CHILD DEVELOPMENT

Child Development, January/February 2021, Volume 92, Number 1, Pages e1–e19

Creating a Developmental Scale to Account for Heterotypic Continuity in Development: A Simulation Study

Isaac T. Petersen and Brandon LeBeau
University of Iowa

Daniel Ewon Choe
University of California, Davis

Many psychological constructs show heterotypic continuity—their behavioral manifestations change with development but their meaning remains the same. However, research has paid little attention to how to account for heterotypic continuity. A promising approach to account for heterotypic continuity is creating a developmental scale using vertical scaling. A simulation was conducted to compare creating a developmental scale using vertical scaling to traditional approaches of longitudinal assessment. Traditional approaches that failed to account for heterotypic continuity resulted in less accurate growth estimates, at the person- and group level. Findings suggest that ignoring heterotypic continuity may result in faulty developmental inferences. Creating a developmental scale with vertical scaling is recommended to link different measures across time and account for heterotypic continuity.

Developmental psychology seeks to elucidate processes of continuity and change across the life span and not just transitory outcomes at some stages in life. There are major challenges, however, in studying lengthy spans of development including cost, time, and most importantly, there are difficult conceptual, and statistical issues to address with respect to measurement. Many constructs change in their behavioral expression with development while retaining their meaning or underlying function, a phenomenon known as heterotypic continuity. This poses challenges for measurement.

Consider the construct of externalizing problems—challenging behaviors including aggression and oppositionality. Externalizing problems are often expressed as overt acts in early childhood, such as physical aggression, but externalizing problems are more often expressed in adolescence as covert and indirect or relational forms of aggression, rule-breaking, and drug use (Miller, Vaillancourt, & Boyle, 2009). When the construct of externalizing problems changes in behavioral expression across development, different measures should be used at different ages to retain construct validity of measurement. However, among two widely used measures of externalizing problems, the Child Behavior Checklist (Achenbach, 2009) and the Eyberg Child Behavior Inventory (Eyberg & Pincus, 1999), only the former uses different item content across ages to capture their changing manifestation. The challenge when comparing measurements across ages is accurately inferring whether differences in a measure's scores across time reflect true changes or differences in the measure's meaning (e.g., scores on an oral vocabulary measure could primarily reflect speech perception ability at one time point but could primarily reflect vocabulary knowledge at a later time point).

Although we present measurement challenges with examples of externalizing problems, considerable research has demonstrated many psychological constructs' changing behavioral expression with development. However, surprisingly little research has adopted measurement and statistical schemes that account for such changes when examining how people develop. That is, few studies have examined people's growth using different measures across development to maintain construct validity when the construct changes in its behavioral expression, with statistical and measurement tools to link the different measures across time. Thus, there is an inconsistency between the theory of how constructs develop and approaches to assess and study their development. Accounting for changes in the behavioral manifestation of constructs can improve the

We graciously thank Bob McMurray for providing helpful feedback on an early draft of the manuscript.

Correspondence concerning this article should be addressed to Isaac T. Petersen, Department of Psychological and Brain Sciences, University of Iowa, 175 Psychological and Brain Sciences Building, Iowa City, IA 52242. Electronic mail may be sent to isaac-t-petersen@uiowa.edu.

0009-3920/2021/9201-0028
DOI: 10.1111/cdev.13433

construct validity of measurement strategies and the accuracy of developmental inferences.

Heterotypic Continuity

Heterotypic continuity refers to the persistence of an underlying construct with behavioral manifestations that change across development (Caspi & Shiner, 2006; Cicchetti & Rogosch, 2002). To empirically establish heterotypic continuity of a construct, one must first identify developmental changes in the characteristics of that construct's content. Content of a measure includes facets assessed by a given measure (purportedly reflecting a given construct), and could include individual behaviors, questionnaire items, or sub-dimensions of a broader construct (e.g., aggression and oppositionality are content of externalizing problems). Content of a measure can be compared to content of the construct to evaluate the measure's content validity. Content validity reflects the extent to which a measure assesses all facets of a construct (i.e., there are no content gaps), without assessing facets of other constructs (i.e., there are no content intrusions). To empirically establish heterotypic continuity of a construct, its content should show cross-time changes in: (1) magnitude of rank-order stability of people's scores on the content across time (i.e., changes in the stability of individual differences across time), (2) the content's level on the construct, and/or (3) how strongly the content reflects the construct.

First, the content's cross-time changes in magnitude of rank-order stability of people can be examined with correlation or regression of the content across time. Second, one can consider whether the content shows cross-time changes in its level on the construct, referred to as difficulty or severity in item response theory (IRT). In (two-parameter) IRT, an item's difficulty parameter describes the construct level at which the probability of endorsing the item is 50%. For example, if a child sets fires, the child is likely to be higher in externalizing problems than children who argue, because fire-setting occurs less frequently than arguing and is a more severe form of externalizing behavior (Petersen, Bates, Dodge, Lansford, & Pettit, 2016). Thus, "sets fires" is more infrequent, severe, and has a higher difficulty parameter (level on the construct) than "argues."

Third, one can consider whether the content shows cross-time changes in how strongly it reflects the construct (i.e., stability of construct validity across time). How strongly the content reflects the construct is referred to as discrimination in IRT. An item's discrimination parameter describes how well the item distinguishes between low and high levels of the assessed construct (i.e., how well the item relates to the construct). For example, because an item asking how often a child attacks people is more relevant to externalizing problems than an item asking how often a child brags; "attacks people" has a higher discrimination parameter (construct validity) for externalizing problems than "brags" (Petersen et al., 2016). Examining how strongly the content relates to a latent construct can help determine which behaviors most strongly reflect a construct at a given point in development (e.g., Lee, Bull, & Ho, 2013).

If the content shows cross-time stability in (1) the magnitude of rank-order stability of people, (2) level on the construct, and (3) how strongly the content reflects the construct, the construct shows a stable factor structure across time. To the extent that the factor structure of the construct changes across development, the construct shows heterotypic continuity.

Extensive research in developmental psychology demonstrates how many constructs change in expression over time and show heterotypic continuity. However, surprisingly little research has considered how to examine people's developmental trajectories in constructs that change in manifestation (i.e., how to *account for* heterotypic continuity when examining development). Despite a proliferation of studies that examine people's growth trajectories, very few studies have examined trajectories in ways that account for heterotypic continuity by using different, age-appropriate measures across time to maintain construct validity (e.g., Petscher, Justice, & Hogan, 2018), and even fewer have done so in ways that allow researchers to examine absolute change rather than just relative, rank-order change (McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009; Petersen et al., 2018). This is a major problem for the field of developmental psychology. The study of development (i.e., change *and* continuity over time) is based on assessing the same or similar measures at multiple points in time with repeated assessments or cross-sectional age comparisons. The assumption of repeated measures is that scores are conceptually and statistically comparable across time, and therefore, different scores for the same person at different ages reflect true change (i.e., change in the person's level on the construct). If, however, the construct changes in manifestation over time and the measures do not accommodate these changes, the measures differ in their validity for the same construct across time—that is, they lack construct validity invariance. Thus, a failure to account for heterotypic continuity may result in invalid measures (with respect to the same construct) over time and, therefore, faulty inferences about development. Failing to account for heterotypic continuity results in measures that are less able to detect developmental change (Petersen et al., 2018) and in misidentified growth trajectories (Chen & Jaffee, 2015). Thus, heterotypic continuity is a characteristic of many psychological constructs, but failure to account for heterotypic continuity is a serious problem in developmental psychology because it presents challenges to the validity of our measures and inferences.

Accounting for Heterotypic Continuity in Development

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). Thus, a consequence of heterotypic continuity is that different measures across time may be necessary to assess the same construct over time (Widaman, Ferrer, & Conger, 2010). There are three primary approaches to assessing a construct over time: (1) all possible content (e.g., observable behaviors, questionnaire items) across all ages, (2) only common content across all ages, and (3) only construct-valid content at each age. To describe the three approaches, consider three content sets in Figure 1: content set A refers to content that is construct-valid at only T1, content set B is construct-valid at both T1 and T2, and content set C is construct-valid at only T2. For instance, in a longitudinal study of externalizing problems from early childhood (T1) to adulthood (T2), "biting others" may be in content set A, "noncompliance" and "oppositionality" may be in content set B, and "drug use" may be in content set C. The three approaches would be as follows: (1) using all possible content across all ages: ABC at T1 and T2, (2) using only common content across all ages: B at T1 and T2, or (3) using only construct-valid content at each age: AB at T1 and BC at T2. Traditionally, developmental psychologists have used all possible content (Approach 1) or only common content (Approach 2) across all ages when assessing a construct over time. However, we argue that using the construct-valid content (Approach 3) is important to account for heterotypic continuity. This would mean using different measures across time to assess age-relevant content and to establish construct validity invariance. Once conceptual equivalence of the measures has been established, statistical equivalence of the different measures is a crucial consideration.

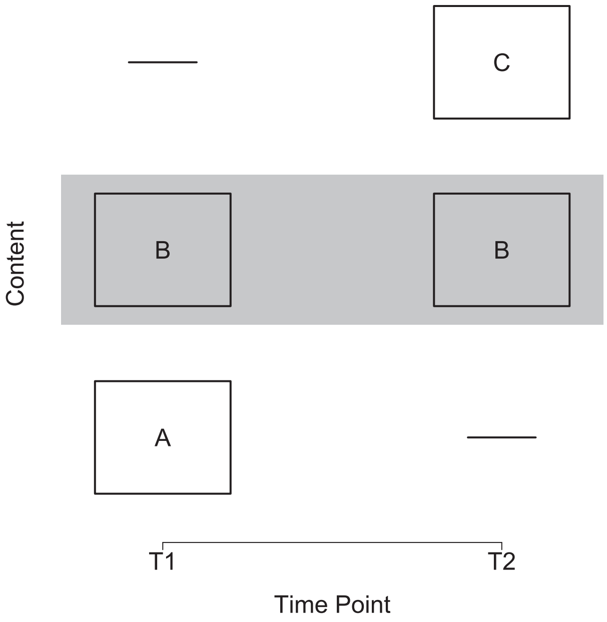


Figure 1. Depiction of using only the construct-valid content at each age. Content set A corresponds to content that is construct-valid at only T1. Content set B corresponds to content that is construct-valid at both T1 and T2. Content set C corresponds to content that is construct-valid at only T2. The "common content" (content set B) is highlighted in gray. The three approaches to assessing a construct over time are as follows: (1) using *all possible content* across all ages: ABC at T1 and T2, (2) using *only common content* across all ages: B at T1 and T2, or (3) using *only construct-valid content* at each age: AB at T1 and BC at T2.

Creating a Developmental Scale to Account for Heterotypic Continuity

There are key challenges to ensuring the statistical equivalence of different measures. A promising approach to linking different measures across time to account for heterotypic continuity is the creation of a developmental scale using vertical scaling. We propose the following approach to create a developmental scale and account for heterotypic continuity when examining developmental trajectories. First, select construct-valid content at each age that, ideally, partially overlap at adjacent ages (see Figure 1). Second, ensure construct validity invariance of the different measures across ages. Third, test longitudinal factorial invariance of the different measures across ages. Fourth, use vertical scaling to link the different measures across ages on the same developmental scale. We describe Steps 2–4 later. Fifth, estimate people's growth trajectories using their vertically scaled scores on this developmental scale.

In vertical scaling, measures that assess the same construct but differ in difficulty are placed on the same scale. The goal of vertical scaling is to assemble and link a construct-valid set of content at each age that have some overlap in content at adjacent ages (i.e., common content) on the same scale. Although vertical scaling uses the common content to put two different measures on the same scale, researchers have used Bayesian approaches to link different measures with no common content (Oleson, Cavanaugh, Tomblin, Walker, & Dunn, 2016). In general, the lesser the amount of unique content and the greater the amount of common content, the more likely the different measures will be successfully linked (Hanson & Béguin, 2002; McArdle et al., 2009). Scores on the construct-valid content at the reference age set the scale, the common content adjusts subsequent scores to that scale, and all construct-valid content (i.e., both common and unique content) at a given time point is used to estimate each person's score on that scale. Thus, the common content is used to determine the general form of change on an identical scale, but all developmentally relevant, construct-valid content is used to estimate each person's construct level on this scale.

Multiple vertical scaling approaches exist. In the present study, we focus on the IRT approach to vertical scaling to account for heterotypic continuity in development. The IRT approach to vertical scaling uses scaling parameters that put people's construct scores from different measures on the same metric. The scaling parameters are determined as the linear transformation (i.e., intercept and slope parameter) that, when applied to the second measure, minimizes differences between the probability of a person endorsing the common content across two measures

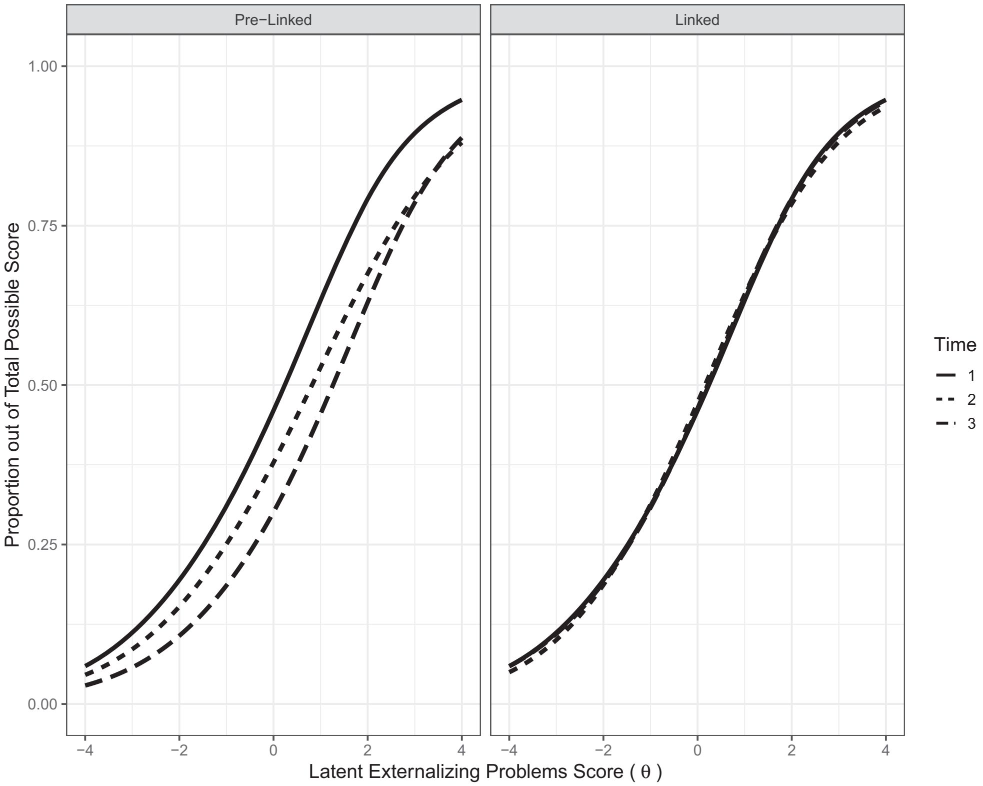


Figure 2. The figure illustrates the effect of linking latent externalizing problem scores, θ, across three time points. The left panel illustrates test characteristic curves representing the model-implied proportion out of total possible scores across latent externalizing problem scores at T1, T2, and T3, before the linking process. The right panel illustrates the test characteristic curves after the linking process, which minimizes differences between common items' discrimination and severity. Discrimination is depicted as the steepness of the slope at the inflection point of the test characteristic curve. Severity is represented by the value on the *x*-axis at the inflection point of the test characteristic curve. The left panel shows that the measures increase in severity with age. The right panel shows considerably smaller differences between the three test characteristic curves, which provide empirical evidence that linking successfully placed the latent externalizing problem scores across time on a more comparable scale (i.e., more similar discrimination and severity of the common content).

(see Figure 2). That is, IRT links scales of measures based on the difficulty and discrimination of the common content, and is often employed for vertical scaling, especially in cognitive and educational testing. For instance, McArdle et al. (2009) and McArdle and Grimm (2011) examined the development of cognitive ability from 2 to 72 years of age using different measures across time and an IRT approach to vertical scaling. The authors used developmentally appropriate, construct-valid content for vocabulary and memory span, and linked the different measures based on the difficulty of common content. Thus, vertical scaling can be used to capture people's unique trajectories on a construct across a lengthy developmental span—assessing people's absolute growth in the construct, unlike in other approaches. Trajectories derived from vertical scaling can then be linked to risk factors, protective factors, and downstream outcomes.

Description of an Empirical Example Using Vertical Scaling

Many studies have used vertical scaling in the fields of education and cognitive testing to assess growth with different measures over lengthy developmental spans (Kenyon, MacGregor, Li, & Cook, 2011; McArdle & Grimm, 2011; McArdle et al., 2009; Wang, Jiao, & Zhang, 2013). We are aware of only two studies that used vertical scaling to study social development (Petersen & LeBeau, in press; Petersen et al., 2018). Petersen et al. (2018) used different measures to assess the development of internalizing problems—one measure in adolescence (31 items) and another in adulthood (23 items). The measure in adolescence included content assessing depression, anxiety, and somatic complaints. Given the changing manifestation of internalizing problems, the measure in adulthood included content assessing depression and anxiety, but not somatic complaints. The authors used vertical scaling to link the different measures (i.e., the construct-valid content) to be on the same scale across development to account for heterotypic continuity. To do this, the authors fit separate IRT models at each age that estimated each item's discrimination and difficulty. They then linked people's internalizing factor scores across time on the same scale by calculating scaling parameters that linked the discrimination and difficulty of the two measures' common content (i.e., 17 items that assess depression and anxiety; see formulas in the Method section of the present simulation study). A growth curve was fit to each person's vertically scaled internalizing factor scores. The authors compared the average trajectory when using vertical scaling of the construct-valid content to the traditional approach of using only common content. The authors observed a group-level decrease in internalizing problems from adolescence to adulthood when using vertical scaling, but they observed no change when using only common content. The authors replicated this pattern with a second approach to vertical scaling (Thurstone scaling). Thus, not only does vertical scaling permit studying lengthier spans of development, but it yields inferences that are better able to detect developmental change. However, to our knowledge, no study has compared the three approaches to assessing a construct over time, and no study has conducted a simulation to better understand the factors that influence the accuracy of these approaches.

The Present Study

In the present study, we conducted a simulation that compares different approaches to assessing a construct over time. This allowed us to determine whether vertical scaling is more accurate than traditional measurement approaches when studying growth trajectories in the context of heterotypic continuity. We compared the three approaches to assessing a construct across time: (1) all possible content, (2) only common content, and (3) only construct-valid content, using a simulation. If items on the construct differed across time only in difficulty, the results from all three approaches would theoretically be the same. This is because all items would be construct-valid at all ages and, therefore, Approaches 1, 2, and 3 would administer the same items. Rather, we were more interested in understanding the influence of heterotypic continuity where the items that are construct-valid for a construct change with development. Therefore, we considered externalizing problems, where items would be expected to differ in difficulty and discrimination across development. We conducted the simulation to map onto the construct of externalizing problems from early childhood to adolescence because externalizing problems show heterotypic continuity across this span (Petersen, Bates, Dodge, Lansford, & Pettit, 2015). We simulated content in accordance with hypothetical measures of externalizing problems across this developmental span—some common content across ages to capture the theoretical core of externalizing problems, and some unique content across ages to capture the changing manifestation of externalizing problems (see example items in Table S1). To allow comparing growth curves derived from the three approaches, we simulated people's externalizing problem scores for the three approaches at three time points (because three-wave designs are common longitudinal studies for growth curves), corresponding to early childhood (age 3 years; T1), middle childhood (age 8 years; T2), and adolescence (age 13 years; T3). Because of the changing manifestation of externalizing problems, the construct-valid content differed across ages. Therefore, for using only construct-valid content (Approach 3), we used vertical scaling to link the different measures across ages. We generated estimates of people's factor scores on the construct-valid content, and then compared the vertically scaled factor scores of the construct-valid content to the traditional scoring approaches—sum scores of all possible content (Approach 1), and sum scores of only common content (Approach 2).

Method

We simulated people's externalizing problem scores for the three approaches at each of three time points (T1, T2, and T3). We used the IRT approach to vertical scaling, which is flexible in that it allows items to differ across measurement occasions. The technique relies on common or anchor content (Holland & Dorans, 2006; Kolen & Brennan, 2014) to put different measures on the same scale, by linking the difficulty and discrimination of the common content. For common content, the same participant receives a score for the same content across two (or more) time points. In our approach, scores on the construct-valid content at T1 set the scale, the common content adjusts subsequent scores to that scale, and all construct-valid content (i.e., both common and unique content) at a given time point is used to estimate each person's score on that scale. Next, we provide more detail about the vertical scaling approach used in this simulation. In the context of externalizing problems, a higher difficulty parameter reflects a higher, more severe level of externalizing problems, so we refer to the difficulty parameter as severity when describing the simulation. The analytical code for this simulation is available on Open Science Framework: https://osf.io/ewmzd.

The simulation varied the following conditions: (a) number of items (two levels; 20 and 40 items) at each time point, (b) the proportion of all content that is common across time points (three levels; .2, .5, .8), (c) the change in severity of the common content (seven levels; −.5, −.25, −.1, 0, .1, .25, .5), and (d) the average severity of the unique content (five levels; −2, −1, 0, 1, 2). These simulation conditions were fully crossed in a factorial design which resulted in a total of $2 \times 3 \times 7 \times 5 = 210$ simulation conditions and each condition was replicated 1,000 times. By default, conditions were "balanced" such that the number of construct-valid items was the same across all time points. Sample size remained fixed at 1,000 respondents across all conditions.

In addition to these balanced cases, we also explored several unbalanced cases. The unbalanced simulation did not manipulate the number of items; instead, we fixed the number of construct-valid items at 15 items at T1 (5 unique items), 25 items at T2 (15 unique items), and 40 items at T3 (30 unique items). We held the number of common items fixed at 10 items. The two manipulated conditions included the change in severity of the common content (two levels; −.5 and .5) and the average severity of the unique content (three levels, −2, 0, 2). These two manipulated simulation conditions were fully crossed resulting in a total of six simulation conditions, each of which was replicated 1,000 times. These conditions were chosen based on descriptive analysis of the balanced simulation conditions and represented the cases where differences were expected to be more pronounced. Sample size again remained fixed at 1,000 respondents across all conditions.

We conducted all analyses in R version 3.5.1 (R Core Team, 2019). We performed IRT model fitting with the mirt R package (Chalmers, 2012), and performed linking with the plink R package (Weeks, 2010).

Simulation Procedure

We simulated data from an IRT framework with the following general steps:

1. We randomly generated item parameters, discrimination and severity, for each common item (e.g., "argues with others") at T1. Item discrimination parameters followed a log normal distribution with a mean log of 0 and a standard deviation on the log scale of .25. The log normal distribution is commonly used as a prior distribution for the discrimination parameter when estimating IRT item parameters because the discrimination parameter is commonly > 0 and skewed (Harwell & Baker, 1991). We generated the item severity parameters from a random normal distribution with a mean of 0 and a standard deviation of 1.
2. We adjusted (i.e., linked) the item parameters of the common content to reflect changes in

severity that would occur due to shifts in the construct of externalizing problems. We applied the adjustment once between T1 and T2 and again between T2 and T3. For example, if the change in severity of the common content was .5, this means that the same common item would become half a standard deviation more severe at T2 and a full standard deviation more severe at T3 compared to T1, however, the item would only be half a standard deviation more severe between T2 and T3. We held the discrimination of the common content to be the same across the time points because the common content remained construct-valid.

3. We randomly generated item parameters, discrimination and severity, for the unique items at each time point that the items were construct-valid (e.g., "uses illegal drugs" at T3). We generated the discrimination parameters from the same distribution as the common content. We generated the severity parameters from a standard normal distribution with a mean that was manipulated by the simulation conditions and a fixed standard deviation of 1.

   We then transformed item parameters for these unique items to reflect that these items would be more severe and less discriminating at times when they are not construct-valid. For example, the item "uses illegal drugs" would be expected to be more severe and less construct-valid for externalizing problems in early childhood compared to adolescence, consistent with IRT estimates from other studies (Petersen et al., 2016). We assumed that the severity would increase by five standard deviations and the discrimination would be half as discriminating at ages when they were not construct-valid (compared to ages when they were construct-valid). For example, for unique items most relevant at T1 (e.g., "temper tantrums"), we added these same items to the instrument at T2, but the item parameters for these items at T2 had five standard deviations added to their severity and the discrimination was cut in half; we also made the severity of these same items more severe at T3 where another five standard deviations were added to the severity, but we kept the discrimination the same as at T2. We followed similar logic for items most relevant at T2 (e.g., "vandalizes property") and T3 (e.g., "uses illegal drugs") where we transformed those item parameters to reflect that the items are less discriminating and more severe at other time points when they were not construct-valid. In the all possible content approach, we included these unique items at all ages (even when the item was not construct-valid). Thus, the all possible content approach included transformed items. In the construct-valid content approach, we included a given unique item only at time points when the item was construct-valid (i.e., we dropped items at ages when they were construct-invalid). In the common content approach, none of these unique items was included at any time point.
   These adjustments to severity of items and discrimination reflect changes that occur in real-world items that assess constructs when they are not endorsed frequently and are not construct-valid. For example, if an item is not endorsed frequently, the severity parameter increases. If an item is not construct-valid, the discrimination parameter decreases. Also, when an item becomes high or low in severity, the item tends to be more difficult to estimate because there are fewer endorsed options or more generally less variation, and, as such the discrimination is also more difficult to estimate and becomes smaller. In order to simulate responses that would reflect the severity of the items that are not construct-valid at a given time, we made these adjustments in item parameters to adequately generate responses to items that we would expect to observe if these assessments were administered in practice (e.g., asking a parent if their 3-year-old child drinks alcohol, which is clearly construct-invalid for externalizing problems at that age).

4. Next, we simulated latent externalizing problem scores (factor scores) for the population from a standard normal distribution with a mean of 0 and a standard deviation of 1.
5. Upon generation of population item parameters and factor scores, we simulated people's item responses. We calculated the probability of a person endorsing an item (i.e., receiving a score of 1) based on the item parameters and the person's factor score. The model-based probability of item endorsement as a function of a person's level on the construct is the item characteristic curve (de Ayala, 2009). Once we calculated the item characteristic curve, we evaluated the probability of endorsing the item compared to the probability of endorsing a random uniform value. If the random uniform value was less than the probability of endorsing the item, we recorded the item as being

endorsed (a score of 1) for that person. If the random uniform value was greater than the probability of endorsing the item, we recorded the item as not being endorsed (a score of 0) for that person. We repeated this procedure for every item and every person (N = 1,000).

Model Fitting

Once we generated the simulated responses for each item and person, we generated scores from the three different measurement approaches—summed scores of all possible content, summed scores of only common content, and vertically scaled factor scores of only construct-valid content. The all possible content scores and common content scores were summed scores across items administered at each time point, because an item sum is the most common way these approaches are used in the literature. The primary difference between the all possible content scores and common content scores is the number of items and which items were included in the calculation of each person's sum score of externalizing problems. Only the common items were included in the common content score, whereas all possible items were included in the all possible content score, including items that were not construct-relevant at a given age. Sum scores were not calculated for the construct-valid content, because the number of construct-valid items differed at each age, which would result in different, noncomparable metrics across ages. Therefore, for the construct-valid content approach, we used vertical scaling to generate estimates of people's level of externalizing problems on the same scale, as described next.

We also conducted intermediate steps to remove potential confounding effects and clarify the results. As an intermediate step in the construct-valid content approach, we fit IRT models, but did not perform the linking procedure (i.e., vertical scaling), and generated the factor scores based on the separate IRT model at each time point. This would represent a case where any developmental shift in the severity or discrimination of the items would be ignored and the assumption would be that the items would be equivalent across the developmental span. When the severity of the common content remains similar across time, the IRT modeling with vertical scaling and the IRT modeling without vertical scaling would be expected to produce similar results. Similarly, to aid in the interpretation of results, for the all possible content approach, we fit IRT models with and without vertical scaling. This would allow for a comprehensive comparison of the three approaches despite the all possible content approach having slightly different data generation procedures (to be consistent with heterotypic continuity). We also used these intermediate steps to disentangle potential confounding effects and isolate the reasons why the accuracy may differ between the approaches (e.g., IRT vs. vertical scaling).

IRT and Vertical Scaling

In the IRT approach to vertical scaling, we followed a three-step procedure. First, we fit separate IRT models at each time point, because it is considered safer than fitting all items across all ages in the same model, which is more likely to violate IRT assumptions (Kolen & Brennan, 2014). Second, we linked people's externalizing problem factor scores across time on the same scale by calculating scaling parameters that linked the estimated item parameters. The scaling parameters linked the discrimination and difficulty of the different measures' common content by minimizing differences between the probability of a person endorsing the common content across the measures (see the following formulas). This step removes severity and discrimination differences of the common content across time points and the item parameters are linearly rescaled to a single unified metric. Finally, we calculated each person's factor score (level of externalizing problems) using linked item parameters and item responses. Each step is described in more detail as follows.

We fit a two-parameter logistic (2pl) IRT model to the item-level data that were simulated for respondents (de Ayala, 2009). The 2pl IRT model takes the following form:

$$p(x_i=1|\theta,a_i,b_i)=\frac{1}{1+e^{-a_i(\theta-b_i)}}, \tag{1}$$

where the probability of endorsing the item is based on three parameters, θ, a_i, and b_i representing the latent variable, item discrimination, and item severity. The i subscript for the item discrimination and severity indicate that they are estimated uniquely for each item. When fitting the IRT model, the default assumption is that the latent variable has a standard normal distribution. This means the factor scores, and also item severity, would be assumed to have a mean of 0 and a standard deviation of 1 at each time point due to the separate IRT model estimation at each time point. The item parameters and latent variable would not necessarily be on a comparable scale; thus we used linking to create a vertical scale and ensure that the factor scores and item parameters were comparable.

We linked the item parameter estimates from the 2pl model shown in Equation 1 to remove differences in item severity and discrimination over time. Linking is an iterative procedure that estimates linking constants that minimize differences in the characteristic curves between adjacent time points (i.e., comparing T1–T2 and T2–T3). There are two commonly recommended linking methods, the Stocking-Lord (SL) procedure (Stocking & Lord, 1983) and the Haebara procedure (Haebara, 1980). Both procedures link the severity and discrimination of the common content, but the SL procedure performs linking at the test-level of the common content based on the test characteristic curve, whereas the Haebara procedure performs linking at the item-level of the common content based on item characteristic curves. The test characteristic curve of the common content is the summed likelihood of having the common items endorsed (or correct) given the item parameters at specific construct scores (Kolen & Brennan, 2014). We used the SL procedure because we were most interested in construct-level scores instead of scores on specific items (Kolen & Brennan, 2014; LeBeau, 2017). The SL scaling parameters minimize the differences between the probability of a person endorsing the common content (at the aggregate level) across the two measures. Even though we used the SL procedure, simulation studies have shown little difference between these two methods (Hanson & Béguin, 2002; Kim & Lee, 2006; LeBeau, 2017).

We then estimated linking constants, including a slope and intercept, at adjacent time points, and we set the reference age to set the scale of the latent variable at the first time point in the simulation. We did this to transform the item parameters at all time points to be on the same scale as the item parameters at T1. We estimated linking constants to minimize differences in the test characteristic curve of the common content between T1 and T2 and between T2 and T3. After estimating the linking constants, we transformed the item parameters at T2 and T3 according to the following equations:

$$a(\text{time}_k) = \frac{a(\text{time}_j)}{A}, \tag{2}$$

$$b(\text{time}_k) = A \times b(\text{time}_j) + B, \tag{3}$$

where $a(\text{time}_j)$ and $a(\text{time}_k)$ are discrimination parameter estimates for the common items at adjacent time points j and k respectively; $b(\text{time}_j)$ and $b(\text{-time}_k)$ are severity parameter estimates for the common items at adjacent times j and k, respectively; A represents the slope scale parameter, and B represents the intercept scale parameter. To shift all item parameters to the same scale as those of T1, we applied all previous adjacent scaling constants to the item parameters in a process called linking and chaining. For example, when transforming item parameter estimates at T2 to the T1 scale, we used a single set of linking constants. However, when transforming item parameter estimates at T3 to the T1 scale, we used two sets of linking constants: first, linking constants to move T3–T2, and second, linking constants from T2–T1 to move the newly transformed T3 item parameters to the T1 scale. Figure 2 shows a visualization of the linking process for a single replication and the simulation condition where we generated the common items to be half a standard deviation apart between time points.

Finally, once we linked the item parameters, we used the newly transformed and linked item parameters in tandem with the individual response string to calculate the person's factor score, which represents their estimated level on the latent construct of externalizing problems (i.e., construct scores), commonly referred to as ability scores in educational measurement terminology. We placed the factor scores on the same scale by the linking procedure described above using the following equation:

$$\theta(\text{time}_1) = A \times \theta(\text{time}_j) + B, \tag{4}$$

where $\theta(\text{time}_1)$ represents the construct factor score at T1 (the reference scale) and $\theta(\text{time}_\text{j})$ represents the construct factor scores at subsequent time points. We used a process of linking and chaining (described above) to link the factor scores at T3 to the T1 scale.

We calculated factor scores with the expected a posteriori (EAP) scoring method (Thissen, Pommerich, Billeaud, & Williams, 1995) which is a Bayesian estimation procedure. The primary benefits of the EAP approach are that the factor scores are more stable due to the use of a prior distribution and the factor scores are able to be estimated for people who endorse positively or negatively all items at each time point (i.e., they have a summed score of 0 or equal to all of the items on the survey). We estimated EAP factor scores separately at each time point.

Growth Curve Modeling

We fit a growth curve to each person's externalizing problem scores using the three measurement approaches. We used the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in R to test hierarchical linear models. Estimates for the fixed effects were standardized for direct comparison across the different metrics. The unconditional linear mixed model took the following form:

$$Y_{tp} = \beta_0 + \beta_1 \text{time}_{tp} + b_0 + \varepsilon_{tp}, \tag{5}$$

where Y_{tp} is the outcome at time t for person p, β_0 and β_1 are the fixed effects for the intercept and linear slope respectively, b_0 is the random intercept, and ε_{tp} are person- and time-specific residuals. We did not model random slopes due to convergence issues resulting from small variances of the slopes across people. If there is variation in the slopes across people (i.e., the variance of the slopes across people does not equal 0 in the population), prior simulation evidence indicates that the fixed effect estimates will be unbiased (Kwok, West, & Green, 2007; LeBeau, 2016), however, the standard errors for the linear slope may be biased, resulting in inflated Type I errors (LeBeau, 2018; LeBeau, Song, & Liu, 2018). Because the primary interest in this study was the direction and magnitude of the slope estimates (rather than tests of whether the slopes differed reliably from 0), we deemed this approach satisfactory.

Outcomes

The outcomes for this simulation include the standardized linear slope estimate across the different methods and correlations between the construct estimate and the construct population value. We calculated the standardized linear slope based on an unconditional linear mixed model (Fitzmaurice, Laird, & Ware, 2011) that included a random intercept and fixed effects for the intercept and linear slope. The linear slope was standardized to turn the slope into a standard deviation metric across all methods, thus removing differences in variation across the different methods. The expected change in the standardized slope would be the inverse of the change in the severity of the common content as depicted in Equation 4. Therefore, we computed slope bias as the difference between the standardized slope estimate and the inverse of the change in severity of the common items.

We used analysis of variance (ANOVA) to determine which simulation conditions explained variation in the estimated standardized linear slopes, consistent with generalizability theory (Shavelson, Webb, & Rawley, 1989). This approach has been used by prior studies (Kwok et al., 2007; LeBeau, 2016, 2018) and is helpful for exploring interactions among the simulation conditions. In the ANOVA model, the standardized linear slopes served as the outcome, and the simulation conditions were the predictors. We explored up to four-way interactions among the predictors. We used effect sizes, rather than p-values, to guide which simulation conditions were important. Eta-squared statistics represented the proportion of variance the simulation condition explained in the outcome. We set eta-squared statistics > .01 or 1% of explained variation as the a priori threshold for identifying simulation conditions that explained practically useful amounts of variation in standardized linear slopes.

We also calculated criterion-related validity of the scores from each measurement approach. We calculated criterion validity as the correlation between people's population values (i.e., "truth") and estimates of people's level on the construct from each measurement approach (i.e., summed all possible content scores, summed common content scores, and vertically scaled IRT factor scores of the construct-valid content). We estimated Pearson correlation coefficients at T1 because true construct scores were known for that time point. We expected that correlations would be positive in all cases; however, a stronger correlation indicates a stronger degree of concurrent criterion-related validity when estimating the person's level of externalizing problems with a particular method. We hypothesized that the vertically scaled scores of the construct-valid content would provide the strongest criterion validity.

Finally, we calculated bias and mean absolute error (MAE) statistics that compared the score estimates from each method to the "true" construct scores at T1, as a validation check. We calculated bias and MAE according to the following formulas:

$$\text{bias} = \theta - \mathrm{\theta} \tag{6}$$

$$\text{MAE} = \frac{\sum_{l=1}^{n} \left| \theta - \mathrm{\theta} \right|}{n}, \tag{7}$$

where θ and $\mathrm{\theta}$ are the estimated construct score and true population score, respectively, at T1, and l indexes the number of replications going from 1 to n. We calculated bias and MAE at T1 because the true construct scores were known for that time point. Due to differences in the scale of the construct scores compared to the all possible content and only common content approaches, we standardized the construct score estimates from these approaches by converting them to *z*-scores. This transformation reflected the same scale of the true population scores and the construct-valid scores, with a mean of 0 and a standard deviation of 1. In general, bias scores that are closer to 0 indicate that the method is less biased (i.e., neither systematically underestimates nor overestimates the construct scores). Smaller MAE statistics indicate estimates that are more accurate (i.e., precise in relation to the true construct scores).

Results

Standardized Slope Estimates

Table 1 shows ANOVA results and subsequent eta-squared effect sizes of the model that explains variation in the standardized slope estimates with the simulation conditions. Eta-squared statistics > .01 are shown in the table; other terms where eta-squared is < .01 are omitted for a clearer understanding of important predictors.

Overall, the ANOVA model fit the data well, accounting for about 98% of the total variation in the standardized slope estimates. The predictors in Table 1 (that explained more than 1% of the variation) collectively explained about 96% of the total variation. The strongest predictor was the change in severity of the common content, which explained just over 56% of the variation in standardized linear slope estimates. The next strongest predictor was the interaction between the change in severity of the common content and the method to obtain construct scores, which explained about 35% of variation. The number of common items did not show up as a main effect, but rather as an interaction with the change in severity of the common content, explaining about 1.5% of the variance.

Table 1
Simulation Conditions That Explained Greater Than One Percent of Total Variation in the Slope Estimates

Term	Sum of squares	η^2
Change common severity	18,852	.566
Unbalanced × Method	857	.032
Change Common Severity × Number of Common Items	379	.014
Change Common Severity × Method	1,800	.348

Note. Residual η^2 was .02, therefore the model explained about 98% of the variance in slope estimates. "×" denotes an interaction effect between the two variables. "Unbalanced" was coded as 0 if the number of construct-valid items was the same across all ages, and was coded as 1 if the number of construct-valid items differed across ages. "Change Common Severity" indicates the direction and magnitude of change in severity of the common content. "Method" refers to the three approaches for estimating people's level on the construct of externalizing problems —all possible content (Approach 1), only common content (Approach 2), and the construct-valid content (Approach 3), in addition to three methods representing intermediate steps (all possible content using item response theory (IRT) without linking, all possible content using IRT with linking, and the construct-valid content without linking).

Figure 3 shows the effects of the unbalanced item design, change in severity of the common content, and the method used on standardized slope estimates. When item design was balanced, shown in the top row of Figure 3, a similar trend in the direction of the standardized slopes was found across four of the six methods. A negative trend was found between the change in severity of the common content and the standardized linear slope estimate for all possible content (summed scores), all possible content with vertical scaling, only the common content, and construct valid-content with vertical scaling. This was expected due to the linking procedures, as the common content becomes more severe on average (i.e., a positive change in severity of the common content), the linking procedures decrease the person's level on the construct on average for these items to ensure the common content has the same probability of being endorsed at each age (this can be seen mathematically when comparing the linking equations shown above). In other words, if the common items (e.g., "throws objects when upset") are more severe on average at T2 than at T1, the items are endorsed less frequently at T2 than at T1. Thus, in such a case, the linking process would decrease people's level on the construct on average at T2 to place the factor scores from these items on the same scale as the factor scores at T1. Compared to the other approaches, the all possible content approach showed less precision (i.e., more variance) across conditions in the standardized slope estimates and tended to show greater slope bias (i.e., slope estimates were closer to 0) across all values of the change in common item severity. The construct-valid content approach showed similar trends to the common content approach; however, when the change in severity of the common content was more extreme (i.e., .5 or −.5), the standardized slope estimates for the construct-valid content approach were larger on average and showed less slope bias.

The greatest slope differences between the methods occurred when the item design was unbalanced. In this context, the construct-valid content approach and the common content approach produced similar results to those of the balanced case.

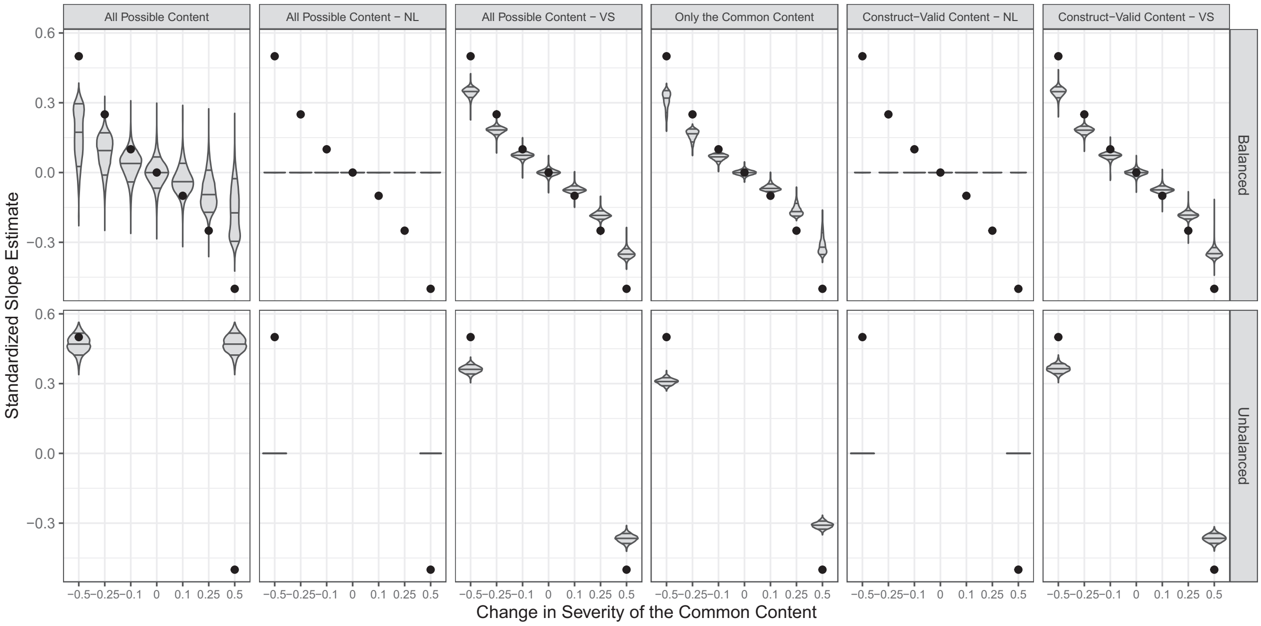


Figure 3. The figure shows the distribution of standardized slope estimates by (a) change in severity of the common content (*x*-axis), (b) the method (columns), and (c) whether the number of construct-valid items was balanced (top row) or unbalanced (bottom row). The 10th, 50th, and 90th percentiles are represented by the horizontal lines within the violin plots. The black points indicate the expected "true" slope change based on the inverse of the change in severity of the common content. "All Possible Content" and "Only the Common Content" were item sum scores, and the "Construct-Valid Content" were estimated with item response theory. "All Possible Content—No Linking (NL)" and "All Possible Content—Vertical Scaling (VS)" were intermediate steps that used item response theory without and with vertical scaling, respectively. "Construct Valid Content—NL" and "Construct Valid Content—VS" used item response theory without and with vertical scaling, respectively. The violin plots were flat in the item response theory approaches that did not perform linking because item response theory estimates the mean to be zero at each time point. "Construct-Valid Content—NL" was an intermediate step.

The primary difference was in the all possible content approach where the unbalanced case gave disproportionate weight to the greater number of construct-valid items at T3, resulting in larger standardized slope estimates that did not follow the expected trajectory. The difference was particularly stark for the change in severity of the common content condition of .5, where the standardized linear slope was *positive* for the all possible content approach, which diverged from the construct-valid content approach and the common content approach, and did not follow the expected *negative* trajectory based on the item generation process.

To disentangle the confounding effect of method, we also explored intermediate steps. When IRT was used without vertical scaling, standardized slope estimates were 0 across all of the change in severity of the common content conditions. This was because IRT commonly estimates the latent construct with a mean of 0 and a standard deviation of 1. Therefore, when the separate IRT models were fit without vertical scaling, the latent construct was estimated to have a mean of 0 at each of the three time points. Thus, IRT without vertical scaling ignored the change in the severity of the common content, and erroneously estimated a standardized slope of 0 (i.e., no mean-level change in the latent construct over time).

Finally, we also conducted IRT with vertical scaling for the all possible content approach, and this yielded similar results to the construct-valid content with vertical scaling. This result was expected because measurement error was not included in the simulation, and the IRT approach to vertical scaling downweighs construct-invalid content in estimating factor scores.

Sensitivity Analysis

We conducted a sensitivity analysis using maximum likelihood estimates to ensure we used the most stable estimator (see Appendix S1).

Criterion-Related Validity

Figure 4 shows correlations between the true population construct scores at T1 and the

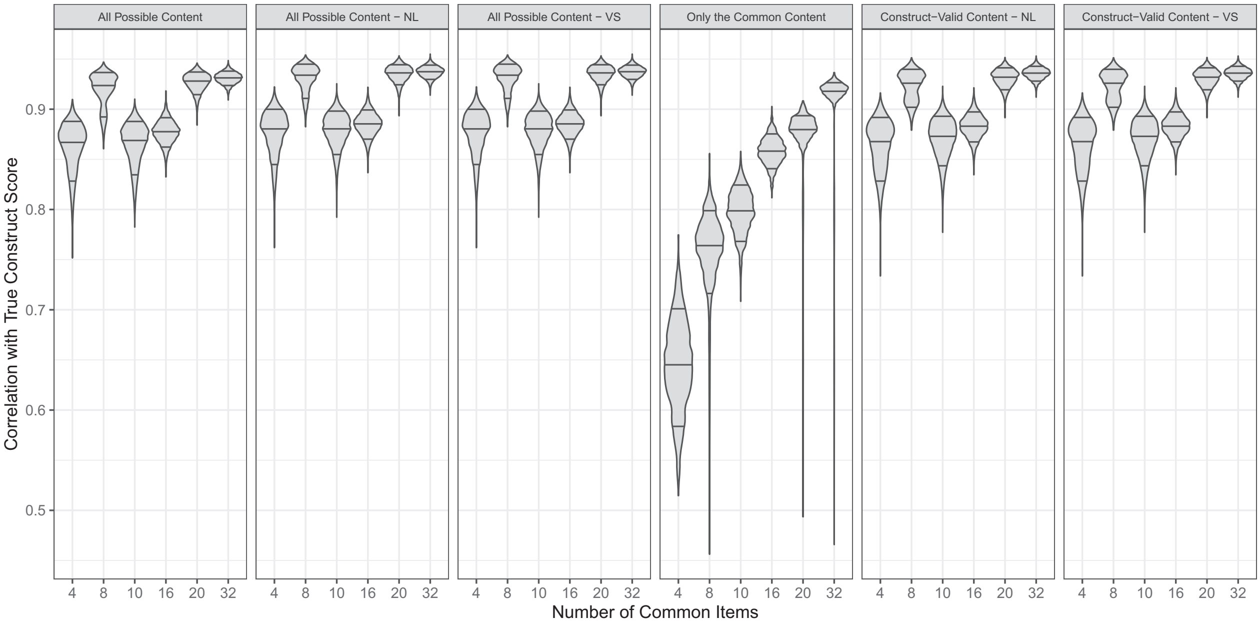


Figure 4. The figure shows the distribution of Pearson correlations in relation to the construct scores (i.e., criterion validity) by (a) the number of common items (*x*-axis) and (b) method (columns). The 10th, 50th, and 90th percentiles are represented by the horizontal lines within the violin plots. “All Possible Content” and “Only the Common Content” were item sum scores. “All Possible Content—No Linking (NL)” and “All Possible Content—Vertical Scaling (VS)” were intermediate steps that used item response theory without and with vertical scaling, respectively. “Construct-Valid Content—NL” and “Construct-Valid Content—VS” used item response theory without and with vertical scaling, respectively. “Construct-Valid Content—NL” was an intermediate step.

estimated scores from each of the six methods. We calculated the correlation at only T1 because this was the only time point when we directly generated the construct scores in the simulation procedure. The number of common items was the main driving factor that showed differences in the correlations between the methods. As shown in Figure 4, using all possible content or construct-valid content resulted in larger correlations with the true construct scores at T1 compared to using only common content. The differences in the correlation were most pronounced with the fewer number of common items, for example, 4, 8, or 10 common items—such correlations were commonly < .8 when using only common content. When the number of common items increased to 20 or 32, differences between the three methods were smaller; however, the correlations remained slightly smaller when using only common content compared to the other two methods.

The estimation methods showed a different impact of the number of common items versus the total number of items (common and unique) on criterion validity in relation to the true scores at T1. When using only common content, correlations with true scores monotonically increased as the number of common items increased from 4 to 32. By contrast, when using all possible content and construct-valid content, correlations with true scores were larger in conditions that had 40 total items (i.e., 8, 20, and 32 common items) compared to conditions that had 20 total items (i.e., 4, 10, and 16 common items), regardless of the number of common items. That is, the criterion validity of the common content approach depended more on the number of common items, whereas the criterion validity of the all possible content and construct-valid content approaches depended more on the total number of items. Finding that the criterion validity of the common content approach depended heavily on the number of common items was not surprising because the number of common items represents the total number of items that are used in the common content approach. That is, all three approaches essentially showed increases in criterion validity as the number of items used to assess the construct increased, which was not surprising because having more items can produce a more accurate and stable estimate of the construct. However, the criterion validity of the approaches that used all possible content and construct-valid content were less dependent on the number of common items, compared to the common content approach.

Bias and MAE

The average bias was very close to 0 across all simulation conditions, indicating that all methods did not systematically overestimate or underestimate people's construct scores at T1. Figure 5 shows the MAE statistics by simulation condition. On average, MAE statistics were smallest for the construct-valid content approach. This indicates that using the construct-valid content approach yielded more accurate estimates of the construct (compared to the all possible content and common content approaches). Using only common content resulted in the largest MAE statistics, indicating that it yielded the least accurate estimates of the construct score. Using all possible content with vertical scaling was equivalent to the construct-valid content approach in terms of MAE. In general, MAE decreased as the number of common items and the total number of items increased. Like the correlational results described above, the MAE of the common content approach depended more heavily on the number of common items, compared to the all possible content and construct-valid content approaches. Finally, for every condition, the unbalanced design resulted in slightly less accurate estimates compared to the balanced design. MAE was smaller when using IRT compared to item sums. This provided important validation that the IRT models accurately estimated the true scores at T1, with and without vertical scaling.

Discussion

Findings from the simulation indicated that using vertical scaling to link the construct-valid content across ages yielded more accurate trajectories, at the group-level and person-level, compared to traditional measurement approaches. This inference

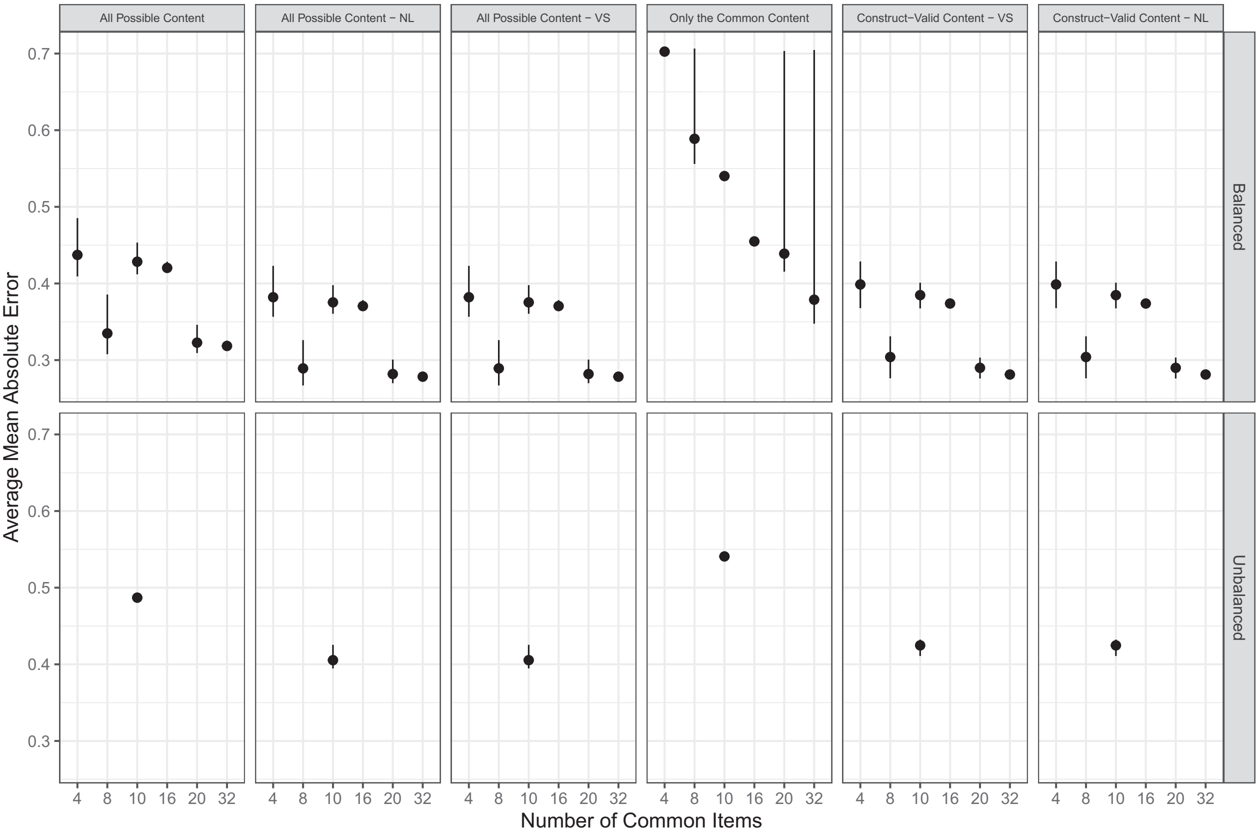


Figure 5. The figure shows the average mean absolute error (MAE) in relation to the construct scores by (a) number of common items (*x*-axis), (b) method (columns), and (c) whether the number of construct-valid items was balanced (top row) or unbalanced (bottom row). The black points indicate the average MAE, and the vertical bars around the points extend to the minimum and maximum MAE scores. "All Possible Content" and "Only the Common Content" were item sum scores. "All Possible Content—No Linking (NL)" and "All Possible Content—Vertical Scaling (VS)" were intermediate steps that used item response theory without and with vertical scaling, respectively. "Construct-Valid Content—NL" and "Construct-Valid Content—VS" used item response theory without and with vertical scaling, respectively. "Construct-Valid Content—NL" was an intermediate step.

was based on the finding that the construct-valid content approach yielded (a) group-level estimates of the standardized slope estimate that were closest to the expected trajectory (see Figure 3) and (b) higher criterion validity estimates (see Figure 4) and more accurate estimates (see Figure 5) in relation to the true scores at the person-level. Although using only common content showed generally similar trajectories at the group-level compared to the construct-valid content approach, using the construct-valid content yielded more accurate estimates at the person-level. Using only common content yielded the least accurate scores of all three approaches, in terms of MAE. Moreover, using only common content resulted in a somewhat smaller slope estimate compared to the construct-valid content approach. This finding is consistent with prior empirical findings (Petersen et al., 2018), and it suggests that the traditional approach of using only common content results in a loss of information that makes the measure less able to detect developmental change compared to using the construct-valid content.

Using all possible content resulted in less precise and more biased estimates of the slopes than the other approaches. We also found that, when the number of construct-valid items differed across ages, using all possible content could provide a completely inaccurate estimate of the slope. When the number of construct-valid items differed across ages (i.e., the unbalanced condition), using all possible content resulted in a positive slope estimate even when the true slope was negative. In addition, criterion validity and accuracy were highest when the measures had a larger number of items, and a greater proportion of the total content was a common content. Moreover, the criterion validity and accuracy of the construct-valid content approach depended less on the number of common items compared to the common content approach. In sum, the vertical scaling approach that linked the construct-valid content across ages was more accurate than the traditional measurement approaches that examined the sum of all possible content and the sum of only common content across ages.

We observed one notable exception. Using all possible content with an IRT approach to vertical scaling showed similar results to using construct-valid content in terms of standardized slopes, criterion validity, bias, and MAE. This suggests that the IRT approach to vertical scaling accurately downweighted the construct-invalid content to obtain accurate estimates when using all possible content. That is, the IRT approach to vertical scaling essentially discarded construct-invalid content from the all possible content approach. If fatigue-related measurement error had been added to the simulation, scores from all possible content would have been expected to be less reliable and accurate than scores from construct-valid content (even with vertical scaling). Moreover, the IRT approach to vertical scaling was more accurate than item sums and IRT without linking. Therefore, IRT approaches to vertical scaling may be especially useful when there are item shifts in severity and discrimination across development due to heterotypic continuity. These results show that equally weighting all content is not appropriate in the context of heterotypic continuity. Thus, observed score approaches to vertical scaling (e.g., Thurstone scaling) would not yield accurate estimates for all possible content in the context of heterotypic continuity because they equally weight content (as opposed to true score methods like IRT).

Traditionally, researchers have studied development using all possible content (Approach 1; Tong & Kolen, 2007) or only common content (Approach 2; Olson, Choe, & Sameroff, 2017). Next, we discuss these approaches more broadly and the importance of using construct-valid content at each age (Approach 3).

The first approach to assessing a construct over time uses the same, all possible content across all ages. One advantage of this approach is its comprehensive assessment of change in each content facet across the developmental span of study. It also makes interpretation of repeated assessments seemingly straightforward, because the content and mathematical metric remain consistent across time. However, the advantage of being straightforward to interpret is obviated in the case of heterotypic continuity, so this approach has key disadvantages. First, it is inefficient, requiring extra time to assess all content across all ages. Second, it could assess developmentally inappropriate or invalid content at a given age, for instance because of changes in difficulty or discrimination. In the case of heterotypic continuity, measurement should account for changes in the construct's manifestation. When the content that is construct-valid for a construct changes with development, using all possible content would include construct-*invalid* content (i.e., content intrusions) at some ages. Thus, aggregating scores on all possible content across all ages is not recommended because the measure would violate content and construct validity, have weaker internal consistency, and erroneously yield lower rank-order stability. Moreover our findings showed that using all possible content resulted in less precise and more biased estimates of the slopes, unless vertical scaling techniques were used that sufficiently downweighted the construct-valid content.

The second approach to assessing a construct over time is to use only the common content across all ages, which has the advantage that it is efficient in only assessing the same information at each assessment while retaining a consistent metric. It also may exclude developmentally inappropriate content at some ages, and permits examining consistent developmentally appropriate content (unlike using all possible content). However, using only common content has key disadvantages, especially in the case of heterotypic continuity. First, using only common content loses information because less content assesses the construct at each time, and this can make measures less able to detect developmental change (Petersen et al., 2018). Using only common content could result in systematic loss of content that reflects either very low or very high levels on the construct. Second, the measure with only common content may lack content validity because it is not assessing all facets of the construct, in particular, the age-specific manifestations. Moreover, our findings demonstrated that using only common content yielded the least accurate scores of all three approaches at the person-level.

The third approach to assessing a construct over time is to use only the construct-valid content at each age, which would mean using different content at a given age that is valid for the target construct in the context of heterotypic continuity. Using only the construct-valid content has several drawbacks. First, it is more time-intensive than using only the common content across all ages. Second, using different measures across time poses a challenge longitudinally, because one goal in longitudinal studies is to describe continuity and change. Developmental inferences are strengthened by establishing measurement invariance (equivalent measures), ensuring that differences over time reflect changes in the phenomenon of interest, not changes in its measurement. As explained, longitudinal assessment of constructs showing heterotypic continuity often requires different measures at different ages. The use of different measures at different ages violates measurement invariance in the strictest sense (same measure, same meaning) and calls into question the comparability of scores across time. Thus, statistical approaches along with theoretical and empirical considerations are necessary to link the different measures on the same conceptual and mathematical metric across time. Yet, to maintain construct validity invariance, developmental theory requires that we use measures that account for changes in the manifestation of a construct (Knight & Zerr, 2010; Widaman et al., 2010).

Because of the importance of accounting for changes in a construct, using the construct-valid content at each age has several key advantages. First, it retains content validity and construct validity invariance. Second, it is more efficient than using all possible content across all ages. Moreover, our findings showed that using the construct-valid content yielded the most accurate results at the group- and person-level. Thus, in the case of heterotypic continuity, using the construct-valid content at each age is the recommended approach. We extend this recommendation by suggesting that researchers account for heterotypic continuity by using vertical scaling to put measures on the same developmental scale. Researchers have used different measures across ages and vertical scaling to examine developmental trajectories across the life span (McArdle & Grimm, 2011; McArdle et al., 2009).

The study had several key strengths. First, we conducted a simulation where the truth was known. This allowed us to compare the three approaches to assessing a construct over time. Second, we examined multiple accuracy metrics including accuracy at the group- and person-level. The approach of using the construct-valid content was more accurate than the other approaches at the group- and person-level, which provides greater confidence in the findings. Third, we conducted intermediate steps to identify the specific reason why the construct-valid content approach was the most accurate (i.e., it used an IRT approach to vertical scaling).

The study also had weaknesses. Although we selected adjustments to the severity and discrimination item parameters based on what has been observed in prior research, future research will benefit from specifying the expected level of change for particular content of different constructs. Furthermore, it will be important for future empirical work to compare the three approaches. Nevertheless, our findings are consistent with prior empirical work that has compared the construct-valid content approach to the common content approach (Petersen et al., 2018), which provides greater confidence in the findings.

Conclusion: Implications for Developmental Psychology

Studies in developmental psychology have largely failed to account for the changing manifestation of constructs in ways that detect meaningful growth. This was the first study to compare the three approaches to assessing a construct over time (i.e., all possible content, only common content, and only construct-valid content). The simulation indicated that item sums of all possible content or only common content provide less accurate estimates of people's levels on the construct and people's trajectories than vertical scaling of the construct-valid content. These findings are consistent with evidence that failing to account for heterotypic continuity results in inaccurate developmental inferences (Chen & Jaffee, 2015; Petersen et al., 2018).

A key goal of developmental psychology is to understand developmental pathways across the life span, and not just limited windows of time or stages in people's lives. By paying greater attention to how constructs change in their expression with development and accounting for heterotypic continuity when it exists, researchers can study development across the life span without violating construct validity invariance. Accounting for heterotypic continuity may require creating a developmental scale using different measures across time with vertical scaling approaches that link different measures on a comparable mathematical metric, rather than traditional item sums. Measurement approaches that accommodate constructs' changes over time are essential for making accurate developmental inferences, especially over lengthy spans of time. Our study provides promising evidence that vertical scaling accounts for heterotypic continuity and is a more accurate measurement approach than traditional approaches. Given how common heterotypic continuity is among psychological constructs, accounting for heterotypic continuity is crucial to advance our understanding of development across the life span.

References

Achenbach, T. M. (2009). *Achenbach System of Empirically Based Assessment (ASEBA): Development, findings, theory, and applications*. Burlington, VT: University of Vermont, Research Center of Children, Youth & Families.

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

Caspi, A., & Shiner, R. L. (2006). Personality development. In N. Eisenberg, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology* (Vol. *3*, 6th ed., pp. 300–365). Hoboken, NJ: Wiley.

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

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

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

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.

Eyberg, S. M., & Pincus, D. (1999). *Eyberg Child Behavior Inventory & Sutter-Eyberg Student Behavior Inventory–revised: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed.). Hoboken, NJ: Wiley.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144–149. https://doi.org/10.4992/psycholres1954.22.144

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*, 3–24. https://doi.org/10.1177/0146621602026001001

Harwell, M. R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement*, *15*, 375–389. https://doi.org/10.1177/014662169101500409

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Praeger.

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

Kim, S., & Lee, W.-C. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement*, *43*, 53–76. https://doi.org/10.1111/j.1745-3984.2006.00004.x

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

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

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

LeBeau, B. (2017). Ability and prior distribution mismatch: An exploration of common-item linking methods. *Applied Psychological Measurement*, *41*, 545–560. https://doi.org/10.1177/0146621617707508

LeBeau, B. (2018). *Misspecification of the random effect structure: Implications for the linear mixed model*. Iowa Research Online. Retrieved from https://ir.uiowa.edu/pq_pubs/2

LeBeau, B., Song, Y. A., & Liu, W. C. (2018). Model misspecification and assumption violations with the linear mixed model: A meta-analysis. *SAGE Open*, *8*, 1–16. https://doi.org/10.1177/2158244018820380

Lee, K., Bull, R., & Ho, R. M. H. (2013). Developmental changes in executive functioning. *Child Development*, *84*, 1933–1953. https://doi.org/10.1111/cdev.12096

McArdle, J. J., & Grimm, K. J. (2011). An empirical example of change analysis by linking longitudinal item response data from multiple tests. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 71–88). New York, NY: Springer Science & Business Media.

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods*, *14*, 126–149. https://doi.org/10.1037/a0015857

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

Oleson, J. J., Cavanaugh, J. E., Tomblin, J. B., Walker, E., & Dunn, C. (2016). Combining growth curves when a longitudinal study switches measurement tools. *Statistical Methods in Medical Research*, *25*, 2925–2938. https://doi.org/10.1177/0962280214534588

Olson, S. L., Choe, D. E., & Sameroff, A. J. (2017). Trajectories of child externalizing problems between ages 3 and 10 years: Contributions of children's early effortful control, theory of mind, and parenting experiences. *Development and Psychopathology*, *29*, 1333–1351. https://doi.org/10.1017/S095457941700030X

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

Petersen, I. T., Bates, J. E., Dodge, K. A., Lansford, J. E., & Pettit, G. S. (2016). Identifying an efficient set of items sensitive to clinical-range externalizing problems in children. *Psychological Assessment*, *28*, 598–612. https://doi.org/10.1037/pas0000185

Petersen, I. T., & LeBeau, B. (in press). Language ability in the development of externalizing behavior problems in childhood. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000461.

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

Petscher, Y., Justice, L. M., & Hogan, T. (2018). Modeling the early language trajectory of language development when the measures change and its relation to poor reading comprehension. *Child Development*, *89*, 2136–2156. https://doi.org/10.1111/cdev.12880

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Shavelson, R. J., Webb, N. M., & Rawley, R. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922–932. https://doi.org/10.1037/0003-066X.44.6.922

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

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*, 39–49. https://doi.org/10.1177/014662169501900105

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*, 227–253. https://doi.org/10.1080/08957340701301207

Wang, S., Jiao, H., & Zhang, L. (2013). Validation of longitudinal achievement constructs of vertically scaled computerised adaptive tests: a multiple-indicator, latent-growth modelling approach. *International Journal of Quantitative Research in Education*, *1*, 383–407. https://doi.org/10.1504/IJQRE.2013.058307

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

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*, 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website:

Figure S1. The Figure Shows the Distribution of Mean Differences Between Adjacent Time Points by (a) Change in Six Severity of the Common Content (*x*-Axis), (b) The Estimation Method (Columns), and (c) Mean Difference Between T2 and T1 (Top Row) or Mean Difference Between T3 and T2 (Bottom Row)

Table S1. Example Items for Hypothetical Measures of Externalizing Problems

Appendix S1. Sensitivity Analysis

Supplementary Appendix S1. Sensitivity analysis.

We conducted a sensitivity analysis using maximum likelihood estimates of the latent construct to ensure the estimator we used (expected a posteriori, EAP) was the most stable estimator. EAP estimates are Bayesian, which include prior distributions, and tend to be biased, particularly for extreme values of the latent construct and strong prior distributional assumptions (Kim, Moses, & Yoo, 2015; Kolen & Tong, 2010; Nicewander & Schulz, 2015). The impact this may have on the present study is that the EAP-derived standardized slope estimates may regress toward the mean or be underestimated. Using the vertical scaling approach with construct-valid content, we calculated the mean differences of the standardized slope estimates compared to the true slope estimates, for both the EAP and maximum likelihood estimates (see Supplementary Figure S1). As expected, the maximum likelihood mean difference estimates were larger compared to the EAP mean difference estimates. These differences were most pronounced when the change in the severity of the common content was larger. The EAP mean difference estimates were more stable and provided estimates that removed some of the extreme values shown in the maximum likelihood estimates, consistent with common findings regarding the estimation methods (Kim et al., 2015; Kolen & Tong, 2010; Nicewander & Schulz, 2015). Moreover, maximum likelihood-derived factor scores are unable to be estimated for people who endorse positively or negatively all items at each time point (i.e. they have a summed score of 0 or equal to all of the items on the survey). Thus, findings suggest that we used the most stable estimator for the present study.

References

Kim, S., Moses, T., & Yoo, H. H. (2015). Effectiveness of item response theory (IRT) proficiency estimation methods under adaptive multistage testing. *ETS Research Report Series, 2015*, 1-19. doi: 10.1002/ets2.12057

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice, 29*, 8-14. doi: 10.1111/j.1745-3992.2010.00179.x

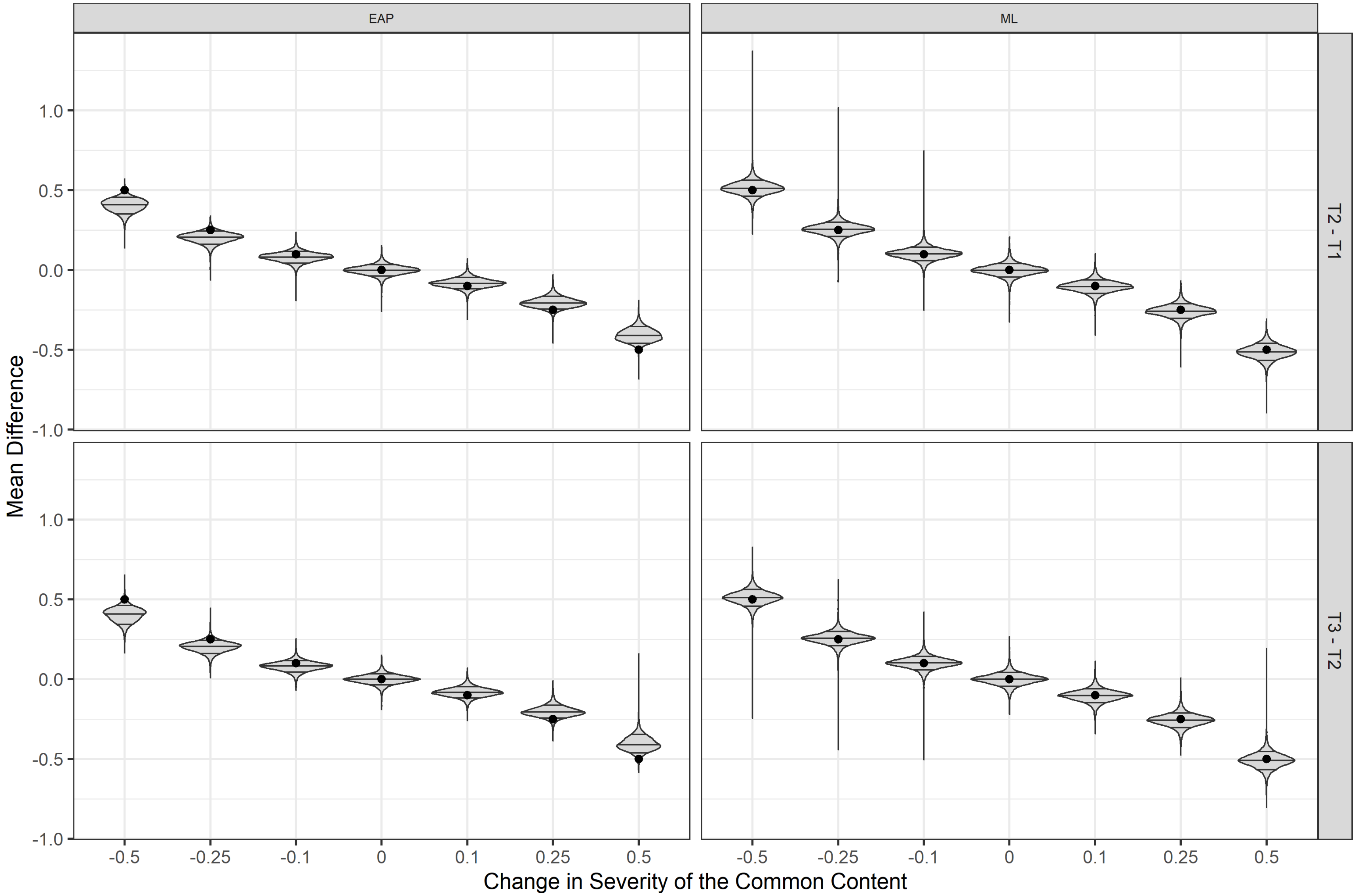
Nicewander, W. A., & Schulz, E. M. (2015). A comparison of two methods for computing IRT scores from the number-correct score. *Applied Psychological Measurement, 39*, 643-655. doi: 10.1177/0146621615601081

Supplementary Table S1. Example items for hypothetical measures of externalizing problems.

Sub-Domain	Item	Early Childhood (age 3 years)	Middle Childhood (age 8 years)	Early Adolescence (age 13 years)
Physical aggression	Temper tantrums	x		
Physical aggression	Destroys own things	x	x	
Physical aggression	**Destroys others' things**	**x**	**x**	**x**
Physical aggression	Bites others	x	x	
Physical aggression	Scratches or pinches others	x	x	
Physical aggression	Spits on others			x
Physical aggression	Hurts or tortures animals		x	x
Physical aggression	**Physically fights other people**	**x**	**x**	**x**
Physical aggression	**Throws objects when upset**	**x**	**x**	**x**
Verbal aggression	**Insults other people/name calling**	**x**	**x**	**x**
Verbal aggression	Threatens others		x	x
Verbal aggression	Screams	x		
Impulsivity/Disinhibition	**Does things without thinking**	**x**	**x**	**x**
Impulsivity/Disinhibition	Explosive behavior	x		
Impulsivity/Disinhibition	Can't tolerate waiting	x	x	
Impulsivity/Disinhibition	**Talks out of turn**	**x**	**x**	**x**
Oppositionality/Defiance	**Disobeys others**	**x**	**x**	**x**
Oppositionality/Defiance	**Argues with others**	**x**	**x**	**x**
Oppositionality/Defiance	**Disrupts others**	**x**	**x**	**x**
Inattention	**Can't concentrate/focus**	**x**	**x**	**x**
Inattention	**Highly distractible**	**x**	**x**	**x**
Hyperactivity	**Can't sit still**	**x**	**x**	**x**
Hyperactivity	Constantly on the move	x	x	
Rule breaking	Truant			x
Rule breaking	Breaks curfew			x

Rule breaking	Breaks rules in school		x	x
Rule breaking	**Lies or cheats**	**x**	**x**	**x**
Bullying	Repeated aggression and intimidation toward a specific individual		x	x
Relational aggression	Spreads rumors about people who cause them trouble		x	x
Relational aggression	Stonewalls or actively ignores/excludes others		x	x
Delinquent behavior	Robbery			x
Delinquent behavior	Vandalizes property		x	x
Delinquent behavior	**Steals**	**x**	**x**	**x**
Delinquent behavior	Starts fires		x	x
Delinquent behavior	Weapon carrying or use			x
Substance use	Uses alcohol or illegal drugs			x
Substance use	Sells illegal drugs			x
Sexual aggression	Physically forces someone to have sex against their will			x
Sexual aggression	Threatens someone to have sex			x

Note: “x” denotes that we consider the item valid for the construct of externalizing problems at that particular age. Items in bold are the common content (i.e., construct-valid at all three ages). Non-bolded items are the unique (non-common) content that reflect the age-specific manifestations of the construct of externalizing problems.



Supplementary Figure S1. The figure shows the distribution of mean differences between adjacent time points by (a) change in severity of the common content (x-axis), (b) the estimation method (columns), and (c) mean difference between T2 and T1 (top row) or mean difference between T3 and T2 (bottom row). The 10th, 50th, and 90th percentiles are represented by the horizontal lines within the violin plots. The black points shown on the figure represent the "true" expected slope change based on the inverse of the change in severity of the common content.