Developmental Review 58 (2020) 100935

Contents lists available at ScienceDirect

Developmental Review

journal homepage: www.elsevier.com/locate/dr

Studying a moving target in development: The challenge and opportunity of heterotypic continuity

Isaac T. Petersen [a,*], Daniel Ewon Choe [b], Brandon LeBeau [c]

[a] Department of Psychological and Brain Sciences, University of Iowa, 340 Iowa Ave., 175 Psychological and Brain Sciences Building, Iowa City, IA 52242, United States
[b] Human Development and Family Studies, University of California, Davis, One Shields Ave., 1312 Hart Hall, Davis, CA 95616, United States
[c] Educational Measurement and Statistics, University of Iowa, 600 Blank Honors Center, Iowa City, IA 52242, United States

ARTICLE INFO

Keywords:
Heterotypic continuity
Changing measures
Longitudinal
Construct validity invariance
Developmental scaling
Factorial invariance

ABSTRACT

Many psychological constructs show heterotypic continuity—their behavioral manifestations change with development but their meaning remains the same (e.g., externalizing problems). However, research has paid little attention to how to account for heterotypic continuity. Conceptual and methodological challenges of heterotypic continuity may prevent researchers from examining lengthy developmental spans. Developmental theory requires that measurement accommodate changes in manifestation of constructs. Simulation and empirical work demonstrate that failure to account for heterotypic continuity when collecting or analyzing longitudinal data results in faulty developmental inferences. Accounting for heterotypic continuity may require using different measures across time with approaches that link measures on a comparable scale. Creating a developmental scale (i.e., developmental scaling) is recommended to link measures across time and account for heterotypic continuity, which is crucial in understanding development across the lifespan. The current synthesized review defines heterotypic continuity, describes how to identify it, and presents solutions to account for it. We note challenges of addressing heterotypic continuity, and propose steps in leveraging opportunities it creates to advance empirical study of development.

Introduction

Developmental science seeks to elucidate processes of continuity and change across the lifespan and not just transitory outcomes at 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 but not in their underlying meaning or function, a phenomenon known as heterotypic continuity. This poses challenges for measurement. Assessments may not capture a construct's features in the same way across time or between groups, contributing to measurement non-invariance, which has the potential to negatively impact both cross-sectional and longitudinal studies. Yet, there is no authoritative guide on what heterotypic continuity is and how to address it when studying development. This synthesized review defines heterotypic continuity, describes how to identify it, and presents solutions to account for

* Corresponding author at: Department of Psychological and Brain Sciences, University of Iowa, 175 Psychological and Brain Sciences Building, Iowa City, IA 52242, United States.
E-mail address: isaac-t-petersen@uiowa.edu (I.T. Petersen).

https://doi.org/10.1016/j.dr.2020.100935
Received 15 April 2020; Received in revised form 10 September 2020;

it. Greater attention to streamlining theoretical, statistical, and empirical issues of measurement relating to heterotypic continuity is necessary to advance the study of development across the lifespan.

Consider the construct of externalizing problems—challenging behaviors including aggression, impulsivity, and other "problems that mainly involve conflicts with other people and with their expectations for the child." (Achenbach & Rescorla, 2001, p. 24). Externalizing problems are often expressed as overt acts in early childhood, such as physical aggression, but externalizing problems are more often expressed covertly in adolescence as indirect or relational forms of aggression, rule-breaking, and drug use (Miller, Vaillancourt, & Boyle, 2009). When the externalizing construct changes in behavioral expression across development, different measures should be used at different ages to retain each measure's construct validity. Among two widely used measures of externalizing problems, the Child Behavior Checklist (Achenbach, 2009) and the Eyberg Child Behavior Inventory (ECBI; Eyberg & Pincus, 1999), only the former uses different item content across ages to capture the construct's changing manifestation across development. 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 a measure of biting could primarily reflect aggression at one timepoint but could primarily reflect neurodevelopmental disorder at a later timepoint).

Although we present measurement challenges with the example of externalizing problems, considerable research has demonstrated many psychological constructs' changing behavioral expression with development (e.g., Schulenberg & Maslowsky, 2009). 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 developmental growth using different measures with statistical and measurement tools to link the different measures across time to maintain construct validity when the construct changes in its

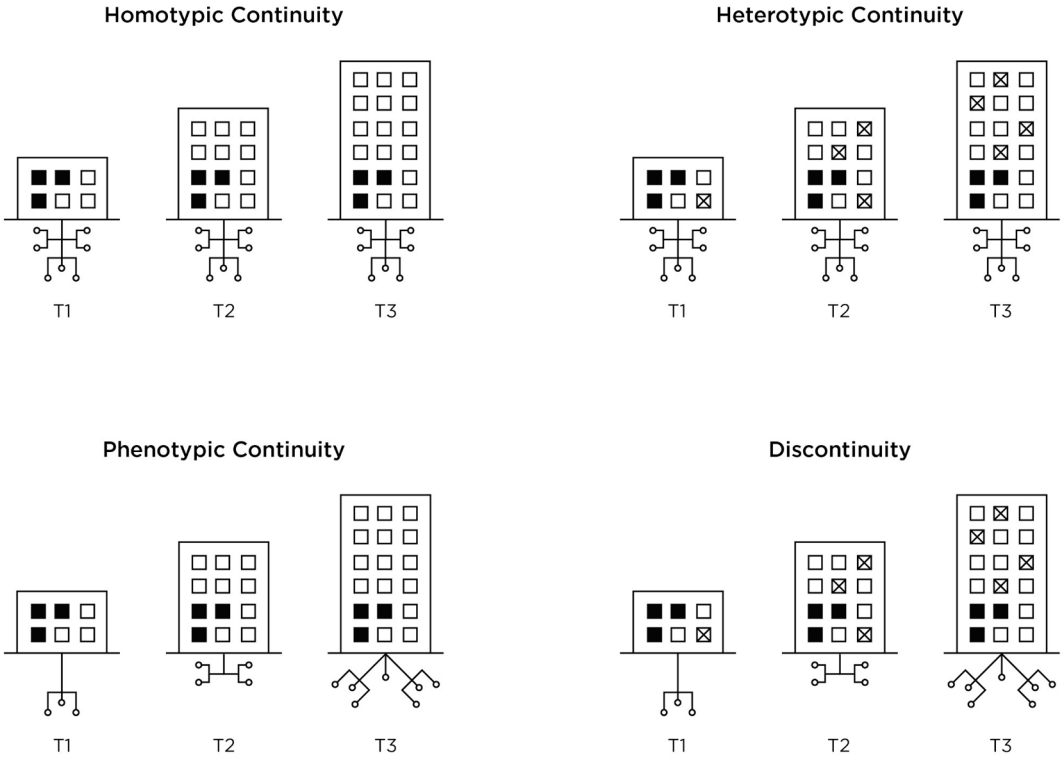

Fig. 1. Depiction of the three types of continuity in addition to discontinuity in the form of a 2 (behavioral manifestation, underlying processes) × 2 (same versus different across time) Latin square. "T1", "T2", and "T3" reflect time points 1, 2, and 3, respectively. The illustrations above the lines are buildings, representing the surface structure (i.e., behavioral manifestation). The illustrations below the lines depict the underlying processes supporting the buildings at each time point. The squares on the buildings are windows. The black windows represent content facets that are active across all time points (i.e., age-common content). The windows that contain X's represent content facets that are active at some but not all time points (i.e., age-unique content). The white windows represent content facets that are inactive, and therefore are not part of the construct at that time point. The increasing size of the buildings at later time points reflects growth with development. The top row of the Latin square involves the same underlying processes across time, whereas the bottom row involves different underlying proccesses across time. The left column of the Latin square involves the same behavioral manifestation across time, whereas the right column involves a different behavioral manifestation across time. Homotypic continuity (top left) describes the same behaviorial manifestation with the same underlying process (i.e., construct) across development. Heterotypic continuity (top right) describes the same underlying process with a different behaviorial manifestation across development. Phenotypic continuity (or functional discontinuity) (bottom left) describes the same behavior with different underlying processes across development. Discontinuity (bottom right) describes different behavioral manifestations with different underlying processes across development. Thus, in both homotypic continuity and heterotypic continuity, the active content facets reflect the same construct or underlying process across time, whereas in phenotypic continuity and discontinuity, the active content facets do not reflect the same construct across time.

behavioral expression. Thus, there is an inconsistency between theory of how constructs develop and approaches to assess and study their development. Measurement considerations are crucial to ensuring the validity of developmental inferences, as reflected in the maxim, "What we know depends on how we know it." Accounting for changes in the behavioral manifestation of constructs can improve the construct validity of measurement strategies and the accuracy of developmental inferences. To date, there has not been a review of research on heterotypic continuity with guidelines for how to identify and address it when studying development.

In the first part of this review, we discuss the historical origins, conceptual definitions, and operationalizations of heterotypic continuity (i.e., how to identify changes in the manifestation of constructs). We also discuss the importance of accounting for changes in the manifestation of constructs, especially for modeling developmental trajectories. In the second part, we discuss theoretical and methodological approaches to accounting for changes in the manifestation of constructs when modeling developmental trajectories by ensuring that different measures are comparable and assess the same construct over time. In the third and final part, we address common questions about measurement and heterotypic continuity. For simplicity and brevity, we refer to birth age as a proxy for developmental time (sometimes synonymously).

Heterotypic continuity

Historical overview

Heterotypic continuity is a developmental phenomenon that reduces the likelihood that a measure has the same meaning over time. The concept of heterotypic continuity has a long and venerable history in developmental science, and is often attributed to Kagan and colleagues (Kagan, 1969, 1971, 1980; Kagan & Moss, 1962). Kagan (1969) described three types of continuity: (a) "complete continuity" (more commonly called homotypic continuity)—when both psychological processes (e.g., reasons, motives, standards, sources of anxiety, or expectancies) and the form of behavior remain the same, (b) "phenotypic continuity" (also called functional discontinuity; Schulenberg & Zarrett, 2006)—when the form of behavior remains the same but not the underlying psychological processes, and (c) heterotypic continuity (also called genotypic continuity)—when the underlying psychological processes remain the same but the form of behavior changes. Historically, many researchers have discussed similar concepts, including Emmerich ("developmental transformation" and "structural development"; 1964, 1968), Yarrow and Yarrow ("dynamic continuity"; 1964), Coan ("factor metamorphosis"; 1966), Baltes and Nesselroade ("structural" change; 1970), Chomsky ("surface structure" vs. "deep structure"; 1971), Bell et al. ("isomorphic continuity" vs. "metamorphic continuity"; 1971), Livson (1973), Buss ("qualitative change"; 1973, 1974), Buss and Royce ("developmental transitions"; 1975), Sroufe ("lawful" change; 1979), Sroufe and Jacobvitz ("coherence"; 1989), and Patterson ("orderly change"; 1993). The three types of continuity (in addition to discontinuity) are illustrated in Fig. 1. The figure depicts the four types of (dis)continuity in the form of a 2 (behavioral manifestation, underlying processes) × 2 (same or different across time) Latin square. In the figure, the illustrations above the lines are buildings, representing the surface structure (i.e., behavioral manifestation). The illustrations below the lines depict the underlying processes supporting the buildings at each time point. The windows represent content facets (see the figure caption for a more detailed description).

A construct shows homotypic continuity when the behavioral manifestation of a construct remains stable across development and can thus be assessed the same way across time. For example, physical growth in height and weight can be assessed using the same ruler and scale, respectively, across time.

According to Kagan, an example of phenotypic continuity or functional discontinuity (i.e., when different psychological reasons underlie the same behavior at different ages) is crying in infancy versus in childhood. Eight-month-olds typically cry when they are hungry or encounter a stimulus that violates their expectations, whereas 8-year-olds typically cry when they want to escape parental restrictions, fear being harmed, or anticipate punishment. Although the behavior (e.g., tears, facial grimaces) is the same, its meaning can differ psychologically at each age. More often than not, an individual's developmental context and experience determine the specific behavior's overt manifestation and underlying psychological meaning (Sroufe, 2013).

Heterotypic continuity, by contrast, occurs when the same psychological reasons underlie different behaviors at different ages—the same construct "looks different" across development. Heterotypic continuity refers to the persistence of an underlying latent construct with behavioral manifestations that change across development (Caspi & Shiner, 2006; Cicchetti & Rogosch, 2002). In the seminal Fels Longitudinal Study, Kagan and Moss (1962) observed that girls who had frequent tantrums at 6 to 10 years of age tended to become women who were more motivated in school, less dependent on others, and more masculine in their interests than women who had fewer tantrums as children. They interpreted the different behaviors at different ages as stemming from the same psychological process: a tendency to avoid adopting "female sex-role standards" (p. 200). Kagan described such findings, when psychological processes remain the same but the form of behavior changes, as examples of heterotypic continuity (Kagan, 1969, 1971, 1980). Heterotypic continuity is analogous to the transformation of water to ice or steam, or of a caterpillar to a butterfly. An underlying core is preserved, but observable manifestations change.

In addition to important historical perspectives, there has been considerable contemporary theoretical discussion on heterotypic continuity (e.g., Caspi & Shiner, 2006; Cicchetti & Rogosch, 2002; Moffitt, 1993; Schulenberg, Patrick, Maslowsky, & Maggs, 2014; Schulenberg & Zarrett, 2006; Weiss & Garber, 2003). Heterotypic continuity is an important issue in the theoretical perspective and study of developmental psychopathology (Rutter & Sroufe, 2000; Sroufe & Rutter, 1984). The framework asserts that our understanding of typical development is informed by understanding atypical development and vice versa (Sroufe, 2013). Similarly, understanding of heterotypic continuity may clarify how and why specific behaviors (e.g., physical aggression) are normative at some ages and deviant at others, as well as the unique developmental trajectories of these behaviors' sub-dimensions (e.g., aggression from hostility, impulsivity, or fear). Although some conceptualizations of heterotypic continuity emphasize genetics (e.g., Caspi & Shiner,

2006; Kagan, 1969), environmental factors are also crucial to consider in heterotypic continuity (Schulenberg et al., 2014). Environmental factors contribute to heterotypic continuity by perpetuating psychological processes across development (Sroufe, 2009). Moreover, age-related experiences (e.g., school entry, transition from reliance on caregivers to closer affiliation with peers) can lead to different behavioral manifestations of the same psychological processes at different ages (Patterson, 1993). Genetic and environmental processes are also intertwined, so their co-actions likely contribute to heterotypic continuity.

A developmentally relevant example of heterotypic continuity is Patterson's (1993) metaphor for antisocial behavior: a chimera with a body of a goat that, with development, grows the head of a lion and then a tail of a snake. This developmental pattern is consistent with Blumberg's (2013) description of a construct that shows elaboration and integration. Patterson (1993) describes antisocial behavior like a chimera in which the underlying essence of oppositionality is maintained while adding more mature features, including the ability to inflict serious damage with violence and covert acts of relational aggression.

Identifying heterotypic continuity in development

Considerable research has examined how psychological constructs change in behavioral expression with development. A search on Google Scholar in March 2020 for "heterotypic continuity" elicited over 2,200 publications, suggesting that many psychological constructs demonstrate heterotypic continuity. Heterotypic continuity has been examined at multiple psychometric levels, including content and construct levels. 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). Even though we refer to content as a singular term, in some cases content applies to multiple facets. Examining continuity of a construct at its content level assesses the extent to which the construct's content (e.g., individual behaviors, behavioral tasks, questionnaire items) shows cross-time stability in three ways.

First, one can consider whether individual differences show the same degree of stability across time when assessed with particular content; that is, whether the content shows the same magnitude of rank-order stability of people across time. The degree to which the content shows cross-time changes in the magnitude of rank-order stability of people can be examined with correlation or regression of people's scores on the content across time. For instance, if the content shows the same magnitude of rank-order stability of people from T1 to T2 (e.g., $r = 0.7$) as from T2 to T3 (e.g., $r = 0.7$), individual differences show the same degree of stability when assessed on the content across T1 to T3.

Second, one can consider whether a person's score, as assessed by particular content, corresponds to the same level on the construct across time; that is, whether the content shows cross-time stability in its level on the construct. The content's level on a construct is similar to the inverse of the content's frequency or base rate (Fan, 1998).[1] The content's cross-time stability in level on the construct can be examined using item response theory (IRT) or factor analysis. The content's level on a construct, such as the item "sets fires" on the externalizing construct, is called difficulty or severity in IRT. In (two-parameter) IRT, an item's difficulty parameter describes the level on a construct 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, et al., 2016). Thus, "sets fires" is more infrequent, severe, and has a higher difficulty parameter (i.e., level on the construct) than "argues" for externalizing problems. In factor analysis, the frequency of a behavior is based on a measure's mean score or intercept, so the behavior's level on the construct would be the inverse of this mean or intercept.

Third, one can consider whether people's scores, as assessed by particular content, show the same association with the construct across time; that is, whether the content shows cross-time stability 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, or can be assessed with factor loadings in factor analysis. 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, how often a child attacks people is more relevant to externalizing problems than how often a child brags, so the item "attacks people" has a higher discrimination parameter or factor loading, and thus construct validity, for externalizing problems than the item "brags" (Petersen, Bates, et al., 2016). Examining how strongly 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 all content facets show 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, evidence suggests that the construct shows a stable factor structure across time. To the extent that the construct shows a stable factor structure across ages, the construct shows *homotypic* continuity. To the extent that the factor structure of the construct changes across development, age-differing content facets show functional discontinuity (i.e., changes in meaning over time) and the higher-order construct shows heterotypic continuity. Thus, we empirically define heterotypic continuity as changes in a construct's factor structure across development. Changes in a construct's factor structure could include cross-time changes in any of the content facets' (1) magnitude of rank-order stability of people (based on correlation or regression across time), (2) level on the construct (based on difficulty in IRT or intercepts in factor analysis), and/or (3) how strongly the content

[1] Note that this deals with a *content*'s level on a construct, where a content that shows lower frequency has a higher level on the construct. By contrast, a *person* with a higher score on the content (who engages in the behavior more frequently) has a higher level on the construct.

facets reflect the construct (based on discrimination in IRT or factor loadings in factor analysis). Continuity of the construct is implicit in this definition because it refers to the same construct across time (despite behavioral manifestations that change across development), which would require some degree of cross-time stability of individual differences and the construct's underlying processes or functions.

The heterotypic continuity of externalizing problems is a clear example of how behaviors of a construct change in manifestation. The sub-dimensions of externalizing problems show different prototypical trajectories in terms of frequency (Olson et al., 2013). For example, physical aggression decreases in frequency from early childhood to adolescence, whereas relational aggression increases across that time span (Miller et al., 2009). Thus, physical aggression would have greater difficulty (i.e., severity, infrequency, or level on the construct) in adolescence than in early childhood. In addition, specific aggressive behaviors are known to change in how strongly they reflect the construct of externalizing problems (i.e., they change in discrimination or factor loadings, and thus, construct validity). Threatening other people is more strongly associated with externalizing problems in adolescence than in early childhood (Lubke, McArtor, Boomsma, & Bartels, 2018). Content has also shown changes in the magnitude of rank-order stability of people, such as increases in the rank-order stability of inattention reflecting more stable individual differences from preschool to school age (Arnett et al., 2012).

Other research has examined the continuity of constructs at the higher-order construct level (Ferdinand, Dieleman, Ormel, & Verhulst, 2007; Lahey, Zald, Hakes, Krueger, & Rathouz, 2014; Lavigne, Gouze, Bryant, & Hopkins, 2014; Miller et al., 2009; Nagin & Tremblay, 2001; Putnam, Rothbart, & Gartstein, 2008; Snyder, Young, & Hankin, 2017). Examining continuity of constructs at the higher-order level assesses the extent to which people's scores on the construct (e.g., a latent syndrome) show rank-order stability across time, versus whether the construct leads to (or predicts) different constructs over time. For instance, anxiety often precedes and predicts depression (Garber & Weersing, 2010), which suggests that anxiety may change in manifestation across development for some people. That is, anxiety and depression may stem from an underlying common liability of negative emotionality (i.e., distress and fear; Eaton et al., 2013), as reflected in the empirically derived internalizing spectrum (Achenbach, 2009). The internalizing spectrum is part of transdiagnostic, hierarchical models of psychopathology such as the p-factor model (Caspi et al., 2014) and HiTOP (Kotov et al., 2017). Thus, as an example of sequential comorbidity, internalizing problems may show heterotypic continuity such that some forms of anxiety change in behavioral expression to also feature depressive symptoms with development (Schleider, Krause, & Gillham, 2014). However, an association between a construct in predicting another construct does not establish heterotypic continuity because it could reflect a spurious association rather than a causal effect. For instance, continuing environmental and contextual factors—rather than a person's inherent disposition—sometimes explain the observed rank-order correlation of different behaviors at different ages (Kagan, 1980). Thus, experimental manipulation may be necessary to establish developmental pathways and heterotypic continuity (Pickles & Hill, 2006).

In sum, extensive research in developmental science demonstrates how many constructs change in expression over time and show heterotypic continuity. Researchers have investigated the heterotypic continuity of internalizing problems (Petersen et al., 2018; Weems, 2008; Weiss & Garber, 2003), externalizing problems (Chen & Jaffee, 2015; Miller et al., 2009; Moffitt, 1993; Patterson, 1993; Petersen, Bates, Dodge, Lansford, & Pettit, 2015; Petersen & LeBeau, in press; Wakschlag, Tolan, & Leventhal, 2010), inhibitory control (Petersen et al., under revise and resubmit; Petersen, Hoyniak, et al., 2016), sleep states (Blumberg, 2013), substance use (Schulenberg & Maslowsky, 2009), and temperament (Putnam et al., 2008). Constructs that are often assessed with different measures across time to maintain developmental relevance also demonstrate heterotypic continuity. For instance, cognitive skills such as language (Petscher, Justice, & Hogan, 2018), nonverbal ability (McArdle & Grimm, 2011), working memory (McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009), and academic skills (Tong & Kolen, 2007) change in manifestation with development. Moreover, Achenbach and Edelbrock (1983) famously developed the Child Behavior Checklist and other measures to be consistent with the heterotypic continuity of behavior problems.

Important gaps in developmental research due to heterotypic continuity

Despite considerable theoretical discussion and empirical identification of heterotypic continuity, the field lacks awareness of appropriate methodological and statistical schemes that account for such changes when examining how people develop. Researchers have stated that the measurement of constructs that show heterotypic continuity should receive greater attention (e.g., Schulenberg & Maslowsky, 2009); yet, surprisingly little research has considered how to *account for* heterotypic continuity when examining people's developmental trajectories in constructs that change in manifestation over time.

Thus, many developmental scientists acknowledge the importance of heterotypic continuity but few address it when examining people's trajectories. Despite a proliferation of studies that examine growth trajectories, 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 et al., 2018), and even fewer have done so in ways that allow researchers to examine absolute change rather than just relative, rank-order change (McArdle et al., 2009; Petersen & LeBeau, in press; Petersen et al., 2018). This is a major problem for the field because the study of development 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 its manifestation over time and the measures do not accommodate these changes, the measures lack 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. Many measures span wide age ranges without changing their content across ages even

when the construct shows heterotypic continuity (e.g., the ECBI measure of externalizing problems spans 2 to 16 years of age; Eyberg & Pincus, 1999). 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). Simulations show that failing to account for heterotypic continuity results in less accurate growth estimates, at the person- and group-levels (Petersen et al., in press). Thus, heterotypic continuity is a characteristic of many psychological constructs, but widespread failure to account for heterotypic continuity is a serious problem in developmental science because it jeopardizes the validity of our measures, data, and inferences.

There are legitimate reasons why researchers have often not addressed heterotypic continuity when examining developmental trajectories. It is easier to claim that one is assessing a construct on the same mathematical scale (i.e., same range of possible scores) if one uses the same measure across time rather than using different measures. Nevertheless, using the same measure across time and failing to account for heterotypic continuity violates construct validity invariance. Thus, even if the measures are on the same *mathematical* metric, they may not be on the same *conceptual* metric—that is, the same score at each age has a different meaning and reflects different levels of the construct. Instead of ignoring heterotypic continuity or treating it as a "nuisance," advancing our understanding of it and accounting for it can help elucidate development. It is both a methodological and theoretical challenge, as well as a theoretical and empirical opportunity to understand and account for heterotypic continuity in development.

Accounting for heterotypic continuity in development

When examining continuity and change in a psychological construct across time, heterotypic continuity can become a confound that should be accounted for (rather than the focus of study). Heterotypic continuity is particularly apparent when examining change over a lengthy timespan in which people experience developmental transitions (e.g., school entry, puberty; Kagan, 1969; Schulenberg et al., 2014). If heterotypic continuity exists and is not properly accounted for, the same measure may not reflect the same construct across time and, therefore, 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). As an example, for developmental reasons, the measurement of internalizing problems should assess somatic problems to a greater degree earlier in development because the content somatic complaints (e.g., headaches, stomachaches, heart pounding) shows decreases over time in how strongly it reflects internalizing problems (Petersen et al., 2018). Thus, the 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, each with its advantages and limitations.

The three approaches to assessing a construct over time include assessing (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 the three content sets in Fig. 2: content set A refers to content that is construct-valid at only T1,

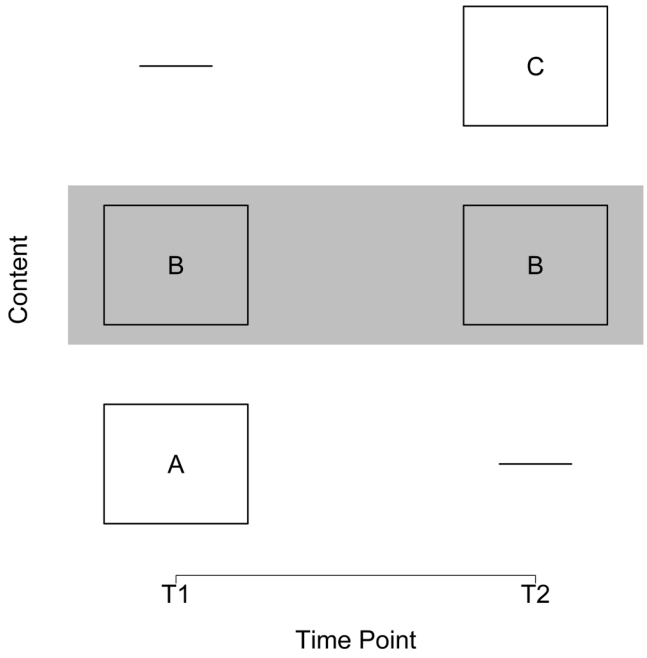

Fig. 2. Depiction of using only the construct-valid content at each age. Reprinted with permission from Petersen et al. (in press; Fig. 1). [Petersen et al. (in press). Creating a developmental scale to account for heterotypic continuity in development: A simulation study. *Child Development*. https://doi.org/10.1111/cdev.13433. Publisher: Wiley]. 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.

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: (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 scientists have used all possible content (approach 1) or only common content (approach 2) across all ages when assessing a construct over time. Below, we discuss these approaches and the importance of using construct-valid content at each age (approach 3; see Fig. 2).

Approaches to measurement

All possible content

The first approach to assessing a construct over time uses all of its content across all ages. For instance, Tong and Kolen (2007) examined the growth of academic skills on the Iowa Tests of Basic Skills from grades 3 to 8, and the same content was administered at all ages, regardless of difficulty. One advantage of this approach is its comprehensive assessment that allows examining 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. The extra time could elicit participant fatigue, increasing measurement error. Second, it could assess developmentally inappropriate content at a given age, for instance because of changes in difficulty. For example, it would be developmentally inappropriate to ask a 7-year-old an advanced calculus question when assessing math skills. Third, the aggregation of scores on all possible content could result in a score that lacks construct validity invariance and therefore becomes incomparable over time if the construct changes in its manifestation (i.e., factor structure). Consider Patterson's (1993) metaphor for heterotypic continuity: a chimera with a body of a goat that, with development, grows the head of a lion and then a tail of a snake. If we tried to assess what a chimera is using a measure of the chimera's lion head, such a measure would not identify a chimera at the earliest ages before the head of a lion is manifested.

Similarly, a measure that includes 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 (Petersen et al., 2018). It is for this reason that many measures differ in content when used at different points in development (e.g., Achenbach's measures of behavior problems and Rothbart's temperament questionnaires; Achenbach, 2009; Rothbart, Ahadi, Hershey, & Fisher, 2001). 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, using or aggregating scores on all possible content across all ages is not recommended because the measure would violate construct validity, have weaker internal consistency, and erroneously yield lower rank-order stability. Moreover simulation findings show that using all possible content results in less precise and more biased growth estimates compared to using the construct-valid content (Petersen et al., in press).

Only common content

The second approach to assessing a construct over time is to use only its common content across all ages. For instance, Sterba, Prinstein, and Cox (2007) examined trajectories of internalizing problems from 2 to 11 years of age using only the same content assessed across ages. Using only the common content is advantageous in efficiently assessing only the same, age-common information each time while retaining a consistent metric. It also may exclude the 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 reliable and 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—content that is crucial for assessing individual differences and typical versus atypical levels. For instance, in a hypothetical study of externalizing problems from 2 to 18 years of age, drug use would not be part of the common content across all ages because it would be developmentally inappropriate to ask whether 2-year-olds use drugs. Omitting drug use across all ages, however, would result in a loss of critical information on externalizing problems with high severity and frequency in adolescence. Second, the measure with only common content may lack content validity because it is not assessing all facets of the construct, specifically the age-specific manifestations. As in Patterson's (1993) metaphor for heterotypic continuity, if we assessed what a chimera is using a measure solely focused on the goat's torso (the common content), we would miss the chimera's changing manifestation with its head and tail. For a construct showing heterotypic continuity, repeated measures focused on its common content would lack content validity. Moreover, simulation findings demonstrate that using only common content yields the least accurate scores of all three approaches at the person-level (Petersen et al., in press).

Only construct-valid content

The third approach to assessing a construct over time is to use only its construct-valid content at each age. In the context of heterotypic continuity, this would mean using different content at a given age that is valid for the target construct. For instance, Petersen et al. (2015) examined trajectories of externalizing problems from 5 to 27 years of age using different construct-valid content across time to account for 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 from a longitudinal perspective. Developmental inferences are strengthened by establishing measurement invariance (equivalent measures), ensuring that differences over time reflect changes in the phenomenon of interest rather than changes in its measurement. Longitudinal assessment of constructs showing heterotypic continuity often requires different measures at different ages, which 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 constructs' manifestation.

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. And in the case of heterotypic continuity, using the construct-valid content is the recommended approach because it is the most accurate of all three approaches, at the group- and person-level (Petersen et al., in press). When using different measures (or even repeated measures), statistical and conceptual equivalence must be established to ensure that different scores across time reflect a person's true growth in the construct, as described below.

Ensuring statistical equivalence of different measures across time

There are key challenges in using the construct-valid content at each age when content differs across time. For one, how can one assess people's true change? Different scores on different content across time could reflect (a) a person's change in the construct, and/or (b) an artifactual change resulting from content differences (i.e., different measures). How does one know actual change is being assessed rather than changes in the measures' meaning? Assuming the measures reflect the same construct over time (i.e., construct validity invariance), the next consideration for determining whether different scores for a person over time reflect actual change is the issue of statistical equivalence. There are two primary considerations for determining whether the measures' scores are on the same metric or scale so they can be meaningfully compared. First, the measures should have the same range of possible scores—not necessarily the same range of observed scores or the same number of items. Second, to assess absolute change (rather than solely a person's change relative to others, i.e., rank-order change), a score on the measure at one timepoint should reflect the same level on the construct as the same score on the measure at other timepoints (i.e., measurement equivalence; Hertzog & Nesselroade, 2003). Researchers have developed multiple approaches to ensuring the statistical equivalence of different measures across time, including: (a) age-norming, (b) average or percentage scores, and (c) developmental scaling.

Age-norming

Standard scores, ranks, and percentiles are commonly used to compare scores on different measures because age-normed scores have the same 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 people's relative change compared to other people in the sample or to a norm-referenced sample (Petscher et al., 2018); however, because age-normed scores have a fixed scale, they cannot detect absolute change (i.e., whether a person or group increased or decreased in construct level over time and whether group means or variability changed over time). Standardizing scores with a fixed range or mean and standard deviation does not ensure the scores are on the same metric, so it is not recommended to use age-norming when examining development (Moeller, 2015; Willett, Singer, & Martin, 1998).

Average or proportion scores

Another approach to comparing scores across different measures is by using average (e.g., Owens & Shaw, 2003) or proportion scores (e.g., Petersen et al., 2015) that account for the different number of items in each measure (as opposed to sum scores). A major assumption of average and proportion scores is that the content on the different measures does not differ in difficulty or discrimination; however, it is unlikely that two measures whose content differ will have the same difficulty and discrimination. Thus, average or proportion scores are not typically advisable when comparing scores on different measures across time. Instead, researchers recommend vertical scaling as a form of linking or equating measures across development (e.g., Khoo, West, Wu, & Kwok, 2006; Kolen & Brennan, 2014).

The process of linking measures across development is often called vertical scaling, especially in the educational testing literature, because it involves putting measures that increase in difficulty with age on the same scale (see Fig. 3). For example, because the same test items (e.g., items assessing math or reading achievement) tend to become easier relative to a given level of ability as children get older, educational testing often uses more difficult items as children get older to retain approximately the same difficulty relative to the age level of interest. For some constructs, however, the difficulty of the content do not increase or decrease with age in a strictly "vertical" or linear fashion (e.g., the item "runs away from home" shows a higher difficulty/severity parameter at age 10 compared to early childhood or adolescence; Petersen, Bates, et al., 2016). Thus, to better align with the breadth of constructs in developmental science, we prefer the term "developmental scaling" (over "vertical scaling") when describing the process of placing measures across development on the same scale. Thus, from here on we refer to vertical scaling as developmental scaling.

Developmental scaling to account for heterotypic continuity

We propose the following approach to account for heterotypic continuity when examining developmental trajectories. First, select construct-valid content at each age that, ideally, partially overlap at adjacent ages (see Fig. 2). Second, ensure construct validity invariance of the different measures across ages. Third, test longitudinal factorial invariance of the different measures across ages.

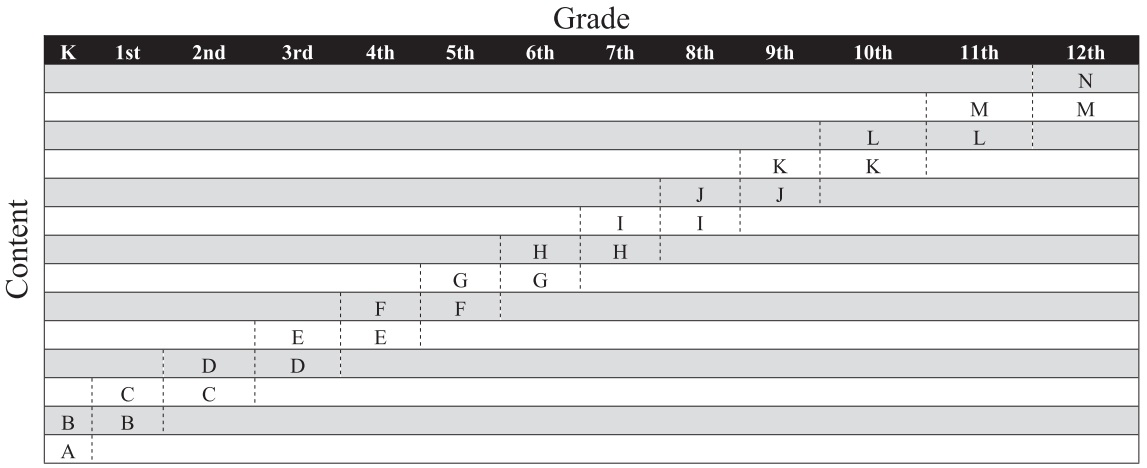

Fig. 3. Illustrative example of a vertical scaling design that uses common content to link the different measures at adjacent ages to be on the same scale. For example, content set B is administered at both kindergarten and 1st grade, and is the common content used to link scores at 1st grade to the same scale as kindergarten scores. Content set A is the unique content at kindergarten; content set C is the unique content at 1st grade (but it is common content with content set C at 2nd grade). The unique content represents age-specific manifestations of the construct. The procedure of linking measures that differ in difficulty to be on the same scale is called vertical scaling (as opposed to horizontal scaling) because the measures, especially in educational settings, tend to increase in difficulty with age (relative to a given level of ability; as depicted above with the upward-trend). Thus, vertical scaling is particularly useful for linking different measures across time to be on the same scale. Horizontal scaling, by contrast, links different measures that have the same difficulty (commonly at the same age). To better align with the breadth of constructs in developmental psychology, we use the term "developmental scaling" instead of vertical scaling to refer to putting measures across development on the same scale.

Fourth, use a developmental scaling approach to create a developmental scale and link the different measures across ages on the same scale. We describe steps 2–4 later. Fifth, estimate people's scores on the developmental scale. Sixth, estimate people's growth trajectories using their developmentally scaled scores. As we describe below, however, some of these steps can be conducted simultaneously.

In developmental scaling, measures that assess the same construct (but often differ in difficulty and/or discrimination) at different ages are placed on the same scale. The goal of developmental 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 (see Fig. 3). Although developmental scaling typically uses the common content to put two different measures on the same scale, alternative approaches may be useful for linking measures that have no common content. One approach would be to assess at least a subset of participants with each of the measures at a given timepoint, and to use linking methods such as linear, equipercentile, or Thurstone scaling to put the measures on the same scale (Kolen & Brennan, 2014). Assuming the common content sufficiently spans the content of the construct, successful linking can often be obtained with only a few common items (LeBeau, 2017). Bayesian approaches have also been used to link different measures with no common content (Oleson, Cavanaugh, Tomblin, Walker, & Dunn, 2016; Ward et al., in press). 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 with developmental scaling (Hanson & Béguin, 2002; McArdle et al., 2009). In developmental scaling, 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 timepoint 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.

When using latent variable approaches to developmental scaling (e.g., factor analysis or IRT), developmental scaling (step 4 above) can either be conducted in separate models or in a single concurrent calibration model. When conducting developmental scaling in separate models, a separate model is fitted at each age, and linking constants are then used to put the measures (i.e., people's scores on the latent construct) on the same scale. Benefits of concurrent calibration include statistical efficiency and that more accurate estimates if the modeling assumptions, including that the proper dimensionality (e.g., uni-dimensionality) is modeled across all ages, are met because concurrent calibration accounts for longitudinal dependency within-person (Kolen & Brennan, 2014; McArdle et al., 2009). However, separate estimation is more robust to violations of model assumptions, and is considered safer in practice because the modeling assumptions involve only pairs of adjacent ages and not the full age span (Kolen & Brennan, 2014). Also note that in some approaches to developmental scaling, steps 4–6 (described above) can be conducted simultaneously in the same model. For example, IRT models can be re-formulated as mixed-effect models (Chalmers, 2015) that allow estimation of growth curves. Bayesian approaches, in particular, may allow flexibility in this regard. For an example of an IRT approach that conducted the developmental scaling and growth curve estimation in the same model, see McArdle et al. (2009). Structural equation modeling could also be used for simultaneous developmental scaling and growth modeling. A benefit of performing developmental scaling and growth modeling in a single model (in addition to statistical efficiency) is that one does not have to estimate factor scores (McArdle et al., 2009), which have

been known to be indeterminate (factor score indeterminacy; DiStefano, Zhu, & Mindrila, 2009; Millsap, 2011). A potential practical challenge of this simultaneous developmental scaling and growth modeling approach may be in getting the model to converge due to greater computational complexity (Khoo et al., 2006; McArdle et al., 2009).

We also recognize other useful approaches for linking different measures across development. For example, the developmental cascade model allows examining the influences of accumulating risks on a construct that snowball into other risks (Dodge, Greenberg, Malone, & Conduct Problems Prevention Research Group, 2008). We focus on developmental scaling because of its promise for linking different measures of the same construct across development on the same scale to allow observing absolute (not just relative) growth. Multiple developmental scaling approaches can account for heterotypic continuity in studies of development.

Approaches for developmental scaling

Thurstone scaling

This approach to developmental scaling aligns percentile scores based on a range of *z*-scores on the common content. This assumes that the age groups to be linked have the same distribution form (i.e., normally distributed on the underlying construct within group), and groups' scores on the measures might differ in their mean and standard deviation. Thurstone scaling has the advantages of retaining scores on the raw metric for interpretability and does not require a large sample size. For an example of Thurstone scaling with different internalizing problem measures across adolescence to adulthood, see Petersen et al. (2018). Thurstone scaling is an observed score approach, but factor analysis and IRT offer latent variable approaches.

Factor analysis

The factor analysis approach to developmental scaling allows flexibility in estimating a latent variable using different content over time. It allows selectively estimating the same value for parameters, such as within-item intercepts, loadings, and residuals across time (i.e., testing longitudinal factorial invariance) to ensure the content has the same meaning in relation to the latent variable, and the latent variable is anchored to the same metric. Item parcels with common and unique items can be used to link different measures across time (Tyrell, Yates, Widaman, Reynolds, & Fabricius, 2019). People's factor scores on the latent variable at each age are used as

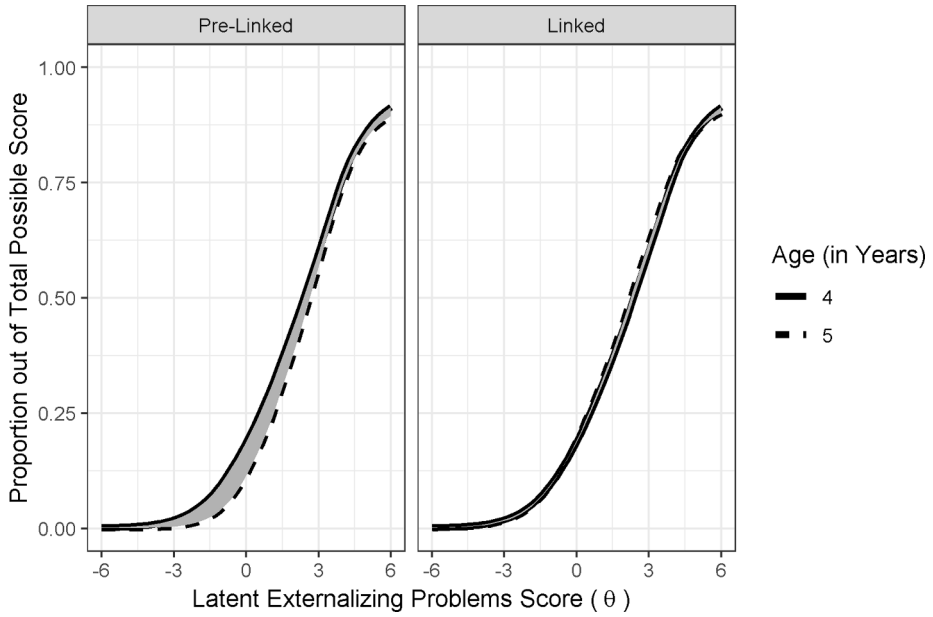

Fig. 4. Depiction of the linking process in the item response theory approach to developmental scaling. Reprinted with permission from Petersen and LeBeau (in press; Fig. 1). [Petersen and LeBeau (in press). Language ability in the development of externalizing behavior problems in childhood. *Journal of Educational Psychology*. doi: https://doi.org/10.1037/edu0000461. Publisher: American Psychological Association]. The figure illustrates the effect of linking the latent externalizing problems scores, θ, across ages, using mother-reported externalizing problems at ages 4 and 5 as an example. The left panel illustrates the test characteristic curves representing the model-implied proportion out of total possible scores across the latent externalizing problems score at age 4 and 5, before the linking process. The right panel illustrates the test characteristic curves after the linking process. The shading between the age 4 and age 5 test characteristic curves represents differences between the two test characteristic curves in terms of discrimination and/or difficulty, where larger differences reflect scores that are less comparable. Linking minimizes differences between the discrimination and difficulty of the common items. Discrimination is depicted by the steepness of the slope at the inflection point of the test characteristic curve. Difficulty is represented by the value on the x-axis at the inflection point of the test characteristic curve. The left panel indicates that the externalizing problem items showed higher difficulty at age 5 than at age 4. The right panel shows considerably smaller differences between the two test characteristic curves, which provides empirical evidence that the linking successfully placed the latent externalizing problem scores across age on a more comparable scale (i.e., more similar discrimination and difficulty of the common items).

the developmentally scaled scores. Factor analysis can accommodate many different types of data, including categorical, ordinal, and continuous data, so it can be useful for linking many types of measures, including behavioral measures. Factor analysis is useful for developmental scaling with moderate sample sizes (typically 120 or more). As an example of developmental scaling using factor analysis, Wang, Jiao, and Zhang (2013) examined trajectories of children's school achievement using different measures across grades 3 to 10. They used computerized adaptive tests of reading and math skills (i.e., developmentally appropriate, construct-valid content), averaged multiple items to create an item parcel for each sub-dimension of reading and math skills, and used the item parcels as the content in growth curve models. Other factor analysis approaches for developmental scaling are discussed in Petersen, Hoyniak, et al. (2016).

Item response theory

The IRT approach to developmental scaling involves 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 (see Fig. 4). That is, IRT links measures' scales based on the difficulty and discrimination of the common content, and is often employed for developmental scaling, especially in cognitive and educational testing. IRT is useful for developmental scaling when the sample is relatively large (typically $\geq$ 500 for two-parameter models that estimate each item's discrimination and difficulty[2]) and the content is dichotomous (e.g., true/false, correct/incorrect) or polytomous (e.g., Likert) item/trial-level data, which are common in questionnaires and performance-based assessments; however, extensions also exist for continuous data (Chen, Prudêncio, Diethe, & Flach, 2019). For instance, McArdle et al. (2009), 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 developmental scaling. They used developmentally appropriate, construct-valid content for vocabulary and memory span, and linked the different measures based on the difficulty of common content.

IRT is essentially equivalent to categorical factor analysis (Kamata & Bauer, 2008; Wirth & Edwards, 2007), and both allow estimating uni- or multi-dimensional models as well as other variants (e.g., hierarchical models, bifactor models). There may be unique benefits of IRT or factor analysis depending on the context, but both approaches to developmental scaling essentially do the same thing: they estimate a latent variable from manifest variables, and attempt to place people's scores on the latent construct at each age on the same scale. So, we have no strong a priori reason to prefer one over the other. Moreover, the latent variable (IRT/factor analysis) approaches to developmental scaling are conceptually similar to integrative data analysis (e.g., Curran et al., 2008); both are methods for data harmonization, except in developmental scaling the *same* participants are sampled (possibly with different measures) over time.

Longitudinal factorial invariance and differential item functioning

One consideration that is important when assessing change over time is longitudinal factorial invariance in factor analysis (i.e., longitudinal measurement invariance), which is analogous to a lack of differential item functioning (DIF) in IRT. Tests of longitudinal factorial invariance allow the researcher to assess whether the content has the same relations to the construct at different ages. Similarly, one can assess DIF in IRT to determine whether content differs in difficulty or discrimination across time. We focus discussion below on longitudinal factorial invariance because of its widespread use in psychology, but similar considerations are relevant for DIF. It is important to test longitudinal factorial invariance or DIF because failed longitudinal factorial invariance could indicate that the latent factor is on a different scale at each age, or it could indicate changes in the factor structure and, therefore, heterotypic continuity (Edwards & Wirth, 2009; Nesselroade & Estabrook, 2009; Widaman et al., 2010).

Traditionally, establishing longitudinal factorial invariance has been considered a pre-requisite for growth curve modeling (Widaman et al., 2010). A goal of longitudinal factorial invariance is to increase confidence that the latent construct at multiple time points is on the same scale so that meaningful changes in means and variances can be observed. In cases of *homotypic* continuity, establishing measurement invariance would be important to help ensure construct validity invariance. Establishing longitudinal factorial invariance at least between adjacent time points, and at least with some items, can help provide greater confidence that one is assessing the construct on the same scale at each age. Traditionally, there are four primary types of invariance that are tested in a sequential way (with successively more constraints at each step because constraints from prior steps are retained): (1) configural invariance (invariance of which items load onto which factors across time), (2) weak factorial ("metric") invariance (invariance of within-indicator factor loadings or discrimination across time), (3) strong factorial ("scalar") invariance (invariance of within-indicator intercepts or difficulty across time), and (4) strict factorial ("residual") invariance (invariance of within-indicator residual variances across time). Little et al. (2007) suggested that researchers establish at least partial weak factorial invariance to examine associations between latent variables, and to establish at least partial strong factorial invariance to examine mean-level change in latent variables. Partial invariance refers to invariance with some indicators but failed invariance with other indicators. The more indicators that show longitudinal factorial invariance, the greater the confidence that the latent construct is on a comparable scale at

[2] However, one-parameter and Bayesian IRT approaches may accommodate smaller sample sizes. For information on linking with smaller sample sizes, see the special issue on "Practical Issues in Linking and Equating with Small Samples" in the journal, *Applied Measurement in Education* Peabody, M. R. (2020). Practical issues in linking and equating with small samples. *Applied Measurement in Education*, *33*(1), 1–2. https://doi.org/10.1080/08957347.2019.1674306.

each age for examining people's growth.

As an example of a study that established longitudinal factorial invariance with different measures across ages, Tyrell and colleagues (2019) examined different measures of participants' depression in adolescence versus early adulthood, with common and unique items. They used item parcels of the common items and unique items, and established partial strong longitudinal factorial invariance by placing cross-time constraints on the factor loadings and intercepts of the common item parcels. It is worth noting that, despite their utility, tests of factorial invariance have important limitations. Tests of factorial invariance (and DIF) rest on fundamentally untestable assumptions related to scale setting (Raykov et al., in press), and results of factorial invariance tests can depend highly on which item is used as the anchor item for setting the scale of the latent factor (Belzak & Bauer, in press).

Nevertheless, longitudinal factorial invariance is important to test. However, in the context of *heterotypic* continuity, longitudinal factorial invariance would not necessarily be expected because the construct's factor structure, by definition, changes with development. In the case of heterotypic continuity, even the common content might not be expected to show invariance in intercepts and factor loadings across all ages (even if the content remains construct-valid). Thus, longitudinal factorial invariance, though useful, should not be the sole marker of whether the model is correct. Theoretical considerations are crucial and establishing longitudinal factorial invariance may not be required in consideration of heterotypic continuity (Knight & Zerr, 2010; Petersen, Hoyniak, et al., 2016; Petersen et al., 2018), because models with failed longitudinal measurement invariance can yield valid inferences in the context of heterotypic continuity (Edwards & Wirth, 2012). Removing content or measures that show DIF or failed measurement invariance over time is not recommended in the case of heterotypic continuity if the measure retains construct validity (Knight & Zerr, 2010). Removing content can result in a less representative sample of content of the construct (i.e., lower content validity), and some content might be expected to change in difficulty or discrimination because of heterotypic continuity, and yet remain construct-valid. Discarding such content would remove meaningful developmental information about the construct. Discarding construct-valid content showing DIF or failed measurement invariance would be akin to using only common content, which we argue is problematic and violates content validity. Lack of DIF is not an assumption of all approaches to developmental scaling. Indeed, one procedure for linking different measures using an IRT approach to developmental scaling is the Stocking-Lord procedure (Stocking & Lord, 1983), which minimizes differences in discrimination and difficulty of the measures at the construct-level (i.e., latent-level) rather than the item-level. Thus, two measures can still be linked on the same scale even if individual items show non-invariance (i.e., differences in discrimination or difficulty) across time. Items showing non-invariance with the Stocking-Lord procedure commonly offset one another such that some items may show positive or negative non-invariance, but when averaged at the construct level, show stronger evidence of construct-level invariance.

Failed longitudinal factorial invariance can complicate interpretation because it could suggest that (a) the construct changes in its manifestation, and/or (b) the measures change in their functioning. If measures change in their functioning or meaning with development, the variables may not be comparable across time, and their growth curves could be comparing "apples" to "oranges." Ideally, the researcher would refer to the literature to disentangle these possibilities, because there is no statistical method for deciding between them. There are few empirical guidelines for determining the severity of failed factorial invariance and whether it needs to be addressed, although Millsap (2010) has proposed using effect sizes. For additional discussion about longitudinal factorial invariance in the context of heterotypic continuity, see Petersen, Hoyniak, et al. (2016).

Ensuring construct validity invariance of different measures across time

Developmental scaling and longitudinal factorial invariance approaches do not ensure variables at different ages have the same conceptual metric, which is critical for modeling developmental trajectories. To ensure this, construct validity invariance is also necessary (Knight & Zerr, 2010). In other words, although identical measures over time are unnecessary, measures' scores should have identical meaning across the study's timeframe (Owens & Shaw, 2003). There are many ways to develop construct validity of measures' scores for a given construct. First, the content selected for the measures should be based on theory—judged to reflect the same construct—and the content should adequately sample different facets of the construct (content validity). Second, despite the possibility for low rank-order stability in the long-term, there should be test–retest reliability of the measures' scores in the short-term. Third, consistent with the nomological network, the measures' scores should show convergent validity with each other and discriminant validity with scores on measures of distinct constructs. Integrating a developmental perspective to the nomological network, the measures' scores should also show associations (or non-associations) with scores on measures of other constructs in ways that accurately reflect stability and change in the nomological network: increasing (convergence), decreasing (divergence), or stable (parallelism) associations among the constructs across time (Buss & Royce, 1975). Fourth, the measures' scores should demonstrate a similar factor structure across time, yet might not be expected to have an invariant structure because of changes in the factor structure with age resulting from heterotypic continuity. Nevertheless, developmental scaling should alleviate many problems with longitudinal factorial non-invariance as long as the measures retain construct validity invariance, because developmental scaling places the scores on the same mathematical metric. Fifth, the measures' scores should have high internal consistency. Sixth, the measures' scores might be expected to be sensitive to change and to show theoretically expected change across development.

Although these and other approaches can help assess the construct validity of measures' scores for a given construct, construct validation is a continual process, not a fixed outcome. Constructs evolve as our understanding changes. Greater attention to the structure of constructs, from theoretical and empirical perspectives, will yield refinements to dynamic constructs and approaches to study them. In sum, in order to model developmental trajectories, it is important for measures' scores to have theoretical relevance to the construct at each age examined and to be on a comparable metric for measurement equivalence (as opposed to measurement invariance, which again may be unnecessary when accommodating heterotypic continuity; Knight & Zerr, 2010).

The importance of construct validity invariance has inspired an approach to testing factorial invariance called idiographic filtering (Molenaar & Nesselroade, 2012; Nesselroade & Estabrook, 2009; Nesselroade, Gerstorf, Hardy, & Ram, 2007; Nesselroade & Molenaar, 2016). Idiographic filtering seeks to filter out idiosyncratic individual differences in behavior or measurement to obtain what is common to the construct across people. In order to establish construct invariance, idiographic filtering attempts to achieve factorial invariance not at the level of measurement but at the latent (i.e., construct) level, with measurement allowed to differ across people, but constraining the relations among constructs (the nomological network) to be the same across people. Traditional factorial invariance requires the same factor loadings of the measures on the latent constructs across people. By contrast, idiographic filtering constrains the interrelations among the latent constructs (or of a latent construct across time) to be the same across people, using person-level factor analysis (e.g., p-technique). Some researchers, however, have questioned the assumption in idiographic filtering of construct invariance across people, arguing that the interrelations among constructs can differ between people (West & Ryu, 2007). In any case, the development of idiographic filtering demonstrates that there are multiple levels at which one can evaluate invariance, including the measurement-level and construct-level.

Using developmentally scaled scores

When there is theoretical and empirical evidence of measurement equivalence across development, developmentally scaled scores can be used to chart people's growth in a construct using growth curve models. The validity of a measure can only be seen in relation to other measures (including the association of a measure with the same measure or other measures at other times). Thus, stability of individual differences on a construct across time can provide evidence of a common process, and therefore continuity (e.g., heterotypic continuity). However, people can change in their level on the construct irrespective of the construct's changing behavioral manifestation, and heterogeneous trajectories can reduce the stability of individual differences on a construct. For instance, Betts et al. (2016) found only modest stability of individual differences in internalizing problems from childhood to adulthood, partly because some children's internalizing problems increase while others' decrease across development. Developmental scaling does not require stability in individual differences on the construct to link scores from different measures across time points, as long as the measures assess the same construct. Thus, developmental 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 developmental scaling can then be linked to risk factors, protective factors, and downstream outcomes.

Studies using developmental scaling to examine growth

Many studies have used developmental scaling in the fields of education and cognitive testing to assess growth with different measures across time (e.g., Kenyon, MacGregor, Li, & Cook, 2011; McArdle & Grimm, 2011; McArdle et al., 2009; Murayama, Pekrun, Lichtenfeld, & vom Hofe, 2013; Wang et al., 2013). Developmental scaling permitted these studies to examine people's change over lengthy developmental spans. We know of only five studies that used developmental scaling to examine growth in social development, in the case of self-esteem (Hancock & Buehl, 2008), underactivity/overactivity (and their sub-dimensions; McDermott, Watkins, Rovine, & Rikoon, 2013), internalizing problems (Petersen et al., 2018) and externalizing problems (Petersen & LeBeau, in press). Tyrell and colleagues (2019) conducted developmental scaling of different depression measures, but the study did not examine participants' change (e.g., slope) in depression across time. Petersen et al. (2018) observed a group-level decrease in internalizing problems from adolescence to adulthood when using developmental scaling, a change that would not have been observed using only the common content. Thus, not only does developmental scaling permit studying lengthier spans of development, it also yields inferences that are better able to detect developmental change. Despite these benefits, researchers may have concerns about studying people's trajectories with different measures across time. Now we discuss potential questions or concerns about measurement and heterotypic continuity.

Common questions about measurement and heterotypic continuity

Is it possible to use the same measure across time to account for heterotypic continuity?

With the goal of assessing development across the lifespan, there have been attempts to use the same measure across time, but such approaches have serious problems. Despite advances in the precision of biological measures (e.g., genotyping, functional magnetic resonance imaging [fMRI]), recent developmental research has paid relatively less attention to improving behavioral measurement strategies. As demonstrated by Petersen, Hoyniak, et al. (2016), behavioral tasks often have restricted time frames of usefulness, owing to changes in developmental appropriateness and/or sensitivity. *Developmentally appropriate* measures are consistent with the capacity of most people of a particular developmental level and with our understanding of the construct at that point in development. If a measure is not appropriate for a given developmental level, it lacks validity for describing people's development in the construct. *Developmentally sensitive* measures yield enough variability to assess individual differences within a narrow window and to compare individual differences across development. Developmental appropriateness and sensitivity help ensure proper measurement of each person's level on the construct and adequate variability between people for rank ordering. It is common for behavioral tasks to be developmentally appropriate and sensitive at one age, but too easy for participants at later ages, resulting in ceiling effects, and too difficult for younger children, resulting in floor effects. For instance, behavioral tasks of inhibitory control have been shown to be useful, on average, for less than three years across childhood due to ceiling and floor effects (Fig. 5). Ceiling and floor effects prevent

researchers from detecting meaningful variability in a construct and result in a restricted range, which increase Type II error rates by reducing power to detect changes in the construct and its associations with other variables. Administering measures that produce ceiling and floor effects also wastes time, material resources, and labor.

To make behavioral tasks useful for a wider developmental range, researchers can increase rule or content difficulty, add rule switches, and reduce the amount of time a child has to respond (Carlson, Faja, & Beck, 2016). Adaptive tests often use these tactics to scale content difficulty to the participant's ability level (Wang et al., 2013); however, a task will lack construct validity invariance if it does not accommodate the construct's heterotypic continuity. Nevertheless, there have been attempts to assess constructs using behavioral tasks over lengthy developmental spans, such as the National Institutes of Health Toolbox (Weintraub et al., 2013) and the Minnesota Executive Function Scale (Carlson & Zelazo, 2014). In light of the restricted time frame of usefulness in behavioral tasks, researchers may be inclined to implement other assessment techniques that allow comparability across time. One option is to use informant ratings on questionnaires; however, many of the same longitudinal challenges with behavioral assessments are present in informant ratings. For a questionnaire to have construct validity and accommodate heterotypic continuity, its items may have to differ across development.

Another option is to use a more precise metric for the measure: a continuous interval or ratio metric as opposed to an ordinal or nominal metric (e.g., pass/fail). Measures with a highly precise metric, including reaction time and psychophysiological techniques, purportedly place all people on the same metric for equivalent comparison within and across ages. Even if no differences are observed in accuracy because of ceiling or floor effects, the same task with a dependent variable of a higher precision metric may detect developmental differences that make the task useful across a wider developmental range. Studies have combined accuracy and reaction time to eliminate floor and ceiling effects (i.e., when relying on accuracy alone) and to enhance the sensitivity of measures to detect developmental change (Carlson & Zelazo, 2014; Magnus, Willoughby, Blair, & Kuhn, 2019; Weintraub et al., 2013). Integrating accuracy and reaction time may be useful for assessing developmental change in performance on a task across a wider age range; however, examining developmental change by assessing reaction time on the same task has several limitations. First, including

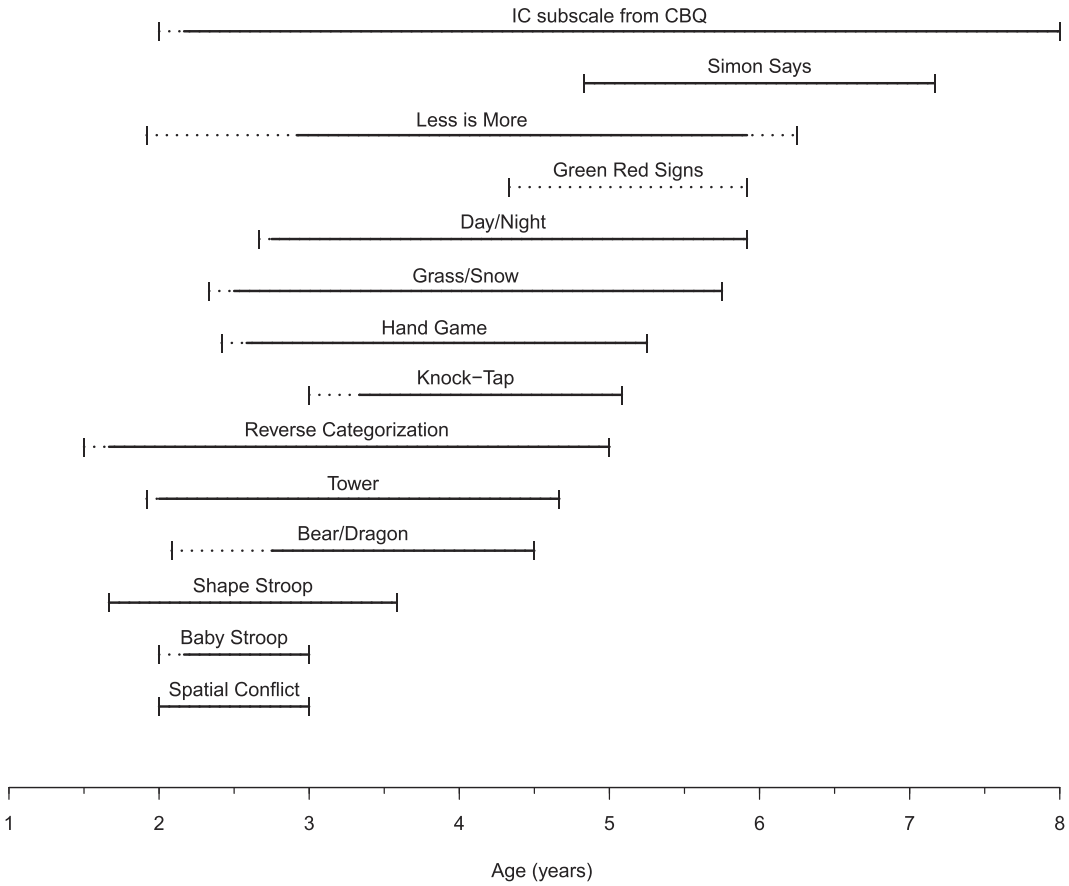

Fig. 5. Depiction of useful age ranges of inhibitory control measures. Reprinted with permission from Petersen, Hoyniak, et al. (2016, p. 42; Fig. 4). [Petersen, Hoyniak, et al., 2016. Measuring the development of inhibitory control: The challenge of heterotypic continuity. *Developmental Review*, *40*, 25–71. doi:https://doi.org/10.1016/j.dr.2016.02.001. Publisher: Elsevier]. Dotted lines represent age ranges with weaker empirical support (fewer than two studies). CBQ = Children's Behavior Questionnaire. Behavioral tasks of inhibitory control were shown to be useful for a developmental span of less than 3 years on average. The limited age range of usefulness of behavioral tasks suggests that different measures across time are necessary to assess change across longer spans.

reaction time with accuracy in the score may alter the score's meaning with respect to the construct. Second, reaction time may change in meaning with age. For example, age-related differences in reaction time do not always reflect developmental differences in processing speed (Anderson, Nettelbeck, & Barlow, 1997). Third, using the same measure across time in which reaction time is assessed, even if it does not yield floor or ceiling effects, would not capture the changing nature of a construct that shows heterotypic continuity. For example, Widaman and colleagues (1991, 1992, 2010) described several studies in which people showed decreased reaction times for basic arithmetic from elementary school to college along with qualitative changes in the strategies they used, including reconstructive strategies (e.g., counting) in early elementary school versus memory retrieval in college. However, the authors noted that relying solely on reaction time would have missed these key qualitative changes, despite having established strict longitudinal factorial invariance. Thus, even if longitudinal factorial invariance is established, relying solely on reaction time may miss the true developmental process when the construct shows heterotypic continuity.

One psychophysiological technique with a precise metric is the event-related potential (ERP), brain potentials recorded at the scalp by electroencephalography. Many ERP components are present across wide age ranges, and important developmental changes in magnitude, timing, and morphology of ERP components are related in part to changes in brain size and skull thickness (Rueda, Posner, & Rothbart, 2005). Thus, it is unclear whether developmental changes in ERP components reflect changes in the cognitive processes that they index, and whether ERPs can truly be compared across development. In addition to physical changes that pose challenges for comparing ERP components across development, the same task may have different task demands (and elicit distinct cognitive processes) at different ages. Thus, similar to reaction times, relying solely on ERPs may mask true developmental changes at a cognitive process level. Challenges in comparability across ages also exist for other psychophysiological techniques. Researchers have questioned the utility of comparing neural activation across development as assessed by fMRI, because age differences could reflect technical aspects of data acquisition rather than developmental change in neural function (Gaillard, Grandin, & Xu, 2001). Moreover, the normative range of heart rate frequency or variability as assessed by electrocardiography changes with development due to age-related changes in breathing and heart rates (Fox, Schmidt, Henderson, & Marshall, 2007). In sum, measures with a precise metric (e.g., reaction time) may not be comparable on the same conceptual metric across development.

The demands of accommodating a construct's heterotypic continuity have frequently led researchers to grapple with developmental equivalence or to avoid examining development across lengthy time spans. Many studies have used all possible content or only common content to maintain the same measurement over time (for a review of such studies on internalizing problems, see Petersen et al., 2018). Many studies have used all possible content of measures outside the ages they were originally designed to assess in developmentally appropriate and sensitive ways or they have discarded construct-valid content to retain only common content.

Thus, many studies address developmental equivalence and, in many cases, resort to using a measure at an age outside the age-range of validation or discarding relevant content, possibly because of a misperception that no simple approach to handle different measures across time exists. Consider the following statement by Broeren et al. (2013) about assessing anxiety trajectories from early to late childhood: "Although the questionnaire was originally developed to measure anxiety in preschoolers, the current study also employed the scale with older children to promote uniformity in measures" (p. 84). There are alternative approaches to handling measurement changes to account for heterotypic continuity, but ignoring heterotypic continuity results in measures that violate construct validity (if using all possible content) or content validity (if using only common content). Using the same measures across time severely restricts our understanding of development across important transitions such as puberty and entry into parenthood. It also limits how lengthy of a developmental span one can examine, because in many domains, especially in early childhood, measures are not useful or designed for a wide age range (Petersen, Hoyniak, et al., 2016). Researchers studying development are doing their best to address developmental equivalence, but it is important for the field to advance its awareness of approaches for handling different measures across time to maintain construct validity invariance.

Why not just study development in a piecewise fashion?

Given the complexities of studying development over a long time frame, why not solely assess development in a short, piecewise manner with greater precision? First, many measures are useful for restricted age ranges (in part because of heterotypic continuity; Petersen, Hoyniak, et al., 2016). Thus, relying on a fixed set of content would severely restrict our ability to see growth over a longer time span. Second, heterotypic continuity is a developmental complexity that arises in many domains of functioning and lifespan phases. Seeking to understand changes across important developmental periods and transitions is better than ignoring the phenotypic complexities associated with meaningful developmental change. Third, the ultimate goal of developmental science is to understand the whole trajectory of a person's life, and not just transitory outcomes at a particular point in development. Examining intelligence longitudinally across the lifespan has provided important insights that growth in some aspects of intelligence continues into midlife, whereas cross-sectional studies suggested that intelligence peaks in early adulthood and declines thereafter (Schaie, 2000).[3] The pathways to outcomes can be just as important as the outcomes themselves, emphasizing the importance of considering developmental

[3] The form of age-related change likely depends on the type of intelligence. Fluid intelligence (e.g., working memory, processing speed) appears to decline into adulthood whereas crystallized intelligence (e.g., learned skills, cultural knowledge, verbal skills, vocabulary) appears to show continued growth past midlife into early stages of aging Horn, J. L., & Blankson, N. (2005). Foundations for better understanding of cognitive abilities. In *Contemporary Intellectual Assessment: Theories, Tests, and Issues.* (pp. 41–68). The Guilford Press. https://psycnet.apa.org/record/2005–09732-003, Horn, J. L., & Donaldson, G. (1980). Cognitive development in adulthood. In J. Kagan & J. Brim (Eds.), *Constancy and change in human development* (pp. 445–529). https://www.hup.harvard.edu/catalog.php?isbn=9780674166257.

science principles such as equifinality and multifinality. Thus, research should strive to build a bridge that spans infancy to adulthood (Rutter & Sroufe, 2000).

Why study higher-order constructs and not just the sub-dimensions or individual behaviors?

Given the complexity of understanding the development of higher-order constructs that change in expression over time, why not just study the underlying sub-dimensions or individual behaviors in isolation to capture more homogeneous processes that demonstrate *homotypic* continuity? There is utility in examining constructs at the higher-order level rather than just at the level of their underlying behaviors, facets, or sub-dimensions. Aggregate scores from multiple measures tend to have better psychometric properties than scores of individual variables or behaviors because of aggregation (Rushton, Brainerd, & Pressley, 1983): having multiple measurements or sub-dimensions is more reliable than a single measurement or sub-dimension (assuming the measurements validly assess the same construct). Moreover, reliability is necessary (even if insufficient) for validity. Nevertheless, validity is a continual process, not an outcome, so constructs must evolve so their measures become more psychometrically sound.

One might argue that sub-dimensions of constructs may be more likely than higher-order constructs to demonstrate homotypic continuity; however, even sub-dimensions of constructs are likely to change in meaning with development (i.e., functional discontinuity). Consider a specific externalizing behavior: disobedience. Disobedience in childhood could be a developmentally appropriate content of externalizing problems. In adulthood, disobedience to authority could also reflect prosocial functions, including protesting against societally unjust actions, and may show weaker construct validity with respect to externalizing problems. Thus, a measure of disobedience might differ in meaning with respect to the construct of externalizing problems in childhood versus adulthood. Even specific behavioral content shows changes in meaning with development.

Some argue that constructs should be uni-dimensional (Strauss & Smith, 2009). However, no constructs in psychology are truly uni-dimensional because they can always be reduced to a lower level unit to be more uni-dimensional and this reductionism can result in less utility. For instance, broadband externalizing problems can be reduced to aggression, impulsivity, rule-breaking, inattention, and hyperactivity (Achenbach, 2009). Aggression can be further reduced to reactive and proactive forms of aggression; which can be further reduced to physical or relational aggression (Dodge, 2006). And so on. Similar points can be made about other constructs. Alternatively, Insel (2014) has argued that instead of focusing on behavior to define syndromes (because behavior is imprecise and provides little information about underlying mechanisms), research should focus on the developmental trajectories of brain mechanisms. As discussed earlier, however, it is unclear in some cases whether measures of brain functioning are capturing actual developmental change in brain functioning. In sum, the researcher should consider their goal(s) when choosing which levels of analysis to examine.

Given the considerable empirical evidence supporting the internalizing-externalizing spectra of psychopathology (Forbes, Tackett, Markon, & Krueger, 2016), there seems to be utility in examining development at this broadband level of analysis. The three-factor view of psychopathology (internalizing, externalizing, and disordered thought processes) captures much of psychopathology, and has established utility (Bates, Schermerhorn, & Petersen, 2014; Caspi et al., 2014). Moreover, at least among externalizing problems, there is evidence that different sub-dimensions of externalizing problems tend to co-occur, and similar developmental processes appear to be involved with the different sub-dimensions (Dodge, 2006; Olson, Bates, Sandy, & Lanthier, 2000; Shaw, Lacourse, & Nagin, 2005). And among internalizing problems, anxiety and depression sub-dimensions show sequentially comorbid courses (Garber & Weersing, 2010). Thus, forms of psychopathology that demonstrate heterotypic continuity may contribute to a better understanding of sequential comorbidity, such that symptoms of one form of mental illness change in manifestation to present as symptoms of other forms of psychopathology. For instance, increases in depressive symptoms in adolescence could follow decreases in separation anxiety across childhood when their underlying meaning is the same. Alternatively, a given construct may influence and evoke changes in other constructs, consistent with a developmental cascade. For example, antisocial behavior could lead to failure in developmentally important tasks and bring about academic failure, peer rejection, and depressed mood (Patterson, 1993). For an empirical example, high levels of externalizing problems throughout childhood have been shown to predict boys' depression in adolescence (Shaw, Hyde, & Brennan, 2012).

Future directions

We see a number of important directions for future research that will advance developmental science. First, future research should pay greater attention to whether and how construct(s) change in manifestation with development. We would especially like to see more research examining stability and change of constructs at the content level to more precisely determine *how* a given construct shows heterotypic continuity. For example, a study could examine whether a particular content such as disobedience shows decreases in discrimination for externalizing problems from childhood to adolescence. Relatedly, we would like to see studies that evaluate whether constructs must have a core set of common content across all ages (e.g., the chimera analogy), or whether it is possible for constructs to show changes in behavioral manifestation across development such that there is no common content (at the behavioral level) that spans all ages. Integrating a developmental perspective with the understanding of the structure of constructs will lead to a more nuanced understanding of constructs, which will improve developmental theory.

Second, future research will benefit from extending and improving the guidelines proposed here for ways to account for heterotypic continuity. Additional information would be helpful to guide researchers to account for heterotypic continuity. It would be helpful for research to determine which methods result in the most accurate way to account for heterotypic continuity, and under which circumstances. For instance, comparisons of different approaches to developmental scaling (Thurstone scaling versus IRT versus factor

analysis) under different circumstances would be helpful. It would also be helpful to know the degree of overlap (e.g., common content) that is required between two measures to accurately link them, and how best to link measures that share little or no content. Further, empirical guidelines for determining the severity of, and how to handle, failed longitudinal factorial invariance would be helpful.

Third, per these and future guidelines, future research should follow appropriate measurement and statistical tools to account for heterotypic continuity of constructs. As a practical tool for researchers, we provide analysis scripts here for using the IRT approach to developmental scaling to account for heterotypic continuity when studying development: https://osf.io/ewmzd (Petersen et al., in press) and https://osf.io/9zd6e (Petersen & LeBeau, in press).

Conclusion: Implications for developmental psychology

This is the first review and synthesis on heterotypic continuity, along with ways to account for it when studying development. Researchers have (a) demonstrated that constructs change in manifestation with development (Patterson, 1993), (b) argued that different measures across development are necessary to account for heterotypic continuity (Widaman et al., 2010), and (c) developed methods for linking different measures across development to be on the same scale (Kolen & Brennan, 2014). Despite this knowledge, studies in developmental science have largely failed to account for the changing manifestation of constructs in ways that detect meaningful growth. What is novel in our review is the synthesis of these theoretical and methodological ideas to advance the study of development across the lifespan in meaningful ways. Failing to account for heterotypic continuity results in inaccurate developmental inferences (Chen & Jaffee, 2015; Petersen et al., in press; Petersen et al., 2018).

A key goal of developmental science is to understand developmental pathways across the lifespan, and not just limited windows of time or stages in people's lives. By (a) paying greater attention to how constructs change in their expression with development, and by (b) accounting for heterotypic continuity when it exists, researchers can study development across the lifespan in ways that are consistent with construct validity invariance. Accounting for heterotypic continuity may require using different measures across time with developmental scaling approaches that link different measures on a comparable mathematical metric across development. Measurement approaches that accommodate changes in the construct over time are essential for making accurate developmental inferences, especially over lengthy spans of development. Given how common heterotypic continuity is among many psychological constructs, accounting for heterotypic continuity is crucial to advancing our understanding of development across the lifespan.

Author note

We have no conflicts of interest to disclose. This research was funded by Grant HD098235 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. 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.

Acknowledgments

We graciously thank Bob McMurray for providing helpful feedback on an early draft of the manuscript. We thank Alyssa Varner for help in designing Fig. 1.

References

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

Achenbach, T. M., & Edelbrock, C. S. (1983). Manual for the child behavior checklist and revised child behavior profile. University of Vermont, Department of Psychiatry. https://store.aseba.org/Manuals/products/40/.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. University of Vermont, Research Center for Children, Youth, and Families. https://store.aseba.org/MANUAL-FOR-THE-ASEBA-SCHOOL-AGE-FORMS-PROFILES/productinfo/505/.

Anderson, M., Nettelbeck, T., & Barlow, J. (1997). Reaction time measures of speed of processing: Speed of response selection increases with age but speed of stimulus categorization does not. *British Journal of Developmental Psychology, 15*(2), 145–157. https://doi.org/10.1111/j.2044-835X.1997.tb00731.x

Arnett, A., Pennington, B., Willcutt, E., Dmitrieva, J., Byrne, B., Samuelsson, S., & Olson, R. (2012). A cross-lagged model of the development of ADHD inattention symptoms and rapid naming speed. *Journal of Abnormal Child Psychology, 40*(8), 1313–1326. https://doi.org/10.1007/s10802-012-9644-5

Baltes, P. B., & Nesselroade, J. R. (1970). Multivariate longitudinal and cross-sectional sequences for analyzing ontogentic and generational change: A methodological note. *Developmental Psychology, 2*(2), 163–168. https://doi.org/10.1037/h0028743

Bates, J. E., Schermerhorn, A. C., & Petersen, I. T. (2014). Temperament concepts in developmental psychopathology. In K. Rudolph & M. Lewis (Eds.), Handbook of developmental psychopathology (3rd ed., pp. 311–329). Springer. https://doi.org/10.1007/978-1-4614-9608-3_16.

Bell, R. Q., Weller, G. M., & Waldrop, M. F. (1971). Newborn and preschooler: Organization of behavior and relations between periods. *Monographs of the Society for Research in Child Development, 36*(1/2), 1–145. https://doi.org/10.2307/1165655

Belzak, W. C. M., & Bauer, D. J. (in press). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. Psychological Methods. https://doi.org/10.1037/met0000253.

Betts, K. S., Baker, P., Alati, R., McIntosh, J. E., Macdonald, J. A., Letcher, P., & Olsson, C. A. (2016). The natural history of internalizing behaviours from adolescence to emerging adulthood: Findings from the Australian Temperament Project. *Psychological Medicine, 46*(13), 2815–2827. https://doi.org/10.1017/S0033291716001495

Blumberg, M. S. (2013). Homology, correspondence, and continuity across development: The case of sleep. *Developmental Psychobiology, 55*(1), 92–100. https://doi.org/10.1002/dev.21024

Broeren, S., Muris, P., Diamantopoulou, S., & Baker, J. R. (2013). The course of childhood anxiety symptoms: Developmental trajectories and child-related factors in normal children. *Journal of Abnormal Child Psychology, 41*(1), 81–95. https://doi.org/10.1007/s10802-012-9669-9

Buss, A. R. (1973). A conceptual framework for learning effecting the development of ability factors. *Human Development, 16*(4), 273–292. https://doi.org/10.1159/000271282

Buss, A. R. (1974). Multivariate model of quantitative, structural, and quantistructural ontogenetic change. *Developmental Psychology, 10*(2), 190–203. https://doi.org/10.1037/h0035845

Buss, A. R., & Royce, J. R. (1975). Ontogenetic changes in cognitive structure from a multivariate perspective. *Developmental Psychology, 11*(1), 87–101. https://doi.org/10.1037/h0076115

Carlson, S. M., Faja, S., & Beck, D. M. (2016). Incorporating early development into the measurement of executive function: The need for a continuum of measures across development. In J. A. Griffin, P. McCardle, & L. Freund (Eds.), Executive function in preschool-age children: Integrating measurement, neurodevelopment, and translational research. American Psychological Association. https://doi.org/10.1037/14797-003.

Carlson, S. M., & Zelazo, P. D. (2014). Minnesota Executive Function Scale. Test manual. Reflection Sciences, LLC. https://ceed.umn.edu/in-person-trainings/intro-to-mefs/.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., … Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science, 2*(2), 119–137. https://doi.org/10.1177/2167702613497473

Caspi, A., & Shiner, R. L. (2006). Personality development. In N. Eisenberg, W. Damon, & R. M. Lerner (Eds.), Handbook of child psychology (6th ed., Vol. 3, pp. 300–365). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470147658.chpsy0306.

Chalmers, R. P. (2015). Extended mixed-effects item response models with the MH-RM algorithm. *Journal of Educational Measurement, 52*(2), 200–222. https://doi.org/10.1111/jedm.12072

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*(3), 269–288. https://doi.org/10.1007/s40865-015-0012-3

Chen, Y., Prudêncio, R. B. C., Diethe, T., & Flach, P. (2019). β3-IRT: A new item response model and its applications. arXiv:1903.04016. https://arxiv.org/abs/1903.04016.

Chomsky, N. (1971). Deep structure, surface structure, and semantic interpretation. In D. D. Steinberg, & L. A. Jakobovits (Eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology* (pp. 183–216). Cambridge University Press. https://www.cambridge.org/us/academic/subjects/languages-linguistics/semantics-and-pragmatics/semantics-interdisciplinary-reader-philosophy-linguistics-and-psychology.

Cicchetti, D., & Rogosch, F. A. (2002). A developmental psychopathology perspective on adolescence. *Journal of Consulting and Clinical Psychology, 70*(1), 6–20. https://doi.org/10.1037/0022-006X.70.1.6

Coan, R. W. (1966). Child personality and developmental psychology. In R. B. Cattell (Ed.), Handbook of multivariate experimental psychology (pp. 732–752). Rand McNally. https://psycnet.apa.org/record/1966-35009-000.

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Developmental Psychology, 44*(2), 365. https://doi.org/10.1037/0012-1649.44.2.365

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical assessment, research & evaluation, 14*(20), 1–11. https://doi.org/10.7275/da8t-4g52

Dodge, K. A. (2006). Translational science in action: Hostile attributional style and the development of aggressive behavior problems. *Development and Psychopathology, 18*(3), 791–814. https://doi.org/10.1017/S0954579406060391

Dodge, K. A., Greenberg, M. T., Malone, P. S., & Conduct Problems Prevention Research Group. (2008). Testing an idealized dynamic cascade model of the development of serious violence in adolescence. Child Development, 79(6), 1907–1927. https://doi.org/10.1111/j.1467-8624.2008.01233.x.

Eaton, N. R., Krueger, R. F., Markon, K. E., Keyes, K. M., Skodol, A. E., Wall, M., … Grant, B. F. (2013). The structure and predictive validity of the internalizing disorders. *Journal of Abnormal Psychology, 122*(1), 86–92. https://doi.org/10.1037/a0029598

Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development, 6*(2–3), 74–96. https://doi.org/10.1080/15427600902911163

Edwards, M. C., & Wirth, R. J. (2012). Valid measurement without factorial invariance: A longitudinal example. In G. R. Hancock, & J. R. Harring (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 289–311). Information Age Publishing.

Emmerich, W. (1964). Continuity and stability in early social cevelopment. *Child Development, 35*(2), 311–332. https://doi.org/10.2307/1126699

Emmerich, W. (1968). Personality development and concepts of structure. *Child Development, 39*(3), 671–690. https://doi.org/10.2307/1126978

Eyberg, S. M., & Pincus, D. (1999). Eyberg Child Behavior Inventory & Sutter-Eyberg Student Behavior Inventory-Revised: Professional Manual. *Psychological Assessment Resources*. https://www.parinc.com/products/pkey/97.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Ferdinand, R. F., Dieleman, G., Ormel, J., & Verhulst, F. C. (2007). Homotypic versus heterotypic continuity of anxiety symptoms in young adolescents: Evidence for distinctions between DSM-IV subtypes. *Journal of Abnormal Child Psychology, 35*(3), 325–333. https://doi.org/10.1007/s10802-006-9093-0

Forbes, M. K., Tackett, J. L., Markon, K. E., & Krueger, R. F. (2016). Beyond comorbidity: Toward a dimensional and hierarchical approach to understanding psychopathology across the life span. *Development and Psychopathology, 28*(4pt1), 971–986. https://doi.org/10.1017/S0954579416000651

Fox, N. A., Schmidt, L. A., Henderson, H. A., & Marshall, P. J. (2007). Developmental psychophysiology: Conceptual and methodological issues. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* ((3rd ed.,, pp. 453–481). Cambridge University Press. https://doi.org/10.1017/CBO9780511546396.020.

Gaillard, W. D., Grandin, C. B., & Xu, B. (2001). Developmental aspects of pediatric fMRI: Considerations for image acquisition, analysis, and interpretation. *NeuroImage, 13*(2), 239–249. https://doi.org/10.1006/nimg.2000.0681

Garber, J., & Weersing, V. R. (2010). Comorbidity of anxiety and depression in youth: Implications for treatment and prevention. *Clinical Psychology: Science and Practice, 17*(4), 293–306. https://doi.org/10.1111/j.1468-2850.2010.01221.x

Hancock, G. R., & Buehl, M. M. (2008). Second-order latent growth models with shifting indicators. *Journal of Modern Applied Statistical Methods, 7*(1), 39–55. https://doi.org/10.22237/jmasm/1209614640.

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*(1), 3–24. https://doi.org/10.1177/0146621602026001001

Hertzog, C., & Nesselroade, J. R. (2003). Assessing psychological change in adulthood: An overview of methodological issues. *Psychology and Aging, 18*(4), 639–657. https://doi.org/10.1037/0882-7974.18.4.639

Horn, J. L., & Blankson, N. (2005). Foundations for better understanding of cognitive abilities. In *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 41–68). The Guilford Press.

Horn, J. L., & Donaldson, G. (1980). Cognitive development in adulthood. In J. Kagan & J. Brim (Eds.), Constancy and change in human development (pp. 445–529). https://www.hup.harvard.edu/catalog.php?isbn=9780674166257.

Insel, T. R. (2014). Mental disorders in childhood: Shifting the focus from behavioral symptoms to neurodevelopmental trajectories. *JAMA, 311*(17), 1727–1728. https://doi.org/10.1001/jama.2014.1193

Kagan, J. (1969). The three faces of continuity in human development. In D. A. Goslin (Ed.), Handbook of socialization theory and research (pp. 983–1002). Rand McNally. https://works.swarthmore.edu/alum-books/2913/.

Kagan, J. (1971). *Change and continuity in infancy*. Wiley. https://psycnet.apa.org/record/1973-10863-000.

Kagan, J. (1980). Perspectives on continuity. In O. G. Brim, & J. Kagan (Eds.), *Constancy and change in human development* (pp. 26–74). Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674166257.

Kagan, J., & Moss, H. A. (1962). *Birth to maturity: A study in psychological development*. Wiley. https://doi.org/10.1037/13129-000

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal, 15*(1), 136–153. https://doi.org/10.1080/10705510701758406

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*(3), 383–400. https://doi.org/10.1177/0265532211404190
Khoo, S.-T., West, S. G., Wu, W., & Kwok, O.-M. (2006). Longitudinal methods. In M. Eid, & E. Diener (Eds.), *Handbook of multimethod measurement in psychology*. American Psychological Association. https://doi.org/10.1037/11383-021.
Knight, G. P., & Zerr, A. A. (2010). Informed theory and measurement equivalence in child development research. *Child Development Perspectives, 4*(1), 25–30. https://doi.org/10.1111/j.1750-8606.2009.00112.x
Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking: Methods and practices (3rd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7.
Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., … Zimmerman, M. (2017). The hierarchical taxonomy of psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology, 126*(4), 454–477. https://doi.org/10.1037/abn0000258
Lahey, B. B., Zald, D. H., Hakes, J. K., Krueger, R. F., & Rathouz, P. J. (2014). Patterns of heterotypic continuity associated with the cross-sectional correlational structure of prevalent mental disorders in adults. *JAMA Psychiatry, 71*(9), 989–996. https://doi.org/10.1001/jamapsychiatry.2014.359
Lavigne, J. V., Gouze, K. R., Bryant, F. B., & Hopkins, J. (2014). Dimensions of oppositional defiant disorder in young children: Heterotypic continuity with anxiety and depression. *Journal of Abnormal Child Psychology, 42*(6), 937–951. https://doi.org/10.1007/s10802-014-9853-1
LeBeau, B. (2017). Ability and prior distribution mismatch: An exploration of common-item linking methods. *Applied Psychological Measurement, 41*(7), 545–560. https://doi.org/10.1177/0146621617707508
Lee, K., Bull, R., & Ho, R. M. H. (2013). Developmental changes in executive functioning. *Child Development, 84*(6), 1933–1953. https://doi.org/10.1111/cdev.12096
Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007). New developments in latent variable panel analyses of longitudinal data. *International Journal of Behavioral Development, 31*(4), 357–365. https://doi.org/10.1177/0165025407077757
Livson, N. (1973). Developmental dimensions of personality: A life-span formulation. In P. B. Baltes & K. W. Schaie (Eds.), Life-span developmental psychology (pp. 97–122). Academic Press. https://doi.org/10.1016/B978-0-12-077150-9.50011-3.
Lubke, G. H., McArtor, D. B., Boomsma, D. I., & Bartels, M. (2018). Genetic and environmental contributions to the development of childhood aggression. *Developmental Psychology, 54*(1), 39–50. https://doi.org/10.1037/dev0000403
Magnus, B. E., Willoughby, M. T., Blair, C. B., & Kuhn, L. J. (2019). Integrating item accuracy and reaction time to improve the measurement of inhibitory control abilities in early childhood. *Assessment, 26*(7), 1296–1306. https://doi.org/10.1177/1073191117740953
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). Springer Science & Business Media. https://doi.org/10.1007/978-0-387-98138-3_5.
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*(2), 126–149. https://doi.org/10.1037/a0015857
McDermott, P. A., Watkins, M. W., Rovine, M. J., & Rikoon, S. H. (2013). Assessing changes in socioemotional adjustment across early school transitions-New national scales for children at risk. *Journal of School Psychology, 51*(1), 97–115. https://doi.org/10.1016/j.jsp.2012.10.002
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*(1), 164–180. https://doi.org/10.1111/j.1467-9507.2008.00480.x
Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives, 4*(1), 5–9. https://doi.org/10.1111/j.1750-8606.2009.00109.x
Millsap, R. E. (2011). Statistical approaches to measurement invariance. *Taylor & Francis*. https://doi.org/10.4324/9780203821961
Moeller, J. (2015). A word on standardization in longitudinal studies: Don't. *Frontiers in Psychology, 6*(1389), 1–4. https://doi.org/10.3389/fpsyg.2015.01389
Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review, 100*(4), 674–701. https://doi.org/10.1037/0033-295X.100.4.674
Molenaar, P. C. M., & Nesselroade, J. R. (2012). Merging the idiographic filter with dynamic factor analysis to model process. *Applied Developmental Science, 16*(4), 210–219. https://doi.org/10.1080/10888691.2012.722884
Murayama, K., Pekrun, R., Lichtenfeld, S., & vom Hofe, R. (2013). Predicting long-term growth in students' mathematics achievement: The unique contributions of motivation and cognitive strategies. *Child Development, 84*(4), 1475–1490. https://doi.org/10.1111/cdev.12036
Nagin, D. S., & Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods, 6*(1), 18–34. https://doi.org/10.1037/1082-989X.6.1.18
Nesselroade, J. R., & Estabrook, R. (2009). Factor invariance, measurement, and studying development over the lifespan. In H. B. Bosworth, & C. Hertzog (Eds.), *Aging and cognition: Research methodologies and empirical advances* (pp. 39–52). American Psychological Association. https://doi.org/10.1037/11882-002.
Nesselroade, J. R., Gerstorf, D., Hardy, S. A., & Ram, N. (2007). Idiographic filters for psychological constructs. *Measurement, 5*(4), 217–235. https://doi.org/10.1080/15366360701741807
Nesselroade, J. R., & Molenaar, P. C. M. (2016). Some behaviorial science measurement concerns and proposals. *Multivariate Behavioral Research, 51*(2–3), 396–412. https://doi.org/10.1080/00273171.2015.1050481
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*(6), 2925–2938. https://doi.org/10.1177/0962280214534588
Olson, S. L., Bates, J. E., Sandy, J. M., & Lanthier, R. (2000). Early developmental precursors of externalizing behavior in middle childhood and adolescence. *Journal of Abnormal Child Psychology, 28*(2), 119–133. https://doi.org/10.1023/A:1005166629744
Olson, S. L., Sameroff, A. J., Davis-Kean, P., Lansford, J. E., Sexton, H. R., Bates, J. E., … Dodge, K. A. (2013). Deconstructing the externalizing spectrum: Growth patterns of overt aggression, covert aggression, oppositional behavior, impulsivity/inattention and emotion dysregulation between school entry and early adolescence. *Development and Psychopathology, 25*(3), 817–842. https://doi.org/10.1017/S0954579413000199
Owens, E. B., & Shaw, D. S. (2003). Predicting growth curves of externalizing behavior across the preschool years. *Journal of Abnormal Child Psychology, 31*(6), 575–590. https://doi.org/10.1023/a:1026254005632
Patterson, G. R. (1993). Orderly change in a stable world: The antisocial trait as a chimera. *Journal of Consulting and Clinical Psychology, 61*(6), 911–919. https://doi.org/10.1037/0022-006X.61.6.911
Peabody, M. R. (2020). Practical issues in linking and equating with small samples. *Applied Measurement in Education, 33*(1), 1–2. https://doi.org/10.1080/08957347.2019.1674306
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*(3), 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*(5), 598–612. https://doi.org/10.1037/pas0000185
Petersen, I. T., Bates, J. E., McQuillan, M. E., Hoyniak, C. P., Staples, A. D., Rudasill, K. M., Molfese, D. L., & Molfese, V. J. (under revise and resubmit). Heterotypic continuity of inhibitory control in early childhood: Evidence from four widely used measures. Developmental Psychology.
Petersen, I. T., Hoyniak, C. P., McQuillan, M. E., Bates, J. E., & Staples, A. D. (2016). Measuring the development of inhibitory control: The challenge of heterotypic continuity. *Developmental Review, 40*, 25–71. https://doi.org/10.1016/j.dr.2016.02.001
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., & LeBeau, B. (in press). Creating a developmental scale to chart the development of psychopathology with different informants and measures across time. Journal of Abnormal Psychology.
Petersen, I. T., LeBeau, B., & Choe, D. E. (in press). Creating a developmental scale to account for heterotypic continuity in development: A simulation study. Child Development. https://doi.org/10.1111/cdev.13433.

Petersen, I. T., Lindhiem, O., LeBeau, B., Bates, J. E., Pettit, G. S., Lansford, J. E., & Dodge, K. A. (2018). Development of internalizing problems from adolescence to emerging adulthood: Accounting for heterotypic continuity with vertical scaling. *Developmental Psychology, 54*(3), 586–599. https://doi.org/10.1037/dev0000449

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*(6), 2136–2156. https://doi.org/10.1111/cdev.12880

Pickles, A., & Hill, J. (2006). Developmental pathways. In D. Cicchetti & D. J. Cohen (Eds.), Developmental psychopathology: Theory and method (2nd ed., Vol. 1, pp. 211–243). John Wiley & Sons, Inc. https://psycnet.apa.org/record/2006-03613-000.

Putnam, S. P., Rothbart, M. K., & Gartstein, M. A. (2008). Homotypic and heterotypic continuity of fine-grained temperament during infancy, toddlerhood, and early childhood. *Infant & Child Development, 17*(4), 387–405. https://doi.org/10.1002/ICD.582

Raykov, T., Marcoulides, G. A., Harrison, M., & Zhang, M. (in press). On the dependability of a popular procedure for studying measurement invariance: A cause for concern? Structural Equation Modeling: A Multidisciplinary Journal. https://doi.org/10.1080/10705511.2019.1610409.

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. *Child Development, 72*(5), 1394–1408. https://doi.org/10.1111/1467-8624.00355

Rueda, M. R., Posner, M. I., & Rothbart, M. K. (2005). The development of executive attention: Contributions to the emergence of self-regulation. *Developmental Neuropsychology, 28*(2), 573–594. https://doi.org/10.1111/1467-8624.00392

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94*(1), 18–38. https://doi.org/10.1037/0033-2909.94.1.18

Rutter, M., & Sroufe, L. A. (2000). Developmental psychopathology: Concepts and challenges. *Development and Psychopathology, 12*(03), 265–296. https://doi.org/10.1017/S0954579400003023

Schaie, K. W. (2000). The impact of longitudinal studies on understanding development from young adulthood to old age. *International Journal of Behavioral Development, 24*(3), 257–266. https://doi.org/10.1080/01650250050118231

Schleider, J. L., Krause, E. D., & Gillham, J. E. (2014). Sequential comorbidity of anxiety and depression in youth: Present knowledge and future directions. *Current Psychiatry Reviews, 10*(1), 75–87. https://doi.org/10.2174/1573400509666131217010652

Schulenberg, J., Patrick, M. E., Maslowsky, J., & Maggs, J. L. (2014). The epidemiology and etiology of adolescent substance use in developmental perspective. In M. Lewis, & K. D. Rudolph (Eds.), *Handbook of Developmental Psychopathology* (pp. 601–620). US: Springer. https://doi.org/10.1007/978-1-4614-9608-3_30.

Schulenberg, J. E., & Maslowsky, J. (2009). Taking substance use and development seriously: Developmentally distal and proximal influences on adolescence drug use. *Monographs of the Society for Research in Child Development, 74*(3), 121–130. https://doi.org/10.1111/j.1540-5834.2009.00544.x

Schulenberg, J. E., & Zarrett, N. R. (2006). Mental health during emerging adulthood: Continuity and discontinuity in courses, causes, and functions. In *Emerging adults in America: Coming of age in the 21st century* (pp. 135–172). American Psychological Association. https://doi.org/10.1037/11381-006.

Shaw, D. S., Hyde, L. W., & Brennan, L. M. (2012). Early predictors of boys' antisocial trajectories. *Development and Psychopathology, 24*(3), 871–888. https://doi.org/10.1017/S0954579412000429

Shaw, D. S., Lacourse, E., & Nagin, D. S. (2005). Developmental trajectories of conduct problems and hyperactivity from ages 2 to 10. *Journal of Child Psychology and Psychiatry, 46*(9), 931–942. https://doi.org/10.1111/j.1469-7610.2004.00390.x

Snyder, H. R., Young, J. F., & Hankin, B. L. (2017). Strong homotypic continuity in common psychopathology-, internalizing-, and externalizing-specific factors over time in adolescents. *Clinical Psychological Science, 5*(1), 98–110. https://doi.org/10.1177/2167702616651076

Sroufe, L. A. (1979). The coherence of individual development: Early care, attachment, and subsequent developmental issues. *American Psychologist, 34*(10), 834–841. https://doi.org/10.1037/0003-066X.34.10.834

Sroufe, L. A. (2009). The concept of development in developmental psychopathology. *Child Development Perspectives, 3*(3), 178–183. https://doi.org/10.1111/j.1750-8606.2009.00103.x

Sroufe, L. A. (2013). The promise of developmental psychopathology: Past and present. *Development and Psychopathology, 25*(4pt2), 1215–1224. https://doi.org/10.1017/S0954579413000576

Sroufe, L. A., & Jacobvitz, D. (1989). Diverging pathways, developmental transformations, multiple etiologies and the problem of continuity in development. *Human Development, 32*(3–4), 196–203. https://doi.org/10.1159/000276468

Sroufe, L. A., & Rutter, M. (1984). The domain of developmental psychopathology. *Child Development, 55*(1), 17–29. https://doi.org/10.2307/1129832

Sterba, S. K., Prinstein, M. J., & Cox, M. J. (2007). Trajectories of internalizing problems across childhood: Heterogeneity, external validity, and gender differences. *Development and Psychopathology, 19*(2), 345–366. https://doi.org/10.1017/S0954579407070174

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*(1), 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227–253. https://doi.org/10.1080/08957340701301207

Tyrell, F. A., Yates, T. M., Widaman, K. F., Reynolds, C. A., & Fabricius, W. V. (2019). Data harmonization: Establishing measurement invariance across different assessments of the same construct across adolescence. *Journal of Clinical Child & Adolescent Psychology, 48*(4), 555–567. https://doi.org/10.1080/15374416.2019.1622124

Wakschlag, L. S., Tolan, P. H., & Leventhal, B. L. (2010). Research Review: 'Ain't misbehavin': Towards a developmentally-specified nosology for preschool disruptive behavior. *Journal of Child Psychology and Psychiatry, 51*(1), 3–22. https://doi.org/10.1111/j.1469-7610.2009.02184.x

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*(4), 383–407. https://doi.org/10.1504/IJQRE.2013.058307

Ward, C., Oleson, J., Tomblin, J. B., & Walker, E. (in press). Modeling population and subject-specific growth in a latent trait measured by multiple instruments over time using a hierarchical bayesian framework. Journal of Applied Statistics. https://doi.org/10.1080/02664763.2020.1817346.

Weems, C. F. (2008). Developmental trajectories of childhood anxiety: Identifying continuity and change in anxious emotion. *Developmental Review, 28*(4), 488–502. https://doi.org/10.1016/j.dr.2008.01.001

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., … Gershon, R. C. (2013). Cognition assessment using the NIH Toolbox. *Neurology, 80*(11 Supplement 3), S54–S64. https://doi.org/10.1212/WNL.0b013e3182872ded

Weiss, B., & Garber, J. (2003). Developmental differences in the phenomenology of depression. *Development and Psychopathology, 15*(2), 403–430. https://doi.org/10.1017/S0954579403000221

West, S. G., & Ryu, E. (2007). Commentary: Assumptions and challenges of an idiographic-nomothetic approach to measurement: A comment on Nesselroade, Gerstorf. *Hardy and Ram. Measurement, 5*(4), 259–263. https://doi.org/10.1080/15366360701775995

Widaman, K. F. (1991). Qualitative transitions amid quantitative development: A challenge for measuring and representing change. In J. L. Horn, & L. M. Collins (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 204–217). American Psychological Association. https://doi.org/10.1037/10099-013.

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*(1), 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Widaman, K. F., Little, T. D., Geary, D. C., & Cormier, P. (1992). Individual differences in the development of skill in mental addition: Internal and external validation of chronometric models. *Learning and Individual Differences, 4*(3), 167–213. https://doi.org/10.1016/1041-6080(92)90002-V

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology, 10*(2), 395–426. https://doi.org/10.1017/s0954579498001667

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*(1), 58–79. https://doi.org/10.1037/1082-989X.12.1.58

Yarrow, L. J., & Yarrow, M. R. (1964). Personality continuity and change in the family context. In P. Worchel, & D. Byrne (Eds.), *Personality change* (pp. 489–523). Wiley. https://psycnet.apa.org/record/1965-01691-000.