of means by age similar to the results with the common items, but the age-relevant items resulted in more construct-valid scores of internalizing problems at the *individual-level.* The age trends of the mean values of the *observed* scores differed from the means of individuals' slopes based on *model-fitted* values in growth curve models, which fit lines through all available time points, and essentially interpolate missing values based on the individual's other time points and the other individuals' trajectories (i.e., shrinkage). We found that vertical scaling made small adjustments to the scores (see Table 2), but these subtle adjustments resulted in potentially meaningful differences in the individuals' and group trajectories. Because the common items ignored the age-specific manifestations of internalizing problems, e.g., somatic complaints, some participants showed *decreases* in internalizing problems when we used their ratings on the age-relevant items but *increases* when we used only the common items, and other participants showed the opposite pattern. The differences in individuals' trajectories using the age-relevant items versus the common items could explain differences we observed in the group-level trajectories using the age-relevant items versus the common items. Although prior research is mixed, some studies have shown decreases in the prevalence of internalizing disorders from adolescence to adulthood (Costello, Copeland, & Angold, 2011). We observed group-level decreases in internalizing problems from adolescence to early adulthood using the construct-valid items. Using only the common items, however, we observed no significant change in internalizing problems over time. Discarding items (e.g., somatic complaints) that were relevant to internalizing problems during some developmental periods but not other developmental periods (i.e., using only the common items) resulted in a loss of information that may have made the measure less sensitive to developmental change. Thus, accounting for heterotypic continuity could have theoretical and practical advantages over ignoring heterotypic continuity by using only the common items across time. Future research should further examine the potential reasons why the approaches may differ in their developmental inferences.

Accounting for heterotypic continuity allowed us to examine predictors of individuals' trajectories over a lengthy developmental span. We observed that females showed higher levels of internalizing problems than males at age 14, and there was a trend toward females showing greater decreases over time compared to males. As shown in Figure 2, we found the greatest difference between females' and males' levels of internalizing problems around ages 15–18, which is consistent with Hankin et al. (1998), and the greatest level of internalizing problems around age 15, consistent with Adkins et al. (2009). We also found a trend toward lower levels of internalizing problems among African Americans compared to European Americans.

Despite evidence of several items showing modest DIF over time, the overall theoretical and empirical evidence suggests we measured the same construct in an equivalent way across time. Although longitudinal measurement invariance should be tested, establishing strict longitudinal measurement invariance is unnecessary in the case of heterotypic continuity because the meaning of the measures is expected to change with changes in the manifestation of the construct (Petersen et al., 2016). Research has demonstrated that models with failed longitudinal measurement invariance can yield valid inferences in the context of heterotypic continuity (Edwards & Wirth, 2012). Removing items/measures that show DIF or failed measurement invariance over time is not necessarily recommended in the case of heterotypic continuity (Knight & Zerr, 2010). Removing items or measures can result in a less representative sample of the content of the construct (i.e., lower content validity), and some items might be expected to change in their discrimination or severity over time given heterotypic continuity, and yet remain construct-valid. Discarding them would be removing important and meaningful developmental information about the construct. Discarding construct-valid items showing DIF or failed measurement invariance would be akin to using only the common items, which we argue is highly problematic (and violates content validity). Nevertheless, we observed similar results when we excluded DIF items, suggesting that DIF did not compromise the findings.

In addition to empirical considerations, there are important theoretical considerations regarding whether one is measuring the same construct across time in an equivalent way (construct validity invariance). First, the Achenbach scales are widely used measures of internalizing problems; they were derived empirically, and have strong validity, including content validity, construct validity, and criterion-related validity (Sattler & Hoge, 2006). Second, the items were selected based on theory and on the known heterotypic continuity of internalizing problems—we used the age-relevant items instead of discarding construct-relevant items that were not present in both forms of the measure. Third, we observed strong internal consistency of the items at each age, and the items showed strong cross-time continuity (see Table 1). Fourth, the items showed convergent and discriminant validity with respect to externalizing problems. Fifth, the trajectories showed construct validity: their pattern was similar with prior findings. Finally, we observed similar trajectories with multiple approaches to vertical scaling, including separate IRT estimation with linking, concurrent calibration in IRT, and Thurstone scaling. Thus, we feel there is strong theoretical and empirical evidence for using the Internalizing scales of the YSR/YASR as they are constructed for measuring the same construct of internalizing problems in an equivalent way over time in the present study.

Alternative Approaches to Vertical Scaling

We applied the widely used IRT approach to vertical scaling, which uses a latent variable approach. There are alternative approaches to vertical scaling. Thurstone scaling, an observed score approach, may be more practical than IRT in some situations for vertical scaling. First, IRT requires large sample sizes for accurate estimation. Second, IRT generally requires dichotomous, polytomous, or categorical items instead of continuous measures (unless moving to a SEM framework). Third, except for advanced and cutting-edge multi-dimensional IRT techniques, most IRT applications require items that are uni-dimensional. These requirements pose challenges for psychological constructs, which are often multi-faceted and measured using various metrics. Nevertheless, (uni-dimensional) IRT is often employed for vertical scaling, and the findings are often consistent with Thurstone Scaling (Becker & Forsyth, 1992), as they were in the present study.

Implications for Developmental Psychology

We are unaware of other studies that have examined individuals' trajectories of broadband internalizing problems from adolescence to adulthood. The present results show how internalizing problems developed across an important developmental transition. Broadband internalizing problems peaked in mid-to-late adolescence and decreased into adulthood, similar to patterns shown for depression (Adkins et al., 2009). Further, the decreases in internalizing problems were detected after we accounted for their heterotypic continuity using vertical scaling. The findings of the present study are novel, but the statistical approach is not. Previous research has (a) used vertical scaling to link different measures on the same comparable scale (Kolen & Brennan, 2014), (b) measured change with different measures (McArdle et al., 2009), and (c) used changing items to account for the heterotypic continuity of psychopathology based on theory (Petersen, Bates, Dodge, Lansford, & Pettit, 2015). What is especially novel in the present study is the assembling of these techniques to demonstrate how to use vertical scaling and changing measures to account for heterotypic continuity and measure individuals' change in constructs showing heterotypic continuity. We feel this is a crucial theoretical and empirical advance, especially for developmental theory. The vast majority of research in developmental psychology has examined trajectories using the same measures over time, which is a common practice with some advantages for model building, but which, as we argue next is often highly problematic for developmental theory.

When developmental psychologists have examined individuals' change in a construct using the same measures over time, in the traditional way, using either all available items or only age-common items, this creates a theoretical and empirical problem when the construct shows heterotypic continuity, i.e., change in its manifestation over time. Using all available items over time violates construct validity because, to one degree or another, the same items do not consistently reflect the same construct over time. Using only age-common items violates content validity because the measure is not assessing the construct as a whole, including its age-specific manifestations. Moreover, not only are there theoretical reasons to use different measures over time in the context of heterotypic continuity, there are likely empirical advantages of using different measures over time, as well. We showed that using different measures (i.e., all construct-valid items) over time may be more sensitive to developmental change than using only age-common items.

Strengths and Limitations

The present study had key strengths. First, it examined the development of broadband internalizing problems over a lengthy span of development across the important developmental transition from adolescence to adulthood. Second, it accounted for the heterotypic continuity of internalizing problems when examining individuals' trajectories, and compared the approach to traditional approaches that ignore heterotypic continuity. Third, it examined the form of change of internalizing problems and sex and ethnicity as predictors of the trajectories. Fourth, it considered multiple approaches to vertical scaling, each with different assumptions, and found substantially similar results with each method, providing greater confidence in the findings.

The study also had limitations. We did not examine trajectories of individual items or sub-dimensions of internalizing problems (e.g., anxiety or depression). One can always reduce to a lower level subunit, however. Internalizing problems have an empirically-derived factor structure, so we believe there is theoretical reason for this level of analysis. In addition, the sub-dimensions of anxiety and depression, themselves, like most behavior trait measures, are heterogeneous and involve behaviors whose meaning would depend on age, so they would likely demonstrate heterotypic continuity, as well.

Conclusion

The present study applied vertical scaling to account for the heterotypic continuity of internalizing problems from adolescence to adulthood. Vertical scaling allowed us to place scores from two measures on the same scale. Accounting for heterotypic continuity by using all developmentally relevant items may have been more sensitive to developmental change in internalizing problems than was ignoring heterotypic continuity by using the same items across major stages of development. Using vertical scaling, internalizing problems peaked in mid-to-late adolescence and decreased into adulthood. Vertical scaling may be a useful approach to measuring individuals' developmental trajectories in constructs that change in their manifestation over time.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge Olga Berkout and Rachel Salk for their helpful feedback on a draft of this manuscript. The Child Development Project has been funded by grants MH42498, MH56961, MH57024, and MH57095 from the National Institute of Mental Health; HD30572 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development; and DA016903 from the National Institute on Drug Abuse. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Institutes of Health.

References

Achenbach, TM. Manual for the Youth Self-Report and 1991 profile. Burlington, VT: University of Vermont, Department of Psychiatry; 1991.

Achenbach, TM. Manual for the Young Adult Self-Report and Young Adult Behavior Checklist. Burlington, VT: University of Vermont, Department of Psychiatry; 1997.

Achenbach TM, Edelbrock CS. The classification of child psychopathology: A review and analysis of empirical efforts. Psychological Bulletin. 1978; 85:1275–1301. DOI: 10.1037/0033-2909.85.6.1275 [PubMed: 366649]

Adkins DE, Wang V, Dupre ME, van den Oord JCG, Elder GH. Structure and stress: Trajectories of depressive symptoms across adolescence and young adulthood. Social Forces. 2009; 88:31–60. DOI: 10.1353/sof.0.0238

Avenevoli, S., Steinberg, L. The continuity of depression across the adolescent transition. In: Hayne, WR., Robert, K., editors. Advances in child development and behavior. Vol. 28. San Diego, CA: Academic Press; 2002. p. 139-173.

Becker DF, Forsyth RA. An empirical investigation of Thurstone and IRT methods of scaling achievement tests. Journal of Educational Measurement. 1992; 29:341–354. DOI: 10.1111/j.1745-3984.1992.tb00382.x

Betts KS, Baker P, Alati R, McIntosh JE, Macdonald JA, Letcher P, Olsson CA. The natural history of internalizing behaviours from adolescence to emerging adulthood: findings from the Australian Temperament Project. Psychological Medicine. 2016; 46:2815–2827. DOI: 10.1017/S0033291716001495 [PubMed: 27439384]

Broeren S, Muris P, Diamantopoulou S, Baker JR. The course of childhood anxiety symptoms: Developmental trajectories and child-related factors in normal children. Journal of Abnormal Child Psychology. 2013; 41:81–95. DOI: 10.1007/s10802-012-9669-9 [PubMed: 22836287]

Chalmers RP. mirt: a multidimensional item response theory package for the R environment. Journal of Statistical Software. 2012; 48:1–29.

Chen FR, Jaffee SR. The heterogeneity in the development of homotypic and heterotypic antisocial behavior. Journal of Developmental and Life-Course Criminology. 2015; 1:269–288. DOI: 10.1007/s40865-015-0012-3

Chen WH, Thissen D. Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics. 1997; 22:265–289. DOI: 10.3102/10769986022003265

Costello EJ, Copeland W, Angold A. Trends in psychopathology across the adolescent years: What changes when children become adolescents, and when adolescents become adults? Journal of Child Psychology and Psychiatry. 2011; 52:1015–1025. DOI: 10.1111/j.1469-7610.2011.02446.x [PubMed: 21815892]

Côté SM, Boivin M, Liu X, Nagin DS, Zoccolillo M, Tremblay RE. Depression and anxiety symptoms: onset, developmental course and risk factors during early childhood. Journal Of Child Psychology And Psychiatry, And Allied Disciplines. 2009; 50:1201–1208. DOI: 10.1111/j. 1469-7610.2009.02099.x

Crocetti E, Klimstra T, Keijsers L, Hale WW, Meeus W. Anxiety trajectories and identity development in adolescence: A five-wave longitudinal study. Journal of Youth and Adolescence. 2009; 38:839–849. DOI: 10.1007/s10964-008-9302-y [PubMed: 19636785]

De Los Reyes A, Kazdin AE. Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. Psychological Bulletin. 2005; 131:483–509. DOI: 10.1037/0033-2909.131.4.483 [PubMed: 16060799]

Dodge KA, Bates JE, Pettit GS. Mechanisms in the cycle of violence. Science. 1990; 250:1678–1683. DOI: 10.1126/science.2270481 [PubMed: 2270481]

Eaton NR, Krueger RF, Markon KE, Keyes KM, Skodol AE, Wall M, et al. Grant BF. The structure and predictive validity of the internalizing disorders. Journal of Abnormal Psychology. 2013; 122:86–92. DOI: 10.1037/a0029598 [PubMed: 22905862]

Edwards, MC., Wirth, RJ. Valid measurement without factorial invariance: A longitudinal example. In: Hancock, GR., Harring, JR., editors. Advances in longitudinal methods in the social and behavioral sciences. Charlotte, NC, US: Information Age Publishing; 2012. p. 289-311.

Fanti KA, Henrich CC. Trajectories of pure and co-occurring internalizing and externalizing problems from age 2 to age 12: Findings from the National Institute of Child Health and Human Development Study of Early Child Care. Developmental Psychology. 2010; 46:1159–1175. DOI: 10.1037/a0020659 [PubMed: 20822230]

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

Gilliom M, Shaw DS. Codevelopment of externalizing and internalizing problems in early childhood. Development and Psychopathology. 2004; 16:313–333. DOI: 10.1017/S0954579404044530 [PubMed: 15487598]

Grimm KJ, Ram N, Hamagami F. Nonlinear growth curves in developmental research. Child Development. 2011; 82:1357–1371. DOI: 10.1111/j.1467-8624.2011.01630.x [PubMed: 21824131]

Hale WW III, Raaijmakers Q, Muris P, van Hoof A, Meeus W. Developmental trajectories of adolescent anxiety disorder symptoms: A 5-year prospective community study. Journal of the American Academy of Child & Adolescent Psychiatry. 2008; 47:556–564. DOI: 10.1097/CHI. 0b013e3181676583 [PubMed: 18356762]

Hankin BL, Abramson LY, Moffitt TE, Silva PA, McGee R, Angell KE. Development of depression from preadolescence to young adulthood: Emerging gender differences in a 10-year longitudinal study. Journal of Abnormal Psychology. 1998; 107:128–140. DOI: 10.1037/0021-843x.107.1.128 [PubMed: 9505045]

Knight GP, Zerr AA. Informed theory and measurement equivalence in child development research. Child Development Perspectives. 2010; 4:25–30. DOI: 10.1111/j.1750-8606.2009.00112.x

Kolen, MJ., Brennan, RL. Test equating, scaling, and linking: Methods and practices. 3rd. New York, NY, US: Springer; 2014.

Leadbeater B, Thompson K, Gruppuso V. Co-occurring trajectories of symptoms of anxiety, depression, and oppositional defiance from adolescence to young adulthood. Journal of Clinical Child and Adolescent Psychology. 2012; 41:719–730. DOI: 10.1080/15374416.2012.694608 [PubMed: 22742519]

Letcher P, Sanson A, Smart D, Toumbourou JW. Precursors and correlates of anxiety trajectories from late childhood to late adolescence. Journal of Clinical Child and Adolescent Psychology. 2012; 41:417–432. DOI: 10.1080/15374416.2012.680189 [PubMed: 22551395]

Letcher P, Smart D, Sanson A, Toumbourou JW. Psychosocial precursors and correlates of differing internalizing trajectories from 3 to 15 years. Social Development. 2009; 18:618–646. DOI: 10.1111/j.1467-9507.2008.00500.x

Markon KE, Chmielewski M, Miller CJ. The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. Psychological Bulletin. 2011; 137:856–879. DOI: 10.1037/a0023678 [PubMed: 21574681]

Mathiesen KS, Sanson A, Stoolmiller M, Karevold E. The nature and predictors of undercontrolled and internalizing problem trajectories across early childhood. Journal of Abnormal Child Psychology. 2009; 37:209–222. DOI: 10.1007/s10802-008-9268-y [PubMed: 18766436]

McArdle JJ, Grimm KJ, Hamagami F, Bowles RP, Meredith W. Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. Psychological Methods. 2009; 14:126–149. DOI: 10.1037/a0015857 [PubMed: 19485625]

Meade AW. A taxonomy of effect size measures for the differential functioning of items and scales. Journal of Applied Psychology. 2010; 95:728–743. DOI: 10.1037/a0018966 [PubMed: 20604592]

Meadows SO, Brown JS, Elder GH. Depressive symptoms, stress, and support: Gendered trajectories from adolescence to young adulthood. Journal of Youth and Adolescence. 2006; 35:89–99. DOI: 10.1007/s10964-005-9021-6

Miers AC, Blöte AW, de Rooij M, Bokhorst CL, Westenberg PM. Trajectories of social anxiety during adolescence and relations with cognition, social competence, and temperament. Journal of Abnormal Child Psychology. 2013; 41:97–110. DOI: 10.1007/s10802-012-9651-6 [PubMed: 22723078]

Moeller J. A word on standardization in longitudinal studies: don't. Frontiers in Psychology. 2015; 6doi: 10.3389/fpsyg.2015.01389

Morin AJS, Maïano C, Nagengast B, Marsh HW, Morizot J, Janosz M. General growth mixture analysis of adolescents' developmental trajectories of anxiety: The impact of untested invariance assumptions on substantive interpretations. Structural Equation Modeling: A Multidisciplinary Journal. 2011; 18:613–648. DOI: 10.1080/10705511.2011.607714

Morizot, J., Ainsworth, AT., Reise, SP. Toward modern psychometrics: Application of item response theory models in personality research. In: Robins, RW.Fraley, RC., Krueger, RF., editors. Handbook of research methods in personality psychology. New York, NY, US: Guilford Press; 2007. p. 407-421.

Petersen IT, Bates JE, Dodge KA, Lansford JE, Pettit GS. Describing and predicting developmental profiles of externalizing problems from childhood to adulthood. Development and Psychopathology. 2015; 27:791–818. DOI: 10.1017/S0954579414000789 [PubMed: 25166430]

Petersen IT, Hoyniak CP, McQuillan ME, Bates JE, Staples AD. Measuring the development of inhibitory control: The challenge of heterotypic continuity. Developmental Review. 2016; 40:25–71. DOI: 10.1016/j.dr.2016.02.001 [PubMed: 27346906]

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. the R Core team. nlme: Linear and nonlinear mixed effects models. R package version 3.1-93. 2009. Retrieved from http://cran.r-project.org/web/packages/nlme/index.html

Raju NS. Determining the significance of estimated signed and unsigned areas between two item response functions. Applied Psychological Measurement. 1990; 14:197–207. DOI: 10.1177/014662169001400208

Ryan ND, Puig-Antich J, Ambrosini P, Rabinovich H, Robinson D, Nelson B, et al. Twomey J. The clinical picture of major depression in children and adolescents. Archives of General Psychiatry. 1987; 44:854–861. DOI: 10.1001/archpsyc.1987.01800220016003 [PubMed: 3662742]

Sattler, JM., Hoge, RD. Assessment of children: Behavioral, social, and clinical foundations. 5th. San Diego, CA, US: Jerome M. Sattler, Publisher, Inc; 2006.

Singer, JD., Willett, JB. Applied longitudinal data analysis: Modeling change and event occurrence. New York, NY, US: Oxford University Press, Inc; 2003.

Snyder HR, Young JF, Hankin BL. Strong homotypic continuity in common psychopathology-, internalizing-, and externalizing-specific factors over time in adolescents. Clinical Psychological Science. 2017; 5:98–110. DOI: 10.1177/2167702616651076 [PubMed: 28239532]

Sterba SK, Prinstein MJ, Cox MJ. Trajectories of internalizing problems across childhood: Heterogeneity, external validity, and gender differences. Development and Psychopathology. 2007; 19:345–366. DOI: 10.1017/S0954579407070174 [PubMed: 17459174]

Tomarken AJ, Waller NG. Potential problems with ‘well fitting’ models. Journal of Abnormal Psychology. 2003; 112:578–598. DOI: 10.1037/0021-843X.112.4.578 [PubMed: 14674870]

Toumbourou JW, Williams I, Letcher P, Sanson A, Smart D. Developmental trajectories of internalising behaviour in the prediction of adolescent depressive symptoms. Australian Journal of Psychology. 2011; 63:214–223. DOI: 10.1111/j.1742-9536.2011.00023.x

van der Ark LA. Mokken scale analysis in R 2007. 2007; 20:19.doi: 10.18637/jss.v020.i11

Weeks JP. plink: An R package for linking mixed-format tests using IRT-based methods. 2010; 2010, 35:33.doi: 10.18637/jss.v035.i12

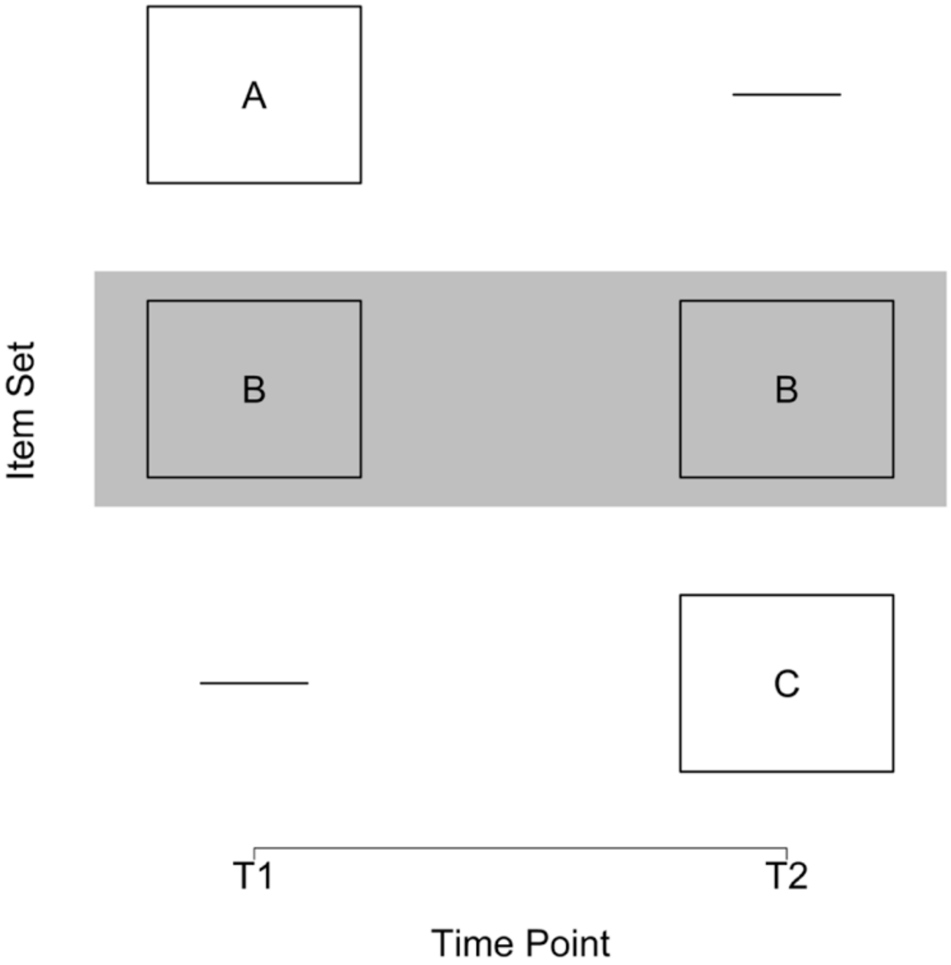


Figure 1.

Depiction of using construct-valid items at each age with a common-item design. Item set A corresponds to items that are construct-valid at only T1. Item set B corresponds to items that are construct-valid at both T1 and T2. Item set C corresponds to items that are construct-valid at only T2. The “common items” (item set B) are highlighted in gray. The present study used the construct-valid items at each age (approach 3: i.e., item sets A and B at T1 and item sets B and C at T2), by using the common items to link the different item sets. Although there were more than two time points (10) in the present study, we used only two different measures (hence we depict the common item-design with T1 and T2 for simplicity).

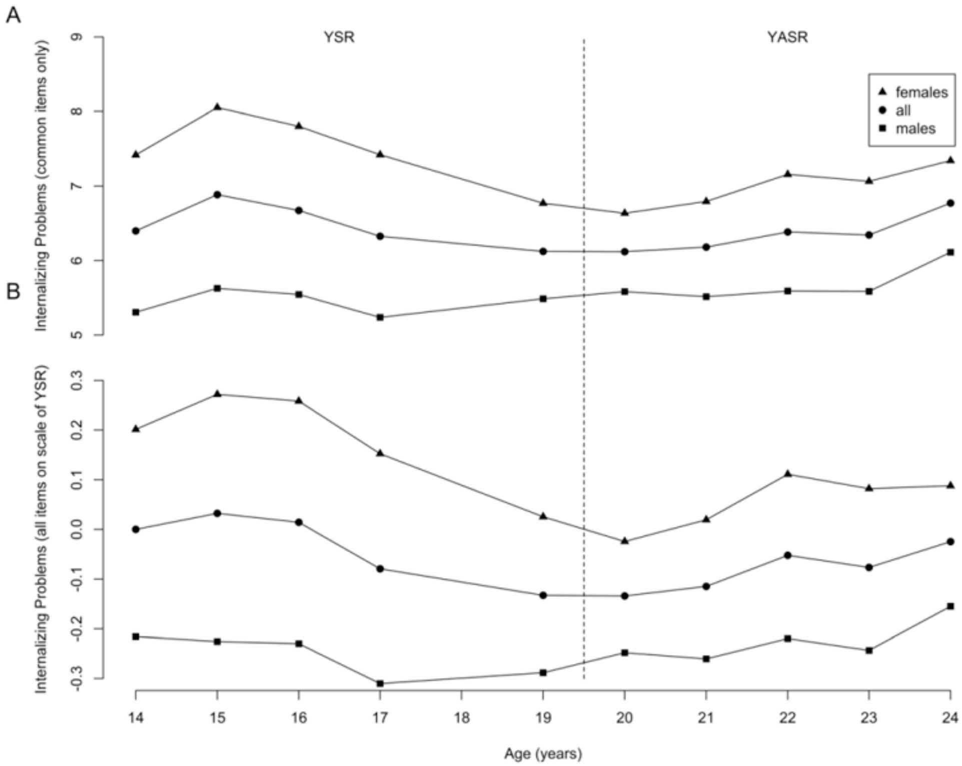


Figure 2.

Panel A depicts participants' mean raw 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 (and YSR) scores to the metric of the YSR (based on the IRT metric of the YSR at age 14). 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.

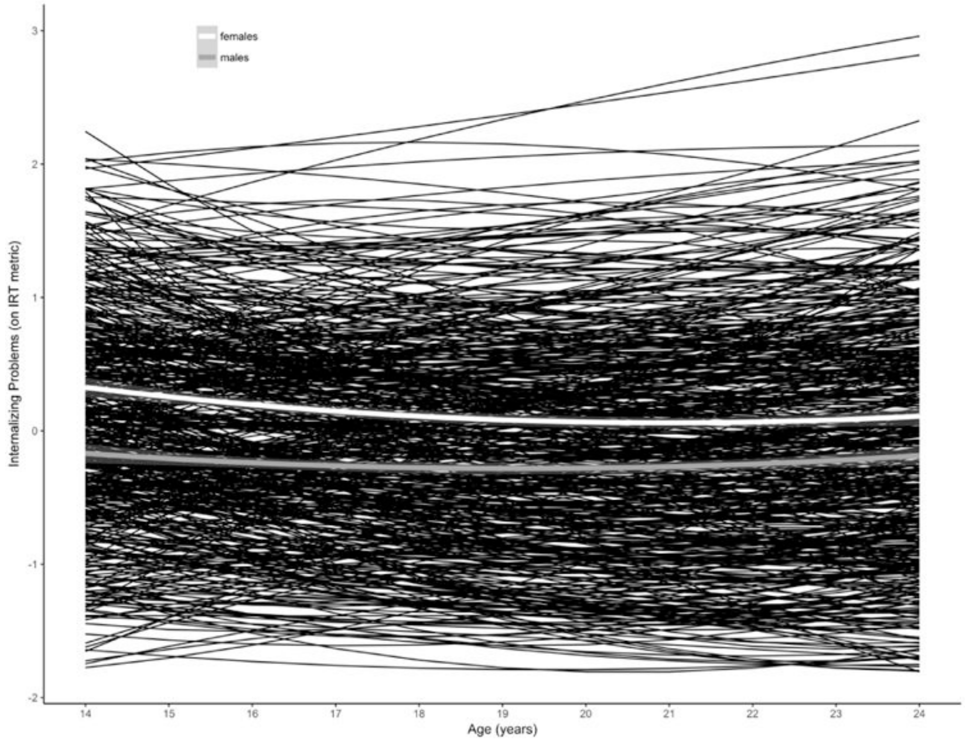
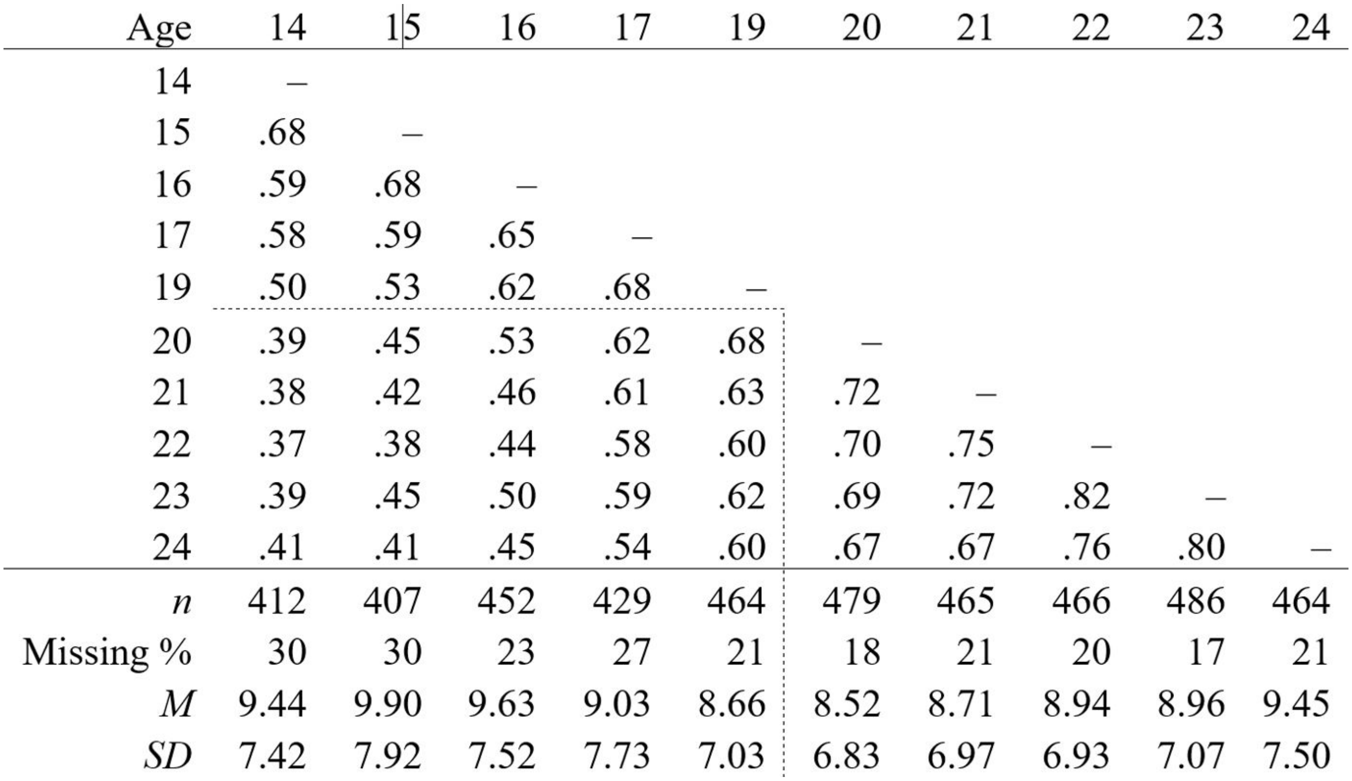


Figure 3.

Individuals' fitted quadratic trajectories of internalizing problems in black (on IRT metric of YSR at age 14). Average quadratic trajectory for females in white. Average quadratic trajectory for males in gray.

Table 1

Pearson correlation matrix (two-tailed) of raw internalizing problem scores and descriptive statistics.



Age	14	15	16	17	19	20	21	22	23	24
14	–									
15	.68	–								
16	.59	.68	–							
17	.58	.59	.65	–						
19	.50	.53	.62	.68	–					
20	.39	.45	.53	.62	.68	–				
21	.38	.42	.46	.61	.63	.72	–			
22	.37	.38	.44	.58	.60	.70	.75	–		
23	.39	.45	.50	.59	.62	.69	.72	.82	–	
24	.41	.41	.45	.54	.60	.67	.67	.76	.80	–
n	412	407	452	429	464	479	465	466	486	464
Missing %	30	30	23	27	21	18	21	20	17	21
M	9.44	9.90	9.63	9.03	8.66	8.52	8.71	8.94	8.96	9.45
SD	7.42	7.92	7.52	7.73	7.03	6.83	6.97	6.93	7.07	7.50

Note: all correlations are significant at $p < .001$ level. Dashed lines separate the scores from the Youth Self-Report (YSR; ages 14–19) from the Young Adult Self-Report (YASR; ages 20–24). Note that the mean scores on the YSR versus YASR are not directly comparable on the same metric because they had different numbers (and types) of items in the calculation of the Internalizing scale (YSR: 31 items, YASR: 23 items).

Table 2

Descriptive statistics and scaling parameters of the vertically scaled scores, along with descriptive statistics of the common items.

	Common Items		Vertically Scaled Scores		Scaling Parameters	
Age	***M***	***SD***	***M***	***SD***	**A**	**B**
14	6.398	4.897	0.000	1.000	–	–
15	6.885	5.437	0.032	1.066	1.091	0.017
16	6.673	5.209	0.014	1.012	0.925	-0.003
17	6.326	5.284	-0.079	1.046	1.064	-0.113
19	6.123	4.954	-0.133	1.004	0.950	-0.043
20	6.119	5.164	-0.134	1.035	1.048	-0.012
21	6.181	5.243	-0.115	1.045	1.001	0.021
22	6.384	5.252	-0.052	1.019	0.960	0.070
23	6.344	5.313	-0.077	1.064	1.068	-0.039
24	6.772	5.606	-0.025	1.115	1.057	0.041

Note: The vertically scaled scores are rescaled to be on the reference scale of the YSR at age 14. The scaling parameters are calculated in reference to the previous year. For example, the age 16 scaling parameters reflect the scaling parameters to link age 16 to age 15. To link age 16 to the reference scale at age 14, however, a process of linking and chaining is necessary (linking age 16 to age 15 using the age 16 scaling parameters, and then chaining them to age 14 by linking age 15 to age 14 using the age 15 scaling parameters). For instance, trait level scores at age 15 were rescaled to the scale at age 14 by multiplying the age 15 trait level scores by 1.091 and adding 0.017. Trait level scores at age 16 were rescaled to the scale at age 14 by first multiplying the age 16 trait level scores by 0.925 and subtracting 0.003 to put them on the age 15 scale, and then multiplying the new scores by 1.091 and adding 0.017 to put them on the age 14 scale.

Table 3

Linear growth curve model of internalizing problems.

Variable	***B***	**β**	***SE***	***DF***	***p***
intercept	-0.197	0.013	0.060	3977	.001
time	0.000	-0.036	0.008	3977	.963
Predictors of the intercepts					
female	0.477	0.183	0.081	539	< .001
African American	-0.189	-0.089	0.112	539	.091
Other Ethnicity	-0.329	-0.024	0.343	539	.337
Predictors of the slopes					
female	-0.018	-0.029	0.010	3977	.078
African American	-0.014	-0.016	0.015	3977	.329
Other Ethnicity	0.022	0.009	0.043	3977	.601
Variance components	*SD*				
intercept	0.83				
time	0.10				
residual	0.58				
Correlation between intercept and slope					$r = -.41$
Model Pseudo-R^2		.753			

Note: The model's pseudo R^2 was calculated as the squared correlation between the model's fitted and observed values (Singer & Willett, 2003).