Developmental Cognitive Neuroscience 33 (2018) 54–72

Contents lists available at ScienceDirect

Developmental Cognitive Neuroscience

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

Longitudinal modeling in developmental neuroimaging research: Common challenges, and solutions from developmental psychology

Kevin M. King[a,*], Andrew K. Littlefield[b], Connor J. McCabe[a], Kathryn L. Mills[c], John Flournoy[c], Laurie Chassin[d]

[a] University of Washington, United States
[b] Texas Tech University, United States
[c] University of Oregon, United States
[d] Arizona State University, United States

ARTICLE INFO

Keywords:
Longitudinal methods
Growth curve models
Change over time

ABSTRACT

Hypotheses about change over time are central to informing our understanding of development. Developmental neuroscience is at critical juncture: although the majority of longitudinal imaging studies have observations with two time points, researchers are increasingly obtaining three or more observations of the same individuals. The goals of the proposed manuscript are to draw upon the long history of methodological and applied literature on longitudinal statistical models to summarize common problems and issues that arise in their use. We also provide suggestions and solutions to improve the design, analysis and interpretation of longitudinal data, and discuss the importance of matching the theory of change with the appropriate statistical model used to test the theory. Researchers should articulate a clear theory of change and to design studies to capture that change and use appropriately sensitive measures to assess that change during development. Simulated data are used to demonstrate several common analytic approaches to longitudinal analyses. We provide the code for our simulations and figures in an online supplement to aid researchers in exploring and plotting their data. We provide brief examples of best practices for reporting such models. Finally, we clarify common misunderstandings in the application and interpretation of these analytic approaches.

1. Introduction

Across multiple disciplines, longitudinal data have been used to elucidate the developmental course of a variety of phenomena and to test a host of hypotheses. As longitudinal data analysis (LDA) becomes more common in the neuroimaging field, imaging researchers can draw on other literatures with a long history of LDA, such as developmental psychology/psychopathology, to inform best practices. Many of our suggestions apply to all longitudinal data, although we highlight where neuroimaging samples may encounter unique challenges. The practicalities of implementing the analytic techniques discussed below have been covered in the extant literature (see Table 1). The purpose of this article, however, is to provide guidance on navigating the major challenges in the design, analysis, interpretation, and communication of longitudinal studies, and to clarify common misunderstandings in the application and interpretation of LDA.

Collins' (2006) seminal review outlined criteria for deciding whether a longitudinal design will provide a strong test of developmental hypotheses about change. These include: 1) a well-articulated theoretical model of change, 2) a design that is able to observe the change process in detail and 3) an analytical framework that operationalizes the theoretical model (Collins, 2006, p. 507). This manuscript follows this framework, with a focus on how design and analytic choices affect researchers' ability to understand developmental processes. We highlight key take-home points of this manuscript in Box 1.

2. Designing longitudinal studies

2.1. Articulate a theory of change

In articulating a theory of change, researchers should describe what change is expected to occur, how developmental processes influence that change, and when in development the change might be observed. First, researchers should decide what is meant by "change" over time. Ram and Grimm (2015) described a taxonomy of how change can unfold. They argue that most developmental studies test change that

* Corresponding author at: Box 351525, Department of Psychology, University of Washington, Seattle, WA, 98195, United States.
E-mail address: kingkm@uw.edu (K.M. King).

https://doi.org/10.1016/j.dcn.2017.11.009
Received 13 February 2017; Received in revised form 16 November 2017; Accepted 16 November 2017
Available online 22 November 2017

Table 1
Summary of statistical models available for addressing different hypotheses regarding change over time.

Number of Time points	Type of Within-Individual Change	Detectable Change	Available Statistical Models	Inferences	Cautions	Key References
One	None	Mean-level	Independent Samples T-test	Between-group mean differences	Confounded by cohort; no estimate of within-individual change	
			ANOVA	As above	As above	
			(Multiple) Regression, Generalized Linear Modeling	As above	As above	
Two	None	Rank-order	Repeated Measures ANOVA	Between-person change	Time treated as a fixed categorical (rather than continuous) predictor	
			(Multiple) Regression, Generalized Linear Modeling	Rank-order change	No estimate of within-individual change	
			Auto-regressive Panel	Rank-order change	No estimate of within-individual or average-level change	Selig and Little (2012)
			Latent Change Score	Latent rank-order change	"Stability" over time may not reflect a lack of change	Ferrer and McArdle, (2003, 2010), Grimm (2012), McArdle (2009)
Three	Linear	Rank-order	Repeated Measures ANOVA	Between-person change	Time treated as a fixed categorical (rather than continuous) predictor	
			Auto-regressive panel	Rank-order change	No estimate of within-individual or average-level change	Selig & Little (2012)
			Latent Change Score	Latent rank-order change		Ferrer and McArdle (2003, 2010), Grimm (2012), McArdle (2009)
		Within- and Between-individual	Multilevel Growth Curve	Between- and within-individual change	Residual variances of predictors at each time point are fixed	Bryk and Raudenbush (1987), Curran (2003), Schuster and von Eye (1998)
			Latent Growth Curve	Between and within-individual latent change; correlations between multiple growth processes		Bollen and Curran (2006), Curran and Hussong (2003)
			Latent Growth Mixture	Latent growth in multiple discrete populations	Sensitive to model specification	Bauer and Curran (2003), Muthén and Shedden (1999), Nagin (1999), Nylund et al. (2007), Ram and Grimm (2009)
			Trait-State	Continuous (trait) and discontinuous (state) change	Latent trait-models may be inappropriate when there is within-individual change over time	Cole et al., (2005), Kenny and Zautra (2001)
Four or more	Linear	Rank-order	Repeated measures ANOVA; auto-regressive panel; latent change score	Between-person and (latent) rank-change	True change process may be non-linear in form	
		Within- and Between-individual	Multilevel growth curve; latent growth curve; latent growth mixture; trait-state	Between and within-individual (latent) change; correlations between multiple growth processes; latent growth in multiple discrete populations; continuous (trait) and discontinuous (state) change	As above	
	Polynomial (e.g., Quadratic)	Non-linear	Multilevel growth curve; latent growth curve; latent growth mixture; trait-state	Non-linear within- and between-individual change		Biesanz et al. (2004), Grimm (2012)
	Piecewise	Discontinuous		Discrete patterns of within- and between-individual	Form of change may be unintuitive; more estimated parameters increases computational complexity of model	Flora (2008), McCoach and Kaniskan (2010)
	Latent basis ("free slope")	Freely-estimated		Model-derived form of within- and between-individual	As above	

Box 1
Key Points for Longitudinal Modeling.

It is important to match a theory of change with design and statistical model to assess change. (p. 4)
Models with two time points cannot provide estimates of within-person change. (p. 5, 15 − 16)
All longitudinal findings are bounded by the start and endpoints of analyses, the selection of temporal intervals, and by the number of time points in analyses (p. 6, 13, 21)
Standardizing *within* a time point actually *removes* time trends from longitudinal data. (p. 9)
It is important to carefully consider the psychometric properties of measures as they are administered over time. (p. 9–10)
"Stability" over time can reflect many things, only some of which mean a lack of change. (p. 14)
Difference scores can be reasonable estimates of change unless the measures have low reliability or high stability over time. (p. 15)
It is important to consider between person variability in repeated measures as a potential confound in auto-regressive models. (p. 18)
Multilevel and latent growth curve models are very similar, but have different strengths and weaknesses. (p. 19)
The intercept in a growth curve model represents the level of an outcome at one time point, and does not have to be the starting point of data collection. (p. 22)
Varying the intercept can provide different estimates of covariances among growth parameters, and will change the estimates of lower-order growth parameters when time is non-linear (e.g. quadratic, cubic). (p. 22 −23)
It is important to interpret all growth model parameters in the context of all others. (p. 25)
It is helpful to describe not only average growth (e.g. intercept and slope means), but variances in the growth parameters. (p. 25)
A non-significant effect of time (i.e. no average growth) does not preclude individual differences in growth (i.e. slope variances or random effect). (p. 26 − 27)
It is important to graph the estimated models of change against the observed data to gain greater insight into your models. (p. 25)
Requiring significant variation in slopes (i.e. random effects) prior to testing *a-priori* hypotheses about predictors of variation may result in under-powered hypothesis tests. (p. 28)
It is important to avoid reifying class solutions from mixture models because numerous methodological factors can influence the number and shape of classes that are found. (p. 32)
It is important to avoid relying on rules of thumb for model fit, and to consider that plausible alternative models may also fit the data well. (p. 36)

Box 2
Considering developmental peaks.

One example of matching theory to design is the search for developmental "peaks". That is, researchers are often interested when, on average, children, adolescents or adults are expected to show maximum levels of some construct, such as reward sensitivity (Braams et al., 2015), cognitive control (Ordaz et al., 2013), or cortical thickness (Walhovd et al., 2016), before those levels begin to decline. Assuming the phenomena is already measured at the correct time scale, peaks may be examined in a number of ways. Average developmental peaks may be discerned from cross-sectional studies with multiple age-cohorts, as has been done with sensation seeking and "self-regulation" (Steinberg et al., 2017). Longitudinal studies may go further by also estimating individual differences in the timing of those peaks, as well as factors that impact individual differences in that timing. Because the functional form of a peak may approximate quadratic growth, splines, or piecewise growth, longitudinal studies hoping to identify individual differences in the timing of developmental peaks require at least four and preferably five repeated observations from most individuals to be properly identified (Bollen and Curran, 2006). Moreover, it is important to include confidence intervals in any peak estimates to characterize the (un)certainty of such peaks.

unfolds as relatively smooth increases or decreases over time. For example, empirical data indicate that cortical thickness decreases roughly 1% per year during adolescence (Tamnes et al., 2017). Conversely, transformational change, where new abilities or characteristics emerge in relatively rapid transitions, or stability-maintenance processes (e.g., homeostatic systems) are less studied. To match developmental theory with the appropriate statistical model, multiple types of change processes, as well as multiple forms of those processes (such as linear, exponential, sigmoidal or spline), should be considered (Ram and Grimm, 2015). See Box 2 for one example of some challenges in matching theory with statistical models. Moreover, change can be assessed in terms of multiple indices (Roberts and Mroczek, 2008), including mean-level change, rank-order consistency (e.g., relative ordering of individuals over time), structural consistency (e.g., similar factor structures across time), and inter-individual differences in intra-individual change (individual differences in within-person change).

Theories about the forms of change should be supported by a theory about the processes that produce change. For example, neurodevelopmental theories attempt to explain developmental changes in risk taking and impulsivity based on the development of the pre-frontal cortex and subcortical structures including the striatum and amygdala (e.g., Casey and Caudle, 2013; Romeo, 2013). Many developmental models have yet to appear in the developmental neuroimaging literature, such as transactional models, bioecological models, or developmental cascade models (e.g., Bronfenbrenner, 1977; Masten and Cicchetti, 2010). For example, "coercion theory" is a transactional model positing that negative parental responses to child misbehavior can be negatively reinforced by the elimination of misbehavior, while accidentally positively reinforcing the misbehavior through the provision of attention (Dishion et al., 1992). Through this cycle of reinforcement, each behavior escalates over time, leading to increasingly maladaptive behaviors by both parents and children (Dishion et al., 1992; Granic and Patterson, 2006). Drawing upon a strong theory that explains how individual differences might emerge over time provides a critical foundation for longitudinal research.

2.2. Design a study to assess change

The next step is to design a study to reflect the theory of change. This includes decisions related to the timing, frequency, and spacing of observations in a longitudinal study. For example, researchers may measure cortical thickness annually across adolescence because it is

Box 3
On overfitting trajectories and the number of time points.

As several recent reviews have noted, neuroimaging studies are only now beginning to acquire samples with more than a single observation (Crone and Elzinga, 2015), Vijayakumar et al., this issue), and studies with three or more observations are exceedingly rare. However, this has not precluded imaging researchers from estimating longitudinal trajectory models in neuroimaging data using samples with only two observations. In these studies, children and adolescents are sampled at different ages (as in an accelerated cohort design) for both a baseline and follow up assessment. Researchers then attempt to estimate both the average trajectory of the outcome by comparing different polynomial shapes (such as linear, quadratic, cubic), as well as testing for individual differences in trajectories.

Most textbooks on longitudinal modeling (Bollen and Curran, 2006; Raudenbush and Bryk, 2002) make it clear that at least three time points are needed to fit a linear trajectory, four (preferably five) to fit a quadratic, and more than five are needed to fit cubic and higher order polynomials. Although one can draw a straight line between only two points, such a line is fitted without any error, and thus individual trajectory estimates will be overly precise. Bollen and Curran (2006) describe the issue of model identification for latent growth models in detail. A model is identified if there is a unique solution for all model parameters. Although identification in SEM can be established in a number of ways (Bollen, 1989), a brief rule of thumb is that fewer parameters must be estimated by a model than are available from the mean and variance/covariance structure of the data. Thus data with two time points provides information about 2 means, 2 variances, and 1 covariance; data with three time points provides 9 unique pieces of information, and so on. A simple linear growth model requires estimation of two means (intercept and slope), two random effect variances (intercept and slope), one covariance (between intercept and slope) and one residual variance. As such, at least three time points are required for a linear latent growth curve model (and any other latent variable approach) to be estimable.

Although multilevel models do not have the same identification requirements as SEMs, growth curve models in MLMs have the same requirements. Although fixed effects of time may be estimable and reasonably accurate with only two time points (provided enough different ages are sampled), the random effects will be unreliable. We provide a demonstration of this in our supplemental material. In our simulations of data with two time points drawn from a population where linear growth was true, linear random effects were only estimable 60% of the time, although fixed effects were accurately estimated. When quadratic growth was the true model, quadratic random effects were never estimable with only two time points, and the linear random effect was mis-estimated (and only 80% of the time), although all fixed effects were estimated without error.

Thus, it is critical that longitudinal imaging studies avoid estimating trajectories for which there are insufficient data to estimate.

expected to change slowly and smoothly during this time period (Tamnes et al., 2017). It is important that the timing of assessments is driven by a theory about the timescale (e.g., yearly, monthly, daily) at which change is thought to occur, as well as the metric of time (e.g. age, grade, time since an event; see Ram and Grimm, 2015 for an excellent overview of these issues). If assessments are spaced too far apart, change could be missed entirely; if they are too frequent, a study could be unduly expensive or hampered by participant reactivity (King et al., 2006).

Study design also limits the type of change that can be modeled. Cross-sectional studies can infer developmental change by measuring individuals with the same measure at specific ages (e.g., Ostby et al., 2009; Somerville et al., 2013), but inferences about age are confounded by cohort and period (Glenn, 1976), and certain examinations of change (e.g., rank-order consistency) are not possible within this design. Studies with two time points can measure within-person change, but only in terms of rank-order changes in levels of a variable. As Rogosa et al. wrote “two waves of data are better than one, but maybe not much better” (Rogosa et al., 1982, p. 744). Conversely, models with more time points allow for the study of both within- and between-person change. Importantly, as the number of data collection points increases beyond a single time point, the ability to model types of change increases substantially, as does the number of available models.

In terms of the number of time points, longitudinal functional neuroimaging data with children and adolescents are relatively rare, with one review reporting 13 longitudinal imaging studies (Crone and Elzinga, 2015). In that report, most studies had data from only two time points, and only a minority of studies followed subjects beyond three time points, and only for subsets of the sample. Although there have been more longitudinal structural imaging studies, with 34 in one recent review (Vijayakumar et al., this issue), only three of those studies reviewed included an average of three or more time points (i.e., scans) from participants, and again, those with more than two observations were usually a subset of the whole sample. Despite this, many of these studies have reported inappropriate longitudinal analyses, such as testing for individual differences in trajectories without enough time points for trajectories to be over-identified (that is, more information is observed than estimated). See Box 3 for a detailed discussion of this issue.

Notably, any observational study is only *sampling* some of the many possible time points that may provide a representation of a larger developmental process. All findings in longitudinal studies are bounded by the start and endpoints of analyses, the selection of temporal intervals, and by the number of time points in analyses. Prior studies have demonstrated that the selection of time points can influence the results from LDA (see, for example, Jackson and Sher, 2006; Rogosa, 1988). Researchers should avoid falling prey to the “jingle” fallacy, where it is assumed that studies cover similar developmental periods because they use the same label (such as “adolescence”) but actually measure different age spans with different time points at different intervals. Rather, each of these factors may have a dramatic impact on the information a model provides. Researchers should carefully consider how their findings may or may not converge with other studies that cover similar age spans. For example, a recent study analyzed longitudinal structural imaging data from four independent samples. Analyses in one dataset, which found *linear* trajectories of change in gray and white matter volume from ages 10–16, actually converged with three other datasets which found cubic trajectories between ages 10 and 30, because both types of trajectories described very similar change during the same developmental period (Mills et al., 2016). (see, for example, Jackson and Sher, 2006; Rogosa, 1988)

Other work highlights the impact of the *coding* of time in LDA, illustrating the importance of matching the coding of time with the assessment frame. If assessments were collected, for example, at a baseline, 6 months, 12 months, and 24 months, time should be coded 0, 1, 2, 4 (not 0, 1, 2, 3), so that 1 unit of time represents the passage of 6 months. Mis-specifying time can also dramatically impact inferences, coefficient estimates, and interpretability of the model (Biesanz et al., 2004; Grimm, 2012; Raudenbush and Bryk, 2002; Schuster and von Eye, 1998).

Thus, it is critical to match the design of the collected data with the hypotheses that researchers wish to test (Collins and Graham, 2002), particularly with secondary data (Brooks-Gunn et al., 1991; Davis-Kean et al., 2015; Greenhoot and Dowsett, 2012; McCall and Appelbaum, 1991). For example, several longitudinal studies of personality change (Littlefield et al., 2010a; Littlefield et al., 2009; Littlefield et al., 2012; Littlefield et al., 2010b) used data originally collected to track developmental change in drinking, though there is no guarantee that personality change follows the same time course of drinking. Sometimes the assessment frame reflects external influences (e.g., available funding for additional assessments) rather than "ideal" designs based on scientific considerations.

2.3. Consider measurement over time

Once the overall longitudinal study is designed, measurement considerations arise that are uncommon for cross-sectional or experimental studies (Grimm et al., 2013). Tests of longitudinal hypotheses about change assume that a study uses measures that are a) valid, b) measure the appropriate time-span, and c) have the same psychometric properties over time. Models of change assume that a one unit change in the observed data reflects the same amount of change in the underlying construct across all time points. This relatively simple assumption raises a number of issues to be considered in longitudinal modeling. First, any data transformations must carefully consider how the transformation may impact the relative scale of the construct across assessments. Standardizing relative to one time point, or all time points, will yield standardized parameters that are equivalent to results with untransformed data. However, many authors have noted that standardizing *within* a time point actually *removes* time trends from longitudinal data (Bollen and Curran, 2006; Stoolmiller, 1995). But not all age-standardized data may be bad; some age-norming (such as with the Child Behavior Checklist; Achenbach and Edelbrook, 1983) actually produces data where a one unit change in the data reflects a similar magnitude of change in the construct across all ages where the raw data does not.

Specific to neuroimaging, methodological approaches should be also used to reduce measurement error, such as maintaining consistency in MRI equipment, software, hardware, and data processing steps (see Vijayakumar et al., this issue).

Psychometrics, which aims to model how observed data are related to the underlying construct, may be especially problematic in the field of neuroimaging. Many of the cognitive and behavioral tasks that are used as stimuli are either ad-hoc or have limited psychometric data. Psychometric analysis is also required to ensure that the measures exhibit structural consistency across time (i.e., for a given construct, the relations between the latent, unobserved construct and the variables used to measure it are consistent across time Pitts et al., 1996). For example, researchers using a Go/No Go task might be concerned about ceiling effects as children get better at the task through either practice or age. If an adaptive threshold is used (i.e., the task becomes more difficult as children improve) used, scores will not be comparable across age. Rather, age-related changes will reflect a mixture of a child's skill improvement and changes in task difficulty (Hamilton et al., 2015). It is important, then, to use measures that are suitable for as broad an age range as possible.

Ensuring structural consistency across time is difficult for behavioral or cognitive tasks, which are often used as a correlate in structural imaging studies or to evoke brain responses in functional imaging studies. Compared to research based on measurement approaches that are amenable to latent variable modeling (e.g., survey data) and includes a relatively mature set of psychometric models (Vandenberg and Lance, 2000), rigorous psychometric evaluations are generally lacking within neuroimaging research. For example, most cognitive tasks (those measuring attention, memory, and cognitive control) included in the Research Domain Criteria (RDoC) were noted by the RDoC committee to have a lack of psychometric data and no standardized administration parameters (National Advisory Mental Health Council Workgroup on Tasks and Measures for Research Domain Criteria (RDoC), 2016), except for a few recent studies of very little is known about the psychometric properties of many cognitive and behavioral tasks used in neuroimaging studies beyond test-retest reliability (c.f., Weafer et al., 2013; Wöstmann et al., 2013). The relative infancy of psychometrics in behavioral tasks used in neuroimaging is not surprising, given the application of traditional psychometric models is much more difficult and expensive compared to survey data.

Further, attempts to utilize psychometric approaches within the cognitive literature have yielded mixed success. For example, the oft-cited latent variable analysis of executive functions (Miyake et al., 2000) identified a latent inhibition trait, which was reflected by an antisaccade, stop-signal, and Stroop task (tasks which are frequently used in fMRI studies). However, subsequent psychometric evaluations cast doubt on the robustness of the originally proposed measurement model (Miyake and Friedman, 2012). Thus, researchers are cautioned against over-reifying constructs in the absence of strong psychometric support.

The extent to which even test-retest reliability reflects practice effects versus true stability (Salthouse, 2014) remains clouded. Many cognitive and behavioral tasks (such as the Erikson Flanker, Stroop, stop-signal, Go/No-Go) may suffer from low between subject variability precisely because they were designed to produce an experimental response from most individuals, rather than to detect individual differences in the magnitude of that response (Hedge et al., 2017). In sum, compared to work based on (primarily self-reported) survey data, robust psychometric evaluations of (largely laboratory-based) tasks often used in neuroimaging are less common. Additional psychometric work, and the development of novel paradigms designed to maximally detect between subject variability, will be required in the neuroimaging field in order to determine whether fundamental prerequisites of LDA are met by these tasks.

3. Improving longitudinal analysis

Finally, researchers are charged with matching the theoretical model of change with the appropriate statistical model. As noted by Collins (2006), "A mismatch of theoretical and statistical models will result in the addressing of irrelevant or even meaningless scientific questions. On the other hand, a close correspondence between theoretical and statistical model can provide an elegant test of a scientific hypothesis and a penetrating look at longitudinal data." (p. 509).

To illustrate common applications of longitudinal models, we simulated a large dataset based on the same basic motivating example, with modifications to the underlying model throughout to illustrate different points about statistical modeling. We simulated data with 10,000 observations so the parameter estimates from our models would approach their true (population) values. For most examples, random samples of 250 participants were used to approximate typical sample sizes in developmental neuroimaging research. The primary outcome of interest is the development of cortical thickness of the right inferior frontal cortex (rIFC), measured annually from age 10–19. The rIFC has been associated with inhibitory control (Aron et al., 2014). We modeled quadratic declines in cortical thickness across adolescence (Tamnes et al., 2017). We assumed that all measurement parameters of the rIFC were identical over time. We also simulated a variable representing a behavioral measure of impulse control (such as performance in a Go/No-Go task) at age 10. Finally, we included a time-varying covariate reflecting the effects of stress on rIFC thickness.

Fig. 1 shows the "true" development of rIFC thickness over time in a subset of cases from these simulated data. As shown in Fig. 1, rIFC thickness exhibited rapid declines in early adolescence before those declines slowed during the latter part of adolescence (i.e., change in rIFC thickness demonstrated a quadratic trend). We injected substantial

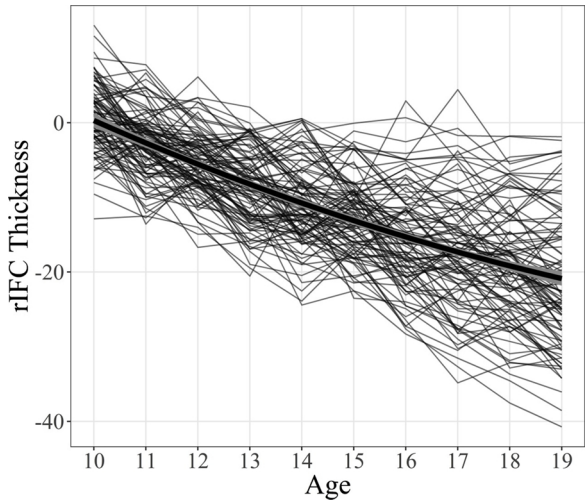

Fig. 1. True change over time in rIFC thickness in the simulated dataset.
Caption: This figure illustrates average change over time in the outcome in a subset of 100 cases, as well as individual lines connecting observations across each time point (reflecting variability in trajectories over time).

intra-individual variation in all aspects of change (i.e., intercept, linear slope and quadratic effects), and allowed this variation to be correlated (r = .20 across all parameters).

3.1. Models for two time points

With two time points of data, general linear models (GLMs) can be used to answer several basic questions regarding change, including the extent to which between-person variability in rIFC thickness at one age (say 10 years) is associated with between-person variability at a later age (say, 14 years). The extent to which the covariate (impulse control) at Age 10 predict individual differences in rIFC thickness at Age 14, adjusting for rIFC thickness at age 10, can also be explored.

3.1.1. Example

Table 2 presents the unstandardized intercept and the standardized coefficients from our simulated dataset. This model was estimated with ordinary least squares regression. When predicting rIFC thickness at Age 14, the intercept means that when Age 10 rIFC thickness, Age 10 impulse control, and Age 10 stress were all at 0, Age 14 rIFC thickness was predicted to be −9.84. If the value of the intercept were important, we could center the predictors so its value could be interpreted as the expected value of Age 14 rIFC thickness at the mean of the predictors. The effect of Age 10 rIFC thickness reflects the degree of rank-order stability of rIFC thickness from Age 10–14. A one SD difference between adolescents in rIFC thickness at Age 10 would be associated with a .35 SD difference between adolescents in rIFC thickness at Age 14. Similarly, a one SD difference between adolescents in impulse control at Age 10 would be associated with a .22 SD change in rIFC thickness at Age14, and so on.

Table 2 illustrates how changing the follow up interval (to Age 15 or Age 18) can change the coefficients we would observe, sometimes dramatically. For example, Age 10 impulse control would be considered to be a relatively minor correlate of rIFC thickness at Age 14 (β = .22), but a very major predictor of Age 18 rIFC thickness (β = .50). Although not illustrated in Table 2, the direction of a coefficient effect can even flip signs, depending on the time of follow-up. Thus, the inferences drawn from longitudinal effects are bounded by the assessment frame, as these effects can change (sometimes substantially) across time intervals.

Table 2
The impact of using different follow-up time points on stability and covariate estimates for regression with two time points. n = 250.

	Age 14 rIFC Thickness	Age 15 rIFC Thickness	Age 18 rIFC Thickness
Intercept	−9.839	−12.545	−20.221
Age 10 rIFC Thickness	0.347	0.304	0.330
Age 10 Impulse Control	0.217	0.318	0.496
Age 10 Stress	0.251	0.235	0.160

Note: This table presents unstandardized intercept values and standardized regression coefficients for two time point regression models predicting rIFC thickness at different ages from Age 10 covariates. Note how as the time interval between the predictor and the outcome changes from a 4 year lag to a 5 and 8 year lag the estimated effects of some predictors can change dramatically.

3.1.2. Cautions in considering stability over time

Importantly, rank-order stability does not preclude mean-level changes (and vice versa). Rank-order stability can result from a number of factors such as how the environment shapes and transacts with individuals to influence both change *and* stability (Fraley and Roberts, 2005). That is, stability could reflect stable *transactions* between an individual and a stable environmental context; given a different environment that stability may evaporate. Accurate inferences about stability require more than two time points, allowing inferences about the degree to which correlations across increasing increments of time asymptote towards a lower bound (Fraley and Roberts, 2005). Others have noted that "stability" coefficients could reflect the amount of true change, the degree of uniformity in change, or the relations between initial status and change over time (Hertzog and Nesselroade, 1987; Selig and Little, 2012). We illustrate this in Fig. 2, showing three ways high stability can manifest in samples with very different patterns of change, and contrast it with a sample exhibiting true variation in change and thus lower stability.

Stability and change may be even more complicated when considering repeated assessments of a cognitive task. In one large study, subjects who had prior experience with a cognitive task exhibited higher scores relative to subjects from the same cohort who were task-naïve (Salthouse, 2014), suggesting that repeated exposure to cognitive tasks may bias stability estimates due to task reactivity. For example, some data suggest various delay discounting measures demonstrate marked test-retest stability across various time intervals (e.g., Odum, 2011). In Kirby (2009), test-retest stability was high (rs > .63) though mean discount rates *increased* across assessments, suggesting (by the interpretation of the test) that the participants were becoming *more* impulsive. An alternate interpretation, however, is that participants were becoming increasingly reactive to the test (Salthouse, 2014).

3.1.3. Cautions in predicting outcomes from rank-order change

It is often of interest to connect rank-order change with some outcome, controlling for common covariates. Although more advanced methods have been recently developed to assist with this problem (McArdle, 2009), more common approaches are to compute difference scores (subtract Time 1 from Time 2 scores), or to residualize the Time 2 score by regressing Time 2 score onto Time 1 score. Burt and Obradović (2013) provide an excellent overview of the true strengths and limitations of these traditional approaches. In short, close attention should be paid to the impact that the reliability, variability, and correlation between the Time 1 and 2 scores have on the psychometric properties of difference and residual scores. Low reliability of either Time 1 or Time 2 score lowers the reliability of the change or residual score, whereas a high correlation between Time 1 and Time 2 scores causes lower reliability of difference scores (Burt and Obradović, 2013). As Rogosa pointed out in a classic chapter on myths on longitudinal

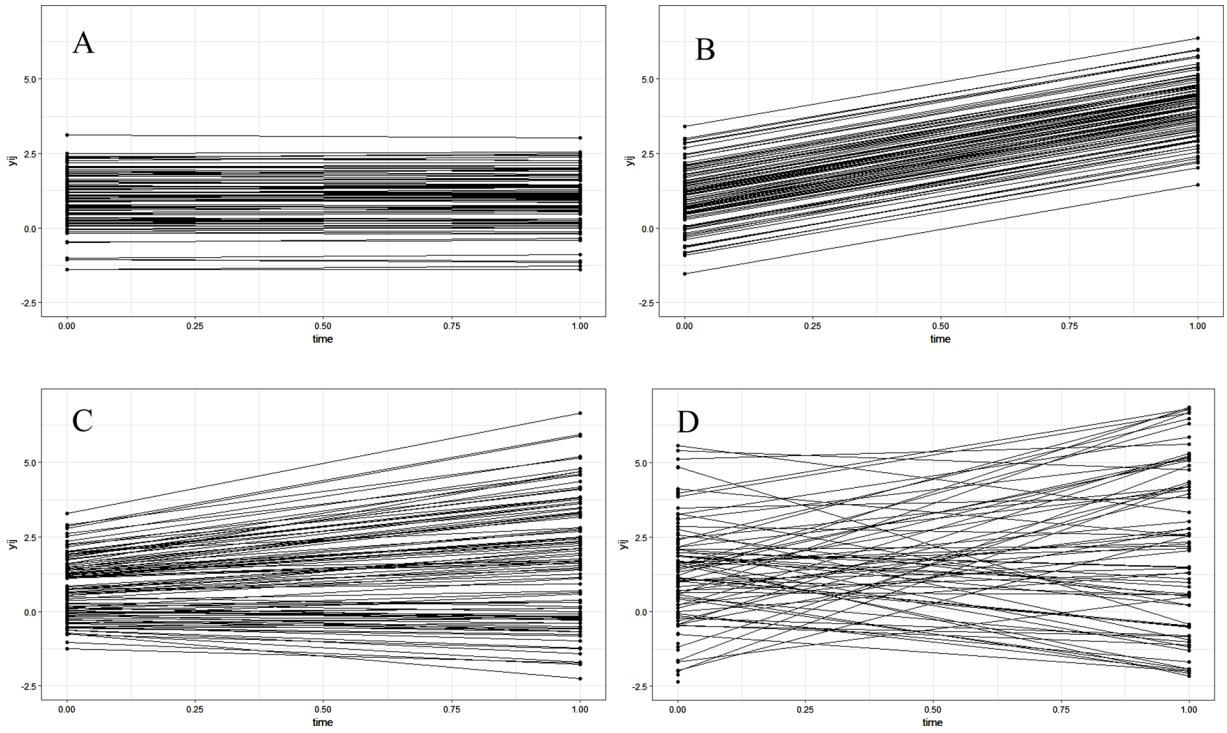

Fig. 2. Four different ways that stability (and "change") can manifest in regression and panel models.
Caption: Displayed are 100 cases from simulated data with two time points, where stability was manifested as A) no average change (β = .99), B) mean change over time, but no individual differences in change (β = .99), C) a high correlation between initial status and change (β = 0.97), contrasted to D) an "unstable" sample with variation in change (β = 0.44).

change (1988), difference scores are unreliable when stability is high for a very good reason: the higher the Time 1–Time 2 correlation, the less that there are individual differences in change, and "you can't detect individual differences that ain't there" (p. 13). As illustrated with two time points in Fig. 2, variability in change can dramatically impact estimates of stability.

3.2. Models for three or more time points

3.2.1. Auto-regressive panel models

Auto-regressive panel models are extensions of the GLM that predict the value of a variable in the future from the same variable in the past. They are based in a structural equation modeling (SEM) framework, and test how between-person differences in levels (i.e., rank-order individual differences) of a variable at one time point are predicted by between-person differences in that variable at a prior time point (Selig and Little, 2012). SEM software solves for a set of two equations simultaneously, in a way that maximizes the fit of the model (i.e., the estimated parameters) to the covariance matrix of the data. The two-time point models we presented above are examples of auto-regressive lag-one effects, where a variable at one time point predicts the same variable at the next observed time point. Auto-regressive lag-two (predicting between-person differences in levels from those two time-points prior) and greater lags are also possible.

These may be extended to two or more variables in cross-lagged panel models (CLPM), where between-person differences in one variable at a time point predict change in another variable at the next (Gollob and Reichardt, 1987; Hamaker et al., 2015). Importantly, these effects will be also expected to change as the time-lag changes, as we demonstrated above with the two time point data. As with two-time point regression models, because panel models do not typically incorporate information about means, inferences can only be made about individuals' standing relative to one another is related to change at the next time point (see equation 2 in Hamaker et al., 2015), but not within-individual or average change in the sample (see Curran et al., 2014 for a method that combines traditional latent curve models with cross-lagged panels).

3.2.1.1. Example. We estimated a bivariate auto-regressive cross-lag panel model for the association between rIFC thickness and stress at Ages 10, 12 and 14. Our auto-regressive panel model provided stability (i.e. auto-regressive) estimates of the effect of Age 10 rIFC thickness on Age 12, and Age 12 rIFC thickness on Age 14 rIFC thickness, and similar stability estimates for stress. We also estimated cross-lagged effects for the effects of Age 10 stress on Age 12 rIFC thickness (controlling for Age 10 rIFC thickness), Age 12 stress on Age 14 thickness, and similar effects for rIFC thickness on stress. For simplicity, we constrained similar effects (such as the stability of rIFC thickness between Age 10 and 12 and Ages 12 and 14, or the effects of stress on rIFC thickness at Ages 12 or 14) to be equal over time, which means we forced the parameters to be equal. This reflects a parsimonious assumption that these associations do not change across time, although this assumption may be relaxed. This model is illustrated in Fig. 3 as an SEM, where boxes represent observed variables, single headed arrows between variables represent regression slopes, and double headed arrows represent correlations. Coefficients are reported in the first column of Table 3, which again illustrates how coefficient estimates can change for the same three time point auto-regressive model with different scales of time.

The interpretation of the coefficients from these models was identical to that from the regression model above: how did between-person differences in the level of stress at one time point predict between-person differences in the level of rIFC thickness at the next, controlling for stability in rIFC thickness? Additionally, how does rIFC thickness at one time point predict stress at the next, controlling for stability in stress? Again, the time interval selected will impact inferences. For example, we would conclude that stress at the prior time point would be weakly or unrelated to rIFC thickness when examining the effect in Age 14, 16 and 18 data, but we would conclude it had a moderate and positive relation with rIFC thickness for the Age 10, 12 and 14 data.

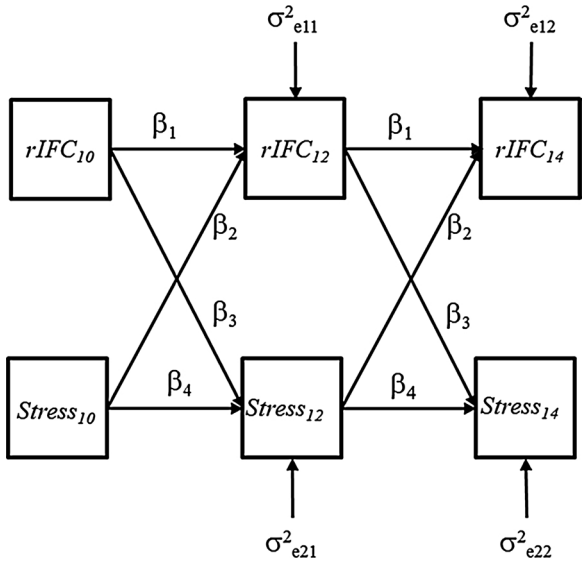

Fig. 3. A three time point cross-lagged panel model.
Note: This figure is a graphical model of an auto-regressive panel model of the relation between rIFC thickness and stress over time. For graphical depictions of structural equation models, boxes represent observed variables, circles (not displayed) represent unobserved (latent) variables, triangles (not displayed) represent estimated means, single headed arrows between variables represent regression slopes, and double headed arrows (not displayed) represent correlations. Each residual error term was estimated independently. IC = Impulse Control. rIFC = right Inferior Frontal Cortex. Coefficients are reported in Table 3.

Table 3
Autoregressive Panel Models at different time intervals. n = 250.

	Age 10, 12, 14	Age 14, 16, 18	Age 10, 14, 18
T2/T3 Outcome	rIFC	rIFC	rIFC
T3 Intercept	−6.39	−4.62	−9.44
T2 Intercept	−3.95	−4.73	−8.97
β_1 Lag 1 rIFC Thickness	0.44	0.71	0.47
β_2 Lag 1 Stress	0.28	0.14	0.24
T2/T3 Outcome	Stress	Stress	Stress
T3 Intercept	−0.15	0.41	0.74
T2 Intercept	−0.13	0.25	0.24
β_3 Lag 1 rIFC Thickness	0.08	0.03	0.08
β_4 Lag 1 Stress	0.73	0.86	0.78

Note: This table displays how coefficients for the cross-lagged panel model illustrated in Fig. 3 would change with different time intervals. Identical effects were fixed to be equal over time.

3.2.1.2. Cautions: considering trait variability. The CLPM has been specifically critiqued for mixing within and between-person variation, leading to mistaken inferences about the nature of transactional processes (Hamaker et al., 2015). For example, Ritchie et al. (2015) applied a CLPM to a sample of twins studied across five time points from ages 7–16, and found associations between reading and later measures of intelligence, even when controlling for prior observations of intelligence. In a reanalysis of the same data, Bailey and Littlefield (2017) showed that a state-trait model, which separated stable (i.e. trait), between-person variation in reading and intelligence over time from state variation in reading and intelligence that varied from time to time (Kenny and Zautra, 2001), could fit those data equally well (or better), and largely reduced the magnitude of the causal paths among reading and intelligence. These findings indicated that stable third variables that contributed to trait variability, such as common genetic or environmental factors, may largely account for the longitudinal association between reading and intelligence. Thus, it is important to carefully consider between-person (i.e., trait) variability in what is thought to be a time-varying construct as a special form of a third variable confound that is particularly pertinent to auto-regressive models.

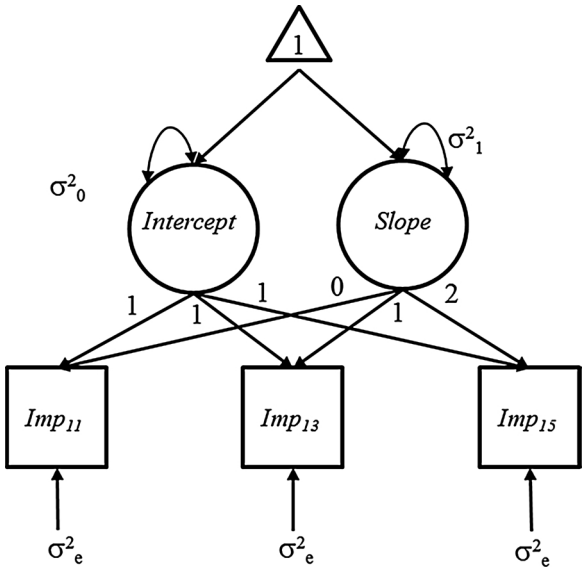

Fig. 4. A latent growth curve model of Impulse control.
Note: For graphical depictions of structural equation models, boxes represent observed variables, circles represent unobserved (latent) variables, triangles represent estimated means, single headed arrows between variables represent regression slopes, and double headed arrows represent correlations and variances.

3.2.2. Growth curve models

Researchers are often interested in modeling and explaining patterns of within-individual change (i.e., how a person changes relative to their own standing over time), as well as how their pattern of change may differ from that of another individual. Growth curve models (GCMs) are common approaches to modeling within-individual change, and can be approached from a multilevel or latent variable framework (see Curran, 2003 for a discussion of the similarities of the two approaches; Wood et al., 2015 for an in-depth review of the relation of various types of longitudinal models).

In multilevel GCMs (or random regression, random coefficient models), observations for individuals at each time point are predicted by a variable coding the passage of time (Bryk and Raudenbush, 1987). By allowing the coefficients of this model (such as the intercept, or level, and the effect of time, or the slope) to be "random" (i.e., assume it was sampled from a larger population in the same way we assume the *sample of individuals* was), we can also estimate between-person differences in those coefficients. These estimates inform the degree to which people exhibit individual differences in change over time. (Ballinger, 2004; McNeish et al., 2016; Zeger et al., 1988)

Latent curve models (LCMs) model the growth process (i.e., intercepts and slopes) as latent variables within an SEM framework (Bollen and Curran, 2006) and are similar to multilevel GCMs (Curran, 2003). In LCMs, the effects of time are coded into structural coefficients for the latent growth factors (i.e., intercept and slope), the average effects of time are estimated from latent means of the growth factors, and random effects (i.e., individual differences in intercepts and slopes) are estimated via the variances of the latent growth factors. LCMs may also be extended to multiple parallel processes, where two or more growth processes are estimated simultaneously, allowing for the estimation of inter-relations among growth factors (Cheong et al., 2003). An SEM representation of the latent growth curve model is illustrated in Fig. 4.

Despite similarities between LCMs and multilevel GCMs, each approach has distinct strengths. Because they model growth as latent variables, LCMs are very useful when researchers use individual

Table 4a
Multilevel Growth Model, Ages 10, 12 and 14, Fixed Effects. n = 250.

	b	S.E.	p-value
Intercept	2.63	0.34	0.000
Slope	−5.80	0.18	0.000
Age 10 Impulse Control	0.92	0.18	0.000
Stress	1.32	0.08	0.000
Slope X Age 10 Impulse Control	0.07	0.11	0.535

Random Effects

	Random Effects	Intercept – Slope Correlation
Intercept	2.21	−0.14
Slope	0.43	
Residual	2.98	

**Note*: This table displays growth model estimates for change in rIFC thickness over time for Ages 10, 12 and 14, conditional on the effects of the covariates.

differences in growth change as a predictor. Moreover, LCMs allow for the estimation of more complicated forms of growth, such as the latent basis model, where the growth function is not completely known (Flora, 2008; McCoach and Kaniskan, 2010). Conversely, multilevel GCMs allow for flexible time specification, such that all participants can have unique time points (such as age measured in days); because time is a factor loading for LCMs, they are much more restricted.

3.2.2.1. Example. Taking our model of rIFC thickness at Ages 10, 12 and 14 from above, we first estimated a linear growth model using a multilevel model, where the outcome was predicted by a variable representing time, coded as Age 10 = 0, Age 12 = 1, Age 14 = 2. The model produced the output displayed in Table 4a.

For simplicity, we present the conditional parameter estimates, which refer to the estimates of model parameters after accounting for covariate effects. The intercept value tells us the predicted level of rIFC thickness (2.62) when time and the other covariates were at zero (i.e., Age 10, and the mean of the covariates), while the random effect (a standard deviation) tells us that we could expect residual variability around that intercept value, with 68% of the population expected to have rIFC thickness values of 2.62 ± 2.21. The slope value tells us that rIFC thickness declined by 5.80 points for every one unit change in time (in this case every two years) at the zero point of the covariates. After accounting for covariate effects, there was variability in change over time, such that 68% of the population exhibited rates of change of 5.80 ± 0.43. Moreover, the intercept and slope were very weakly

Table 4b
Multilevel Growth Model, Full Model, Fixed Effects. n = 250.

	b	S.E.	p-value
Intercept	2.42	0.29	0.000
Linear Slope	−3.13	0.11	0.000
Quadratic Slope	0.05	0.01	0.000
Age 10 Impulse Control	1.02	0.16	0.000
Stress	1.15	0.05	0.000
Linear Slope X Age 10 Impulse Control	−0.21	0.07	0.001
Quadratic Slope X Age 10 Impulse Control	0.07	0.01	0.000

Random Effects

	Random Effects	Random Effect Correlations	
Intercept	2.20	Intercept	Linear Slope
Linear Slope	0.47	−0.48	
Quadratic Slope	0.07	0.50	−0.54
Residual	2.98		

**Note:* This table displays growth model estimates for change in rIFC thickness over time for Ages 10–19, conditional on the effects of the covariates.

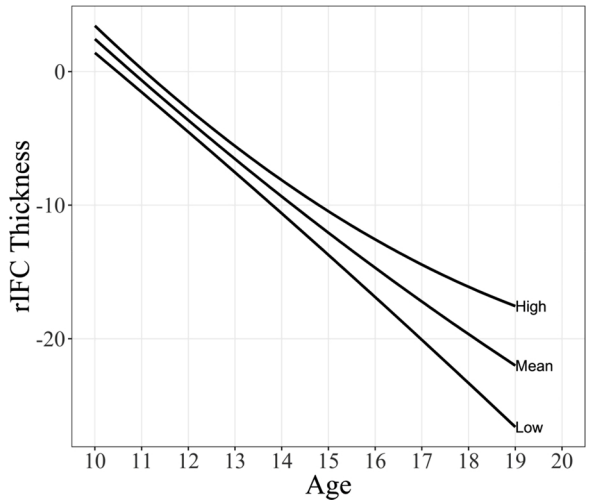

Fig. 5. The Effects of high (+1 SD), mean and low (−1 SD) rIFC thickness on trajectories of impulse control.
Caption: Illustrating variability in change over time of rIFC thickness for High, Mean and Low levels of Impulse Control at Age 10.

correlated (r = −0.14), which means that an individual's level of rIFC thickness at Age 10 was weakly associated with their rate of change over time. The residual term describes unexplained within-person variation in rIFC thickness across all time points.

Covariate effects indicated that greater Age 10 impulse control was related to higher initial levels of rIFC thickness (b = 0.92), but not to change in rIFC thickness over time (b = 0.07). Similarly, stress at each time point was related to higher levels of rIFC thickness at each time point (b = 1.32), meaning a one unit increase in stress at a given age was related to a 1.32 unit increase in rIFC thickness above and beyond that which was predicted by the developmental trajectory. Re-estimating this model in a latent growth curve model framework gave identical results, except the latent variable model would estimate unique residuals at each time point rather than a single residual estimate for all time points.

For comparison, we also re-estimated this model with the full dataset (Table 4b). Because there was now a quadratic effect included in the model, the interpretation of the conditional linear growth coefficient changed: it represented the rate of change *only* when time was equal to 0 (Age 10), controlling for the covariates, while the quadratic effect represented how the linear time effect changes for every one unit change in time (i.e., acceleration or deceleration) controlling for covariate effects. In our simulated model, youth's rIFC thickness at Age 10 was declining by −3.13 per year, but that decline slowed by .05 per year. Accurate estimates of the intercept and slope for different ages can be obtained simply by re-centering the intercept at different time points.

In addition, the covariate effects differed: not only did Age 10 impulse control predict the level of rIFC thickness (to nearly the same degree), but it also predicted the linear and quadratic slope. An increase in Age 10 impulse control was associated with higher initial rIFC thickness (b = 1.02), but also with a greater rate of decline at age 10 (b = −0.21) and a larger quadratic effect (b = 0.07). The predicted trajectories of rIFC thickness at low, average, and high (± one SD) levels of Age 10 impulse control are shown in Fig. 5, which displays (similar to an interaction plot) how trajectories of rIFC thickness would be expected to change for individuals with differing levels of Age 10 impulse control.

Conversely, the time-varying effects of stress (and the residual, which also reflects time-varying, or within-individual variation) were nearly identical to the model above. This is because when there are no between individual differences in within-individual effects, their effects will be generally be unbiased as long as the time interval assessed is identical.

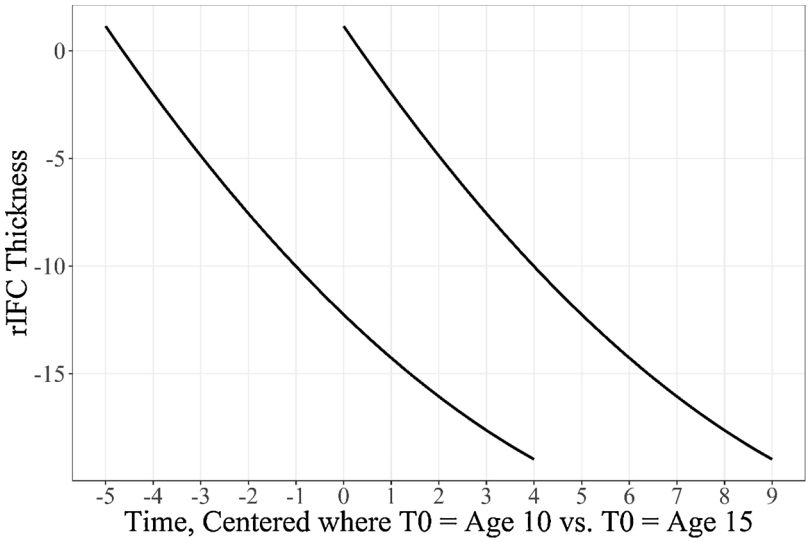

Fig. 6. The effects of centering time at different ages.
Caption: Choosing to center time at age 10, versus centering time at age 15, does not change the average quadratic growth trajectories of rIFC thickness, but only shifts the axis of time.

3.2.2.2. Cautions in interpreting the intercept and the coding of time. As opposed to other models, in growth models the intercept is important because: a) it helps define the trajectory (along with the slope parameters) and b) individual differences (or random effects) in levels of the outcome are of interest. Biesanz et al. (2004) noted that researchers struggled with interpreting the intercept; most commonly, researchers assume the intercept *must* be the first observed time point, though the intercept can be set at any time point. They further noted that many studies wrongly labeled the intercept as the starting point even when it was clear from the coding scheme that this was not true (Biesanz et al., 2004). There are two important consequences to this conflation of the intercept and the starting point. First, it often limits researchers' *inferences* about individual differences in levels, and it limits interpretations of the correlation between the slope and intercept. Both variability in the intercept, and (some degree of) the magnitude of its correlation with the slope, are determined by the placement of the intercept (Biesanz et al., 2004; Mehta and West, 2000; Rogosa and Willett, 1985). Changing the coding of time via centering impacts the estimation of the value of an intercept, its covariance with a linear slope, and the effects of a predictor on an intercept (without affecting estimates of the slope), *all without changing the underlying trajectories or model fit.* Fig. 6 illustrates how choosing to center time at age 10, versus centering time at age 15, does not change the average quadratic growth trajectories of rIFC thickness, but only shifts the axis of time.

Coding time is also important when interpreting non-linear effects of time. Because a quadratic growth effect is estimated as the effect of time squared, the value of the linear effect of time, and its correlation with the intercept are only interpreted *when time is equal to zero*. Just as re-centering data to explore the impact of a moderator changes the simple slope of a predictor on an outcome (Aiken and West, 1991), all other variables in a growth model are impacted by the choice of the intercept. Researchers often incorrectly interpret this in quadratic growth models, and generalize the linear time effect and its correlations to all time points. For example, Harden and Tucker-Drob (2011) estimated correlated change between quadratic growth curves of impulsivity and sensation seeking across adolescence, and reported a moderate (r = .21) correlation between individual differences in linear change in impulsivity and sensation seeking. However, that correlation only applied to age 16 (where the data were centered), because the presence of a quadratic effect meant that the linear effect changed across time, and the correlation was certain to take on other values at different locations. Moreover, if one wanted to most accurately describe how change in one construct was related to another, the correlation between quadratic change observed in the study (r = .41) should be interpreted, because it was not conditional on any other effects in the model. In other words, as the rate of change in impulsivity accelerated by one SD, we might expect a parallel .41 SD acceleration in of change in sensation seeking in Harden and Tucker-Drob (2011).

We illustrate this in Table 5. Changing the location of the intercept changes the estimates of all fixed effects of our quadratic growth model except the quadratic effect itself, as well as the estimates of the correlations among growth factors. Whether a random effect increases or decreases depends on the variability of the trajectories and the location of the point of minimum variability. Fig. 7 illustrates this: around Age 12, there is little variability in levels, so there is likely to be a low

Table 5
The impact of centering at different ages on coefficient estimates. Fixed effects. n = 250.

	Age 10	Age 11	Age 12	Age 13	Age 14	Age 15	Age 16	Age 17	Age 18	Age 19
Intercept	1.15	−1.98	−4.88	−7.56	−10.02	−12.25	−14.27	−16.06	−17.63	−18.98
Linear Slope	−3.24	−3.01	−2.79	−2.57	−2.35	−2.13	−1.90	−1.68	−1.46	−1.24
Quadratic Slope	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11

Correlations among random effects

	Age 10	Age 11	Age 12	Age 13	Age 14	Age 15	Age 16	Age 17	Age 18	Age 19
Intercept: Linear Slope	−0.21	0.18	0.54	0.75	0.82	0.85	0.87	0.88	0.90	0.92
Intercept: Quadratic Slope	0.64	0.56	0.51	0.50	0.53	0.58	0.64	0.69	0.74	0.79
Linear Slope: Quadratic Slope	−0.62	−0.33	0.08	0.46	0.69	0.81	0.88	0.91	0.94	0.95

**Note:* This table displays growth model estimates for change in rIFC thickness over time for Ages 10–19, conditional on the effects of the covariates.

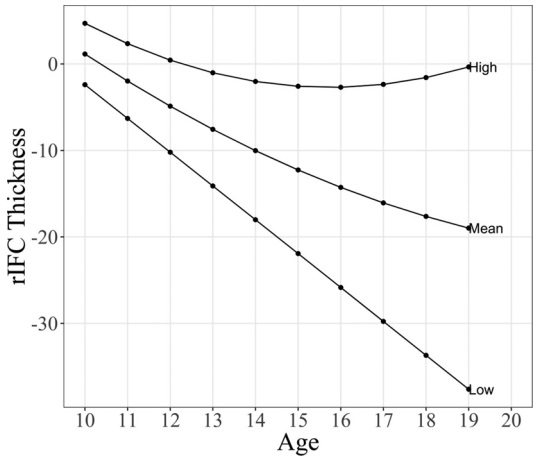

Fig. 7. An illustration of variability in a growth model.
Caption: This figure illustrates average growth over time (Mean), along with model estimated growth for subjects at +1 SD (High) for the intercept, linear and quadratic slopes, and model estimated growth for subjects at −1 SD (Low) for all growth factors.

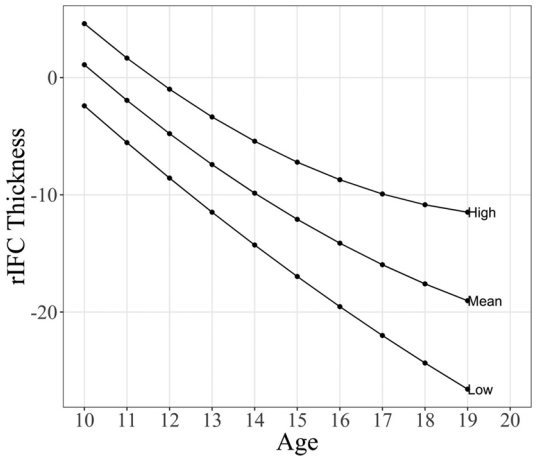

Fig. 8. Illustrating negative correlations (r = −.50) between initial levels and linear and quadratic rates of change.
Caption: Illustrating the association between initial levels and variability in change in rIFC thickness over time, for High, Mean and Low initial levels of rIFC thickness. High initial rIFC thickness is associated with less change over time.

correlation between the level of rIFC thickness and the rate of change over time. Conversely, by Age 19, there is large variability in the level of rIFC thickness, and thus we would expect at Age 19 to observe a large correlation between the level and the rate of change over time.

Unless there is a compelling rationale for choosing a point in time (such as the beginning of high school, or the end of a treatment study), there is value in exploring several locations of the intercept, because the interpretation of slopes, intercepts and their correlations in a growth model are not independent of one another.

3.2.2.3. Cautions in interpreting the slope(s). There is frequent confusion about the interpretation of trajectory parameters in non-linear growth models. As noted above (Ram and Grimm, 2015), growth may be represented by non-linear shapes (e.g., a quadratic slope), and researchers often struggle to interpret even quadratic growth models. Further, assuming sufficient model degrees of freedom, SEMs allow for the free estimation of change parameters (referred to as "latent basis" or "free slope" models), permitting examinations of discontinuous change (Flora, 2008; McCoach and Kaniskan, 2010). When the shape of growth extends beyond the linear case, the interpretation of parameters rapidly becomes less intuitive. Take, for example, the relative difficulty of interpreting the parameter values for Table 5, when the intercept was centered at age 10, with the relative ease of interpreting Figs. 5–9, which represent the same data. Thus, we strongly recommend that all parameters should be interpreted in the context of all others and all models should be graphed.

3.2.2.4. Cautions in interpreting the variances/random effects and covariances. Because of variability in the intercept and slope, it is probably more useful to describe the variability in trajectories than to simply describe the average trajectory. This is where visual displays can be useful (Biesanz et al., 2004). Although it is often recommended, it is rare to find visual displays of variability for growth models, and especially unusual to display variability as a function of the covariation among growth components. It is not uncommon that slopes and intercepts in growth models are correlated. Individual differences in developmental timing can cause children who start at a high level on a particular indicator to demonstrate less pronounced change over time, simply because they enter the study at a more advanced stage of development. This induces a negative correlation between the starting level and the rate of change over time. Researchers regularly fail to interpret these correlations, or (if they do), interpret them incorrectly. Thus, we strongly encourage researchers to plot variability in growth, as well as the correlations between growth factors to better understand the basic properties of their growth models.

Figs. 7–9 provides three examples of how variation and co-variation in our growth example might be illustrated. Fig. 7 illustrates the expected trajectories for individuals at high, mean and low initial levels and linear and quadratic slopes based on the model estimated variability. Fig. 8 illustrates a case where the intercept is negatively correlated with the slope and quadratic effects ($r = -.30$), meaning that individuals who are high on rIFC thickness at Age 10 are expected to change the least over time, while those with the lowest rIFC thickness levels are expected to change the most. Fig. 9 samples 12 cases from the simulated data and displays quadratic curves plotted over individual data points as a means of highlighting between person variability in change over time. These figures provide a more direct, intuitive means of understanding the results provided by the estimated model compared to a text description of variance associated with a growth parameter.

In latent growth curve models, significance tests for variances and covariance of the growth factors are provided as a matter of estimation, as growth factors are estimated as latent variables with means, variances and covariances by default for most software. Thus, it is unlikely that a researcher will explicitly fix the variance of a latent growth factor to zero. However, in multilevel models, variances must be explicitly estimated as a random effect of the intercept and time variables (such as the linear and non-linear time effects). A common limitation that is specific to multilevel models is the failure to test for random effects when the main effect of time (i.e., the fixed effect) is zero. This seems to reflect a misunderstanding that the estimate of within-person change over time applies to all subjects, and thus represents a fundamental misunderstanding of the parameters reported in a GCM. As illustrated in Fig. 2, there can be no average growth, but individual differences in growth (Fig. 2D), and there can also be growth over time, but no individual differences in growth (Fig. 2B). Thus, it should be a matter of course to test for random effects for all trajectory parameters, and they should be fixed to zero only when there is no evidence that they vary in the sample (Raudenbush and Bryk, 2002). Researchers should also use less biased approaches to test parameter significance such as likelihood ratio test/deviance tests (Agresti, 2002; Johnston and Dinardo, 1997), given the more standard Wald-*Z* test may be positively biased when samples are less than 100 (Pawitan, 2001).

By assuming a dependency between the presence of a fixed and a random effect, researchers assume that information about one provides information about another. However, if a random effect for a trajectory parameter is not modeled, variance attributable to that random effect

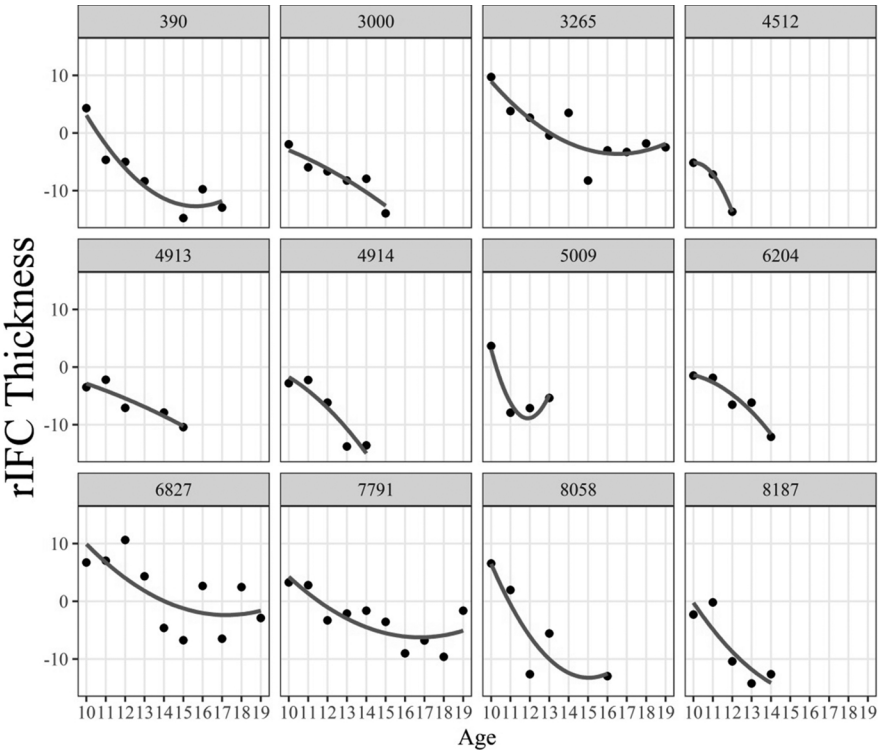

Fig. 9. A random sample of 12 individual growth curves.

will remain in the residuals of the observed variable at each time point, and the standard errors associated with time-varying covariates will in turn be underestimated (Chen et al., 2013; Gibbons et al., 2010; Hedeker et al., 1994). Because of this underestimation, the statistical tests of these parameters will be overly liberal, resulting in inappropriately small *p*-values and inflated Type I Error. Note that the fixed effect parameters are often not substantially altered when random effects are specified (e.g., Hedeker et al., 1994); this is often because the fixed effect estimate (for example, the effect of linear change over time) is simply the sample average, while the random effect simply provides an estimate of how individuals in the sample vary around that average. If researchers observe a non-significant fixed effect (i.e., slope of time), and assume that means no one in their sample is changing over time, this could be a serious mis-characterization of their data.

Some researchers have argued that a stepwise approach, requiring a significant random coefficient before testing the effects of predictors on that coefficient (Raudenbush and Bryk, 2002). Despite intuitive appeal, we argue that this requirement is unnecessarily restrictive (Aguinis et al., 2013). In short, if the covariate effect is true, a stepwise method requires researchers to have a significant random effect for a model that is effectively mis-specified, and it is not clear that the variance attributable to the covariate effect would be accurately represented in the unconditional random effect. (Aguinis, 1995). Thus, we believe it is reasonable to test for theoretically-grounded moderation in growth models, regardless of the presence of a significant random effect.

3.2.2.5. Cautions in interpreting predictor effects. Associations with either the initial levels of rIFC thickness or explained time-varying variance in rIFC thickness can be relatively straightforward to interpret. It is often more difficult to interpret predictor effects when they explain individual differences in a slope, and particularly so for non-linear growth models, as covariate effects on slopes represent interactions between the covariate and the effect of time on the outcome. However, this interpretation does not come readily to most applied researchers, and is further complicated when trajectories include non-linear components. Given that *all* coefficients that identify a trajectory must be interpreted to properly interpret covariate effects, we strongly urge that researchers plot covariate effects on trajectories, in same way they plot other interactions. For example, if a covariate is associated with all aspects of a non-linear trajectory (such as the intercept, linear and quadratic slope), all three effects should be plotted in a single graph (see Fig. 5 for an example of this). Take the associations of Age 10 impulse control with rIFC thickness. Greater impulse control at Age 10 was associated with higher rIFC thickness at Age 10. However, it was also associated with rIFC thickness that declined more quickly and slowed in deceleration over time, exhibiting the highest levels of rIFC thickness by Age 19. Interpreting the effects of a predictor on trajectory parameter (such as the linear slope) without considering its effects on other trajectory parameters can lead to misleading inferences. In the supplementary materials, we have provided R code for all of our figures, as well as an illustration of how several of our figures could be created in spreadsheet software using only model output.

3.2.3. Latent class growth and growth mixture models

Latent class growth models (LCGM; Nagin, 1999) extend LCMs by assuming that multiple discrete populations produced the sample data. That is, individual differences in levels and change are thought to arise from multiple, discrete distributions with different means rather than reflect normal variability around a single distribution. Importantly, estimation of LCGM required within-group variances for intercept and slope to be set to zero. LCGMs assign all individuals a probability of belonging to each latent class, and more optimal solutions exhibit high entropy (i.e., latent classes where individuals have a high probability of belonging to one class and a low probability of belonging to other classes). Latent growth mixture models (LGMMs) extend the Nagin model by allowing the latent classes to have variability around the mean intercept and slopes (Ram and Grimm, 2009). In other words, they can model heterogeneity in trajectories both between classes, as well as within them (Bauer and Shanahan, 2007), although often default settings are to assume equal variances across classes. Bauer and colleagues provide an excellent overview of the similarities and differences between multilevel, latent growth, and latent class growth models (Bauer et al., 2007). In brief, LGMMs should generally be preferred to LCGMs because they model within-class variability in growth.

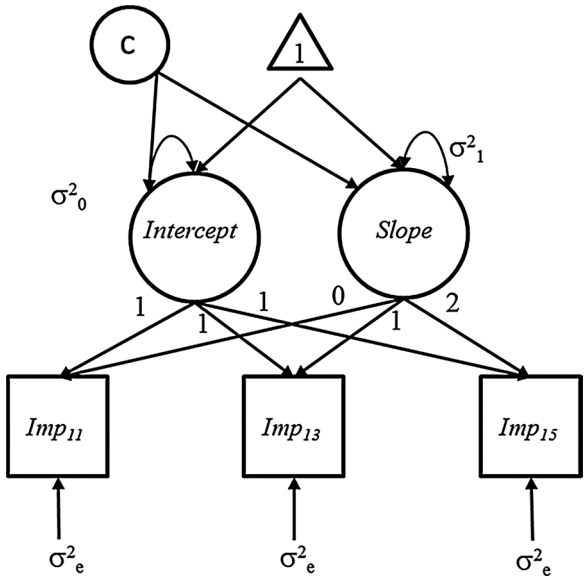

Fig. 10. The growth mixture model.
Caption: As compared to the latent growth model (Fig. 4), note that its only difference is a new latent variable "c", which represents the probability of membership in each latent class.

However, LGMMs may be more difficult to estimate because of the additional parameters required to estimate within class variability. Although these models are exploratory, in that tests of model fit are relative rather than absolute (Jung and Wickrama, 2008), they can be a useful way of testing theories about heterogeneity in development. For example, many early examples of LCGM and LGMMs were motivated by theories about "adolescent limited" versus "persistent" developmental trajectories of criminality and delinquency (Moffitt, 1993; Nagin and Tremblay, 1999) and substance use (Chassin et al., 2002). Fig. 10 provides an SEM illustration of the growth mixture model.

3.2.3.1. Example: LCGM. For the current dataset, we took a random subsample of 500 cases (to improve model estimation larger samples are often required for latent class analysis), and analyzed them as an LCGM. We used the recommended rules to choose the number of latent classes based on a range of model fit indices (Jung and Wickrama, 2008; Ram and Grimm, 2009). We tested one to four classes. Based on recommendations from the literature (Nylund et al., 2007), the final models suggested that three classes best fit the data, because this solution minimized AIC and BIC, the Vuong-Lo-Mendell-Rubin likelihood ratio test (which performs a corrected likelihood ratio test comparing nested models) of the 2 versus 3 class model was significant, entropy, meaning the average classification probability for participants' most likely class, was high, and models with 4 or more classes divided existing classes into non-meaningful sub-groups, and/or became unstable in estimation (see supplementary tables S1). Fig. 11, which illustrates the model estimated trajectories from the LCGM solution, is strikingly similar to Fig. 7, and illustrates how the LCGM parceled variation in growth into separate classes. Effectively, it described three patterns of change in rIFC thickness across adolescence: one of declines which slow over time, one of larger declines that slow less, and one of even greater declines that do not slow over time. However, because the within-class variability in LCGM was forced to be zero, this model actually estimated *less* variability in growth than a latent growth model or growth mixture model would, because both of the latter models allowed for between person variability around trajectory averages (as illustrated in Fig. 1).

3.2.3.2. Example: LGMM. Our simulated data did not provide an optimal illustration of the advantages of LGMM; thus for LGMM we simulated four separate samples of 100 subjects each with different linear trajectory shapes across 5 time points, merged them into a combined dataset, and analyzed them with LGMM to see if we could recover the individual groups (illustrated in Fig. 12a and b). Here, we simulated data to represent the "cat's cradle" pattern of stable-high, stable-low, and increasing/decreasing classes of change over time, which has been found repeatedly in growth mixture modeling studies of risky behavior. In the "cat's cradle", named after a children's game, there are four classes: one high and stable, one low and stable, one increasing and one decreasing (Sher et al., 2011). Although this may be biologically unlikely, it is a useful example because it parallels many common behavioral examples in the literature applying LGMM.

Again, we tested for between one and four latent classes. Model fit indices (supplementary table S2) suggested that the four class solution best fit the data, because AIC and BIC were low, entropy was high, and the Vuong test suggested that the 4 class solution was superior to 3

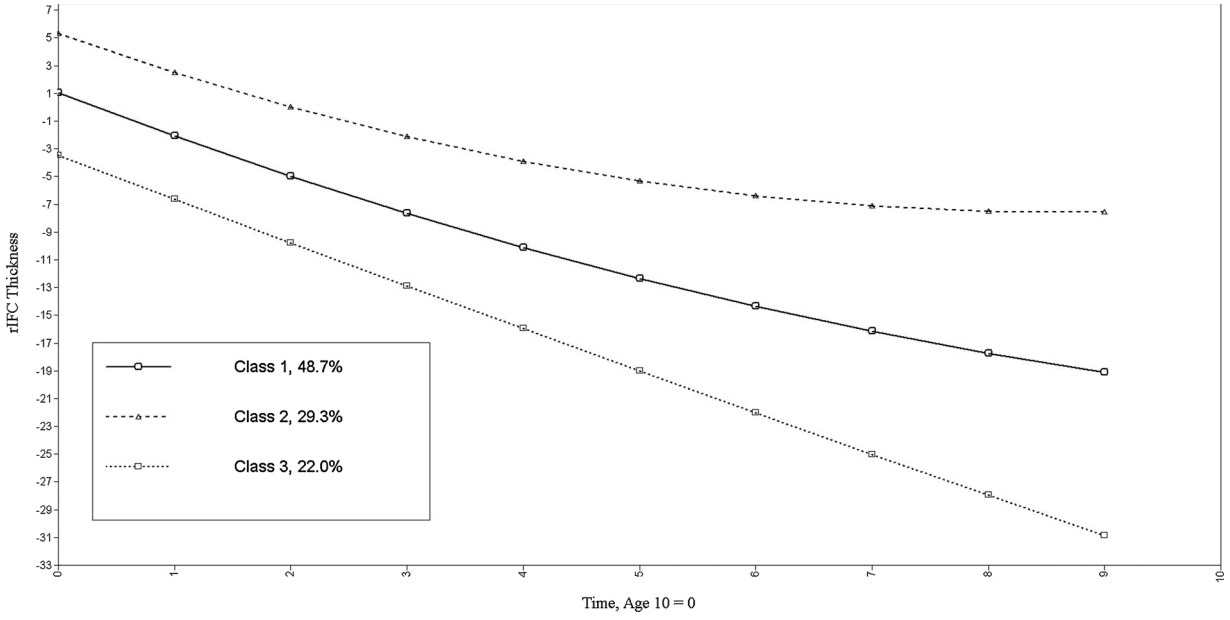

Fig. 11. A three class latent class growth model solution (n = 500).

classes but that 5 classes provided no additional explanatory power. Around 100 participants classified in each of the classes based on their highest probability of class membership. In short, the LGMM procedure, in this case, did a reasonable job replicating the actual sub-populations of the simulated data. We could go further and predict class membership from covariates to understand how individual differences in covariates are related to the probability of belonging to one class or another (Nylund et al., 2007) to further understand the correlates of inter-individual variability in trajectories. They would be interpreted using a multinomial logistic regression framework, with covariates predicting relative probability of membership in different classes (see Asparouhov and Muthen, 2014 for more details).

3.2.3.3. Cautions about over-extraction and reification of latent classes. Researchers should be cautious in the application and interpretation of such LCGMs (Sher et al., 2011). Research has shown that latent class growth models are as vulnerable to differences in model specification. For example, different trajectory solutions are obtained across different scaling of outcome variables (e.g., continious vs. ordinal; Jackson and Sher, 2008), the number and interval of observations (Jackson and Sher, 2006), and the choice of (highly related) outcomes (e.g., heavy drinking vs. alcohol quantity-frequency; Jackson and Sher, 2005). Sher and colleagues also demonstrated that certain types of solutions (i.e., the "cat's cradle") appear to be prototypic in longitudinal data of a given phenomena (e.g., alcohol use), regardless of the developmental period under consideration (see Sher et al., 2011). These findings indicate that methodological artifacts drove at least some of the solutions obtained by these methods. Further, non-normal data may produce trajectory classes when none exist in the true population (Bauer, 2007; Bauer and Curran, 2003). Finally, researchers should take care to recall that all latent classes are probabilistic, and should not be confused with true observed group differences.

In sum, LCGM and LGMMs can be very useful models for testing hypotheses about variation in growth, but caution should be taken in their interpretation, and the latent classes should not be reified. Latent growth classes do not "carve nature at its joints," but may provide a useful means of describing different probabilistic patterns of variability in change over time (Nagin and Tremblay, 2005). By identifying these latent sub-populations who exhibit different trajectories of change over time, LCGM and LGMMs provide a way to explore how subgroups of individuals may differentially respond to interventions, or exhibit differential susceptibility to risk and protective factors. Best practices for testing and fitting LCGM and LGMM begin with following best practices for GCMs, and following recommended practices for model selection (Nylund et al., 2007), as well as being cautious to avoid reifying the classes that are discovered (Sher et al., 2011).

3.2.4. Latent change score models

Latent change score models (LCS; Ferrer and McArdle, 2010; McArdle, 2009) are SEM-based extensions of the auto-regressive panel model that may be used to test auto-regressive, cross-lag models and growth curve models (described in more detail below), while allowing somewhat more dynamic questions about change to be posed about each type of model. LCS models use highly constrained latent variable models to separate stability and change at each time point. Although they can be used to essentially replicate other models described above, their specification allows many flexible representations of change processes that may not be adequately captured by those models. Excellent overviews of LCS models, and their comparison to auto-regressive and latent growth models, are presented by several authors (Ferrer and McArdle, 2003; Grimm et al., 2016).

3.2.4.1. Example. One advantage of LCS models is that they may be used with two time points of data to create a latent change variable that can itself be used to predict latent time points, which is not possible with either regression or auto-regressive and cross-lag panel models. The growth model may be overlaid on the LCS model by loading an intercept factor on the latent factor at the desired time point, and the slope factor(s) on the latent change variables. Fig. 13 provides an SEM

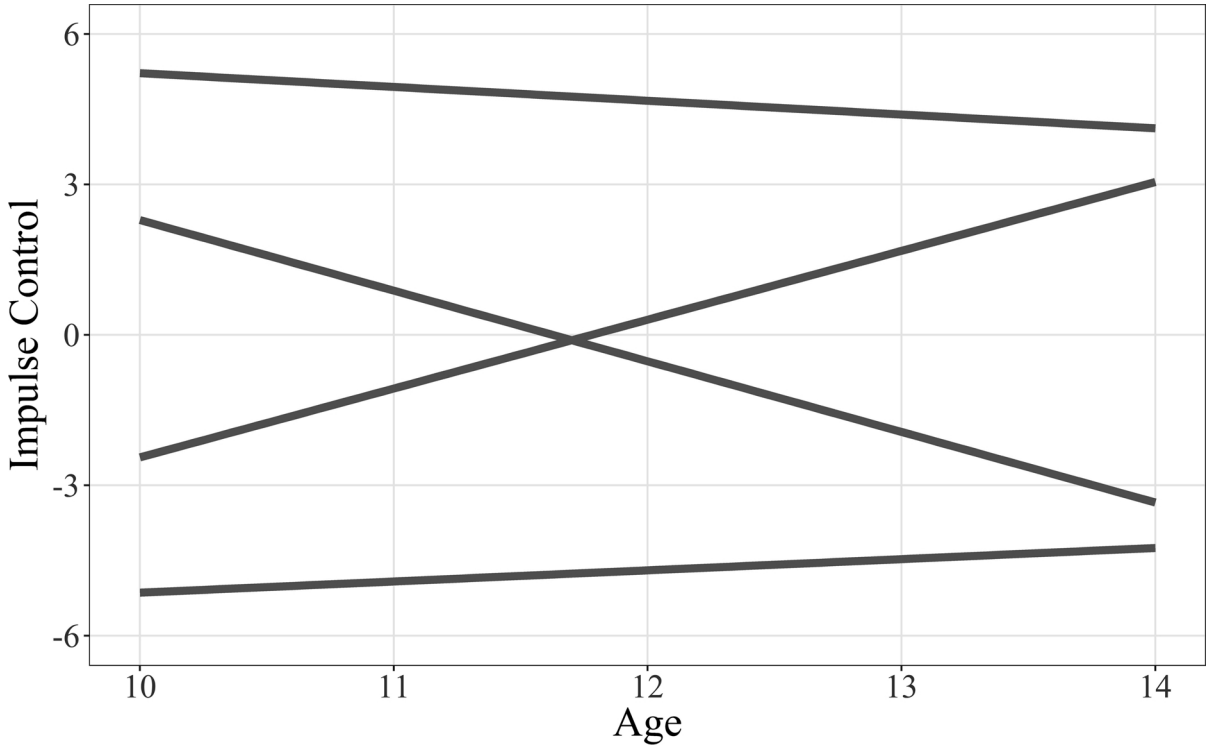

Fig. 12. (a) Four simulated latent growth trajectory groups (n = 100 each) across four time points. (b) Latent growth mixture model "best fitting" 4 class solution of the same data.

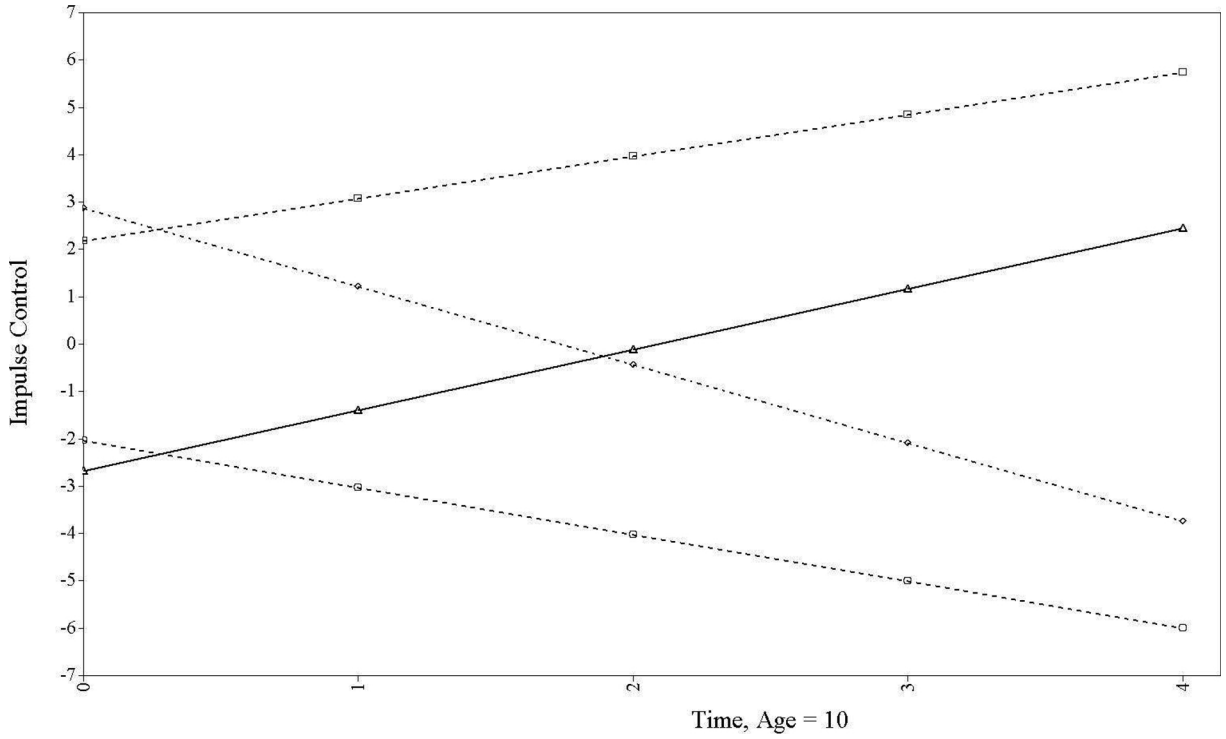

Fig. 12. (*continued*)

representation of a latent change score model from our simulated dataset for ages 11, 13 and 15, and Table 6 providing coefficient estimates.

These results indicate that individuals had an average level of rIFC thickness at age 11 of −1.73, and a variance of 14.37 (SD = 3.79). Above and beyond year-to-year stability, individuals on average declined by −4.53 from age 11 to 13, and varied around that degree of change (SD = 2.81), and declined faster, by −5.90 from age 13 to 15, while also varying in that amount of change (SD = 3.82). Note how these inferences differ from those obtained from the autoregressive model described above in Tables 2 and 3, although they were derived from the same data and are somewhat similar models. We could include predictors or outcomes of those change scores, as well, or use latent variables to summarize between individual differences in change over time and make the LCS akin to a growth model (e.g., see Fig. 6 in McArdle, 2009). This is the powerful flexibility of the LCS framework, see (Kievit et al., 2017) for a more in depth discussion of LCS models.

Table 6
Latent change score results

	Means	Variances	Residual
Y11	−1.73	14.37	6.21
δ 13	−4.53	7.91	
δ 15	−5.90	14.65	

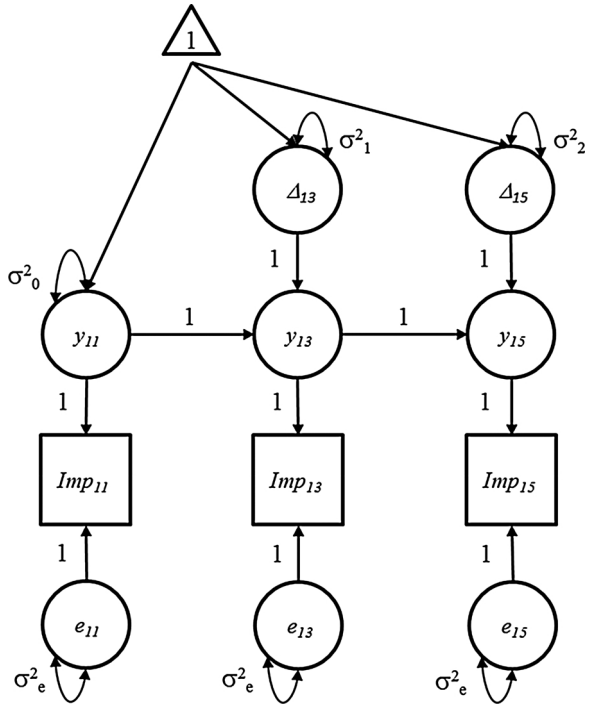

Fig. 13. Estimates for a latent change score in impulse control across ages 11, 13 and 15.

These models may be readily extended to the bivariate model, testing models of reciprocal change (much like CLPMs), or to the bivariate growth modeling case, testing theories about correlated change. There has been substantially less methodological research on LCS models (Ferrer and McArdle, 2010; McArdle, 2009; though see Usami et al., 2015, for a recent examination of the mathematical relation between latent change score and autoregressive cross-lagged factor approaches) and LCS models are only just beginning to be widely applied (relative to LCMs). In our experience, these already highly constrained models tend to require additional constraints to permit estimation, such as constraining estimates of change to be equal over time, and can sometimes be difficult to fit (in our experience). However, LCS models represent a powerful and flexible class of models for change over time that are worthy of additional research attention.

4. General cautions, conclusions and limitations

There are general concerns about fitting longitudinal data to a given model that apply to all longitudinal models. Most applications we have discussed either rely on model fit criteria that may be applied to

sequences of models to allow researchers to make decisions about which parameterization of a model best fits the data. For example, researchers may wish to test whether a quadratic time effect or a free time slope better explains change in rIFC thickness, or whether a three versus four latent class solution explains individual differences in trajectories. Across all maximum likelihood based estimation approaches, relative model fit indices derived from the likelihood function [such as Akaike's Information Criteria (AIC), Bayesian Information Criteria (BIC, and −2 log-likelihood (-2LL)] may be used to guide modeling decisions. SEMs also provide a range of fit indices [e.g., Chi-square, root-mean squared error of approximation (RMSEA), and confirmatory fit index (CFI)] as a means of informing how well model parameters reproduce the observed covariance matrix among the variables (Bollen, 1989).

However, researchers should be cautious about applying simplistic "golden rules" about model fit (Bentler, 2007; Hu and Bentler, 1999; Marsh et al., 2004). Substantial research suggested that using strict cutoffs does not perform well under some circumstances (Marsh et al., 2004). As such, suggestions for what determines a well-fitting model have evolved over time, and currently a broad and holistic approach to model fitting is encouraged (Jackson et al., 2009; see special issue of *Personality and Individual Differences*, May, 2007) SEM.

It is also commonly assumed that one well-fitting model rules out alternative models (Tomarken and Waller, 2003), when in fact model fit indices for SEM *only* provide information on whether the model tested could have plausibly generated the data at hand. Thus, careful consideration of multiple plausible alternative models (e.g., Liu et al., 2012)(e.g. Liu et al., 2012) in the context of the design and analysis of longitudinal data is critical to drawing appropriate inferences. As we have demonstrated, different models of change may be fit to the same data, each answering different questions about change over time. For example, auto-regressive models can test hypotheses about the timing of associations between variables, such as whether one variable predicts another, or vice-versa, or whether an association strengthens or weakens over time. Conversely, growth curve models can test hypotheses about how two variables might change together over time, as in parallel process growth models. In choosing a statistical model to test one's hypothesis, it is important to consider how alternative models might provide different insights about change, and to acknowledge that any given model is only one way of testing theories of change.

Longitudinal data have the ability to establish temporal precedence, thus improving causal inference over cross-sectional studies because observations of an outcome at a prior time point may be controlled. However, longitudinal data are not a panacea for causal inference. As noted by Littlefield and Alquist (2014), non-experimental observational data remain correlational, even if data are longitudinal. That is, models that assume specific sets of causal processes may fit the data as well as models that assume alternative causal relations or no causal processes (Rovine and Molenaar, 2005; Tomarken and Waller, 2003). Thus, though longitudinal observational data may illuminate various patterns of co-relations among variables across time and provide clues regarding functional relations between constructs, strong evidence of causality will require integrating these data with evidence from other approaches (Littlefield and Alquist, 2014). For example, though many authors are aware that inferences about mediation are inappropriate in cross-sectional data, similar issues still arise in longitudinal data (see Cole and Maxwell, 2003 for a detailed guide on mediation in longitudinal data). As with cross-sectional data, longitudinal data can suffer from unmeasured third variables that confound causality (Fraley and Roberts, 2005). Indeed, threats to inferences of mediation are fundamental even in the best experimental contexts (Bullock et al., 2010).

Although examining parallel change is a critical direction for future neuroimaging research, given the nascent state of longitudinal modeling in developmental neuroimaging, we did not go into detail on parallel processes growth models (Cheong et al., 2003), autoregressive latent trajectory (ALT) models (Bollen and Curran, 2004), or latent curves with structured residuals (Curran et al., 2014). Similarly, it is important for researchers to consider issues of variable centering and separating between- and within-person variability in both predictors and outcomes in longitudinal models (Curran and Bauer, 2011; Enders and Tofighi, 2007). State-trait models (Kenny and Zautra, 2001) may be used to test the degree to which between- and within-person variation in covariates contribute to the development of psychopathology over time (e.g., King et al., 2009; McLaughlin and King, 2014). Latent transition models (Bray et al., 2010) may be used to test how individuals transition between different expressions of disorder (Jackson et al., 2006). As these methods become increasingly available to neuroimaging researchers, we encourage their exploration as they fit researchers' hypotheses and available software. Table 1 provides references detailing the use of a number of these models.

The field of neuroscience is at an exciting juncture as an increasing number of research endeavors are incorporating repeated-measures designs to inform critical scientific questions germane to neuroimaging. At the time of this writing, two major efforts are underway. The Adolescent Brain Cognitive Development (ABCD, https://abcdstudy.org/) study aims to collect data on 10,000 participants across the United States who will be tracked over the course of ten years starting at age 9. Individuals who take part of this study will be scanned every other year, potentially resulting in 5 time points of brain data for each individual between the ages of 9 and 19 years. Another developmental major longitudinal neuroimaging initiative, The Lifebrain study (https://www.lifebrain.uio.no/), aims to examine 6000 Europeans across different periods in the human lifespan from ages 0–100 years, with an expected 40,000 time point total (with potentially ~6 time points per individual). The success of these (and similar) projects will depend, in part, on the appropriate utilization of statistical techniques to analyze longitudinal data.

Towards this end, the current paper summarized common problems and issues that arise when applying longitudinal models in the extant developmental literature. Reflecting, to some extent, an analytic embarrassment of riches, a legion of techniques currently exists to examine a variety of types of change as well as explore various functional relations among constructs across time. As shown in our paper, each approach includes a somewhat distinct set of strengths and weaknesses, necessitating a series of (hopefully informed) decisions throughout the research process. It is our intent that the provided suggestions and solutions will serve as a useful guide to those who seek to optimize the design, analysis, and interpretation of longitudinal neuroimaging studies.

Conflict of Interest

None

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.dcn.2017.11.009.

References

Achenbach T.M., Edelbrook C., (1983). Manual for the Child Behavior Checklist: And Revised Child Behavior Profile. Burlington, VT : University of Vermont.
Agresti, A., 2002. An introduction to categorical data analysis. Wiley Ser. Probab. Stat. 45. http://dx.doi.org/10.1198/tech.2003.s28.
Aguinis, H., Gottfredson, R.K., Culpepper, S.A., 2013. Best-Practice recommendations for estimating cross-Level interaction effects using multilevel modeling. J. Manage. 39 (6), 1490–1528. http://dx.doi.org/10.1177/0149206313478188.
Aguinis, H., 1995. Statistical power with moderated multiple regression in management research. J. Manage. 21 (6), 1141–1158. http://dx.doi.org/10.1177/014920639502100607.
Aiken, L.S., West, S.G., 1991. Multiple Regression: Testing and Interpreting Interactions. Sage, Los Angeles, CA.
Aron, A.R., Robbins, T.W., Poldrack, R.A., 2014. Inhibition and the right inferior frontal cortex: one decade on. Trends Cogn. Sci. 18 (4), 177–185.
Asparouhov, T., Muthen, B., 2014. Auxiliary variables in mixture modeling: 3-Step

approaches using mplus. Struct. Equ. Model. Multidiscip. J. 21 (3), 329–341. http://dx.doi.org/10.1080/10705511.2014.915181.
Bailey, D.H., Littlefield, A.K., 2017. Does reading cause later intelligence? Accounting for stability in models of change. Child Dev. 88, 1913–1921.
Ballinger, G.A., 2004. Using generalized estimating equations for longitudinal data analysis. Organiz. Res. Methods 7 (2), 127–150. http://dx.doi.org/10.1177/1094428104263672.
Bauer, D.J., Curran, P.J., 2003. Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. Psychol. Methods 8 (3), 338–363. http://dx.doi.org/10.1037/1082-989X.8.3.338.
Bauer, D.J., Shanahan, M.J., 2007. Modeling complex interactions: person-Centered and variable-Centered approaches. Modeling Contextual Effects in Longitudinal Studies. Taylor & Francis, New York, NY. http://dx.doi.org/10.4324/9780203936825.
Bauer, D.J., Luz, H., Reyes, M., 2007. Modeling variability in individual development: differences of degree or kind? Child Dev. Perspect. 4 (2), 114–122.
Bauer, D.J., 2007. 2004 Cattell award address observations on the use of growth mixture models in psychological research. Multivariate Behav. Res. 42 (4), 757–786.
Bentler, P.M., 2007. On tests and indices for evaluating structural models. Personal. Individual Diff. 42 (5), 825–829. http://dx.doi.org/10.1016/j.paid.2006.09.024.
Biesanz, J.C., Deeb-Sossa, N., Papadakis, A.A., Bollen, K.A., Curran, P.J., 2004. The role of coding time in estimating and interpreting growth curve models. Psychol. Methods 9 (1), 30–52. http://dx.doi.org/10.1037/1082-989X.9.1.30.
Bollen, K.A., Curran, P.J., 2004. Autoregressive latent trajectory (ALT) models a synthesis of two traditions. Sociol. Methods Res. 32 (3), 336–383.
Bollen, K.A., Curran, P., 2006. Latent Curve Models: A Structural Equation Perspective. John Wiley & Sons, Hoboken, New Jersey.
Bollen, K.A., 1989. Structural equations with latent variables. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics Section. John Wiley & Sons, New York, NY.
Braams, B.R., van Duijvenvoorde, A.C.K., Peper, J.S., Crone, E.A., 2015. Longitudinal changes in adolescent risk-taking: a comprehensive study of neural responses to rewards, pubertal development, and risk-taking behavior. J. Neurosci. 35 (18), 7226–7238. http://dx.doi.org/10.1523/JNEUROSCI.4764-14.2015.
Bray, B.C., Lanza, S.T., Collins, L.M., 2010. Modeling relations among discrete developmental processes: a general approach to associative latent transition analysis. Struct. Equ. Model. 17 (4), 541–569. http://dx.doi.org/10.1080/10705511.2010.510043.
Bronfenbrenner, U., 1977. Toward an experimental ecology of human development. Am. Psychol. 32 (7), 513. http://dx.doi.org/10.1007/s10648-006-9029-9.
Brooks-Gunn, J., Phelps, E., Elder, G.H., 1991. Studying lives through time: secondary data analyses in developmental psychology. Dev. Psychol. 27 (6), 899–910. http://dx.doi.org/10.1037//0012-1649.27.6.899.
Bryk, A.S., Raudenbush, S.W., 1987. Application of hierarchical linear models to assessing change. Psychol. Bull. 101 (1), 147–158. http://dx.doi.org/10.1037/0033-2909.101.1.147.
Bullock, J.G., Green, D.P., Ha, S.E., 2010. Yes, but what's the mechanism? (don't expect an easy answer). J. Pers. Soc. Psychol. 98 (4), 550–558. http://dx.doi.org/10.1037/a0018933.
Burt, K.B., Obradović, J., 2013. The construct of psychophysiological reactivity: statistical and psychometric issues. Dev. Rev. 33 (1), 29–57. http://dx.doi.org/10.1016/j.dr.2012.10.002.
Casey, B.J., Caudle, K., 2013. The teenage brain: self control. Curr. Direct. Psychol. Sci. 22 (2), 82–87. http://dx.doi.org/10.1177/0963721413480170.
Chassin, L., Pitts, S.C., Prost, J., 2002. Binge drinking trajectories from adolescence to emerging adulthood in a high-risk sample: predictors and substance abuse outcomes. J. Consult. Clin. Psychol. 70 (1), 67–78. http://dx.doi.org/10.1037//0022-006X.70.1.67.
Chen, G., Saad, Z.S., Britton, J.C., Pine, D.S., Cox, R.W., 2013. Linear mixed-effects modeling approach to FMRI group analysis. Neuroimage 73, 176–190. http://dx.doi.org/10.1016/j.neuroimage.2013.01.047.
Cheong, J., Mackinnon, D.P., Khoo, S.T., 2003. Investigation of mediational processes using parallel process latent growth curve modeling. Struct. Equ. Model. Multidiscip. J. 10 (2), 238. http://dx.doi.org/10.1207/S15328007SEM1002_5.
Cole, D.A., Maxwell, S.E., 2003. Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling. J. Abnorm. Psychol. 112 (4), 558–577. http://dx.doi.org/10.1037/0021-843X.112.4.558.
Cole, D.A., Martin, N.C., Steiger, J.H., 2005. Empirical and conceptual problems with longitudinal trait-state models: introducing a trait-state-occasion model. Psychol. Methods 10 (1), 3.
Collins, L.M., Graham, J.W., 2002. The effect of the timing and spacing of observations in longitudinal studies of tobacco and other drug use: temporal design considerations. Drug Alcohol Depend. 68, S85–S96. http://dx.doi.org/10.1016/S0376-8716(02)00217-X.
Collins, L.M., 2006. Analysis of longitudinal data: the integration of theoretical model, temporal design, and statistical model. Annu. Rev. Psychol. 57, 505–528. http://dx.doi.org/10.1146/annurev.psych.57.102904.190146.
Crone, E.A., Elzinga, B.M., 2015. Changing brains: how longitudinal functional magnetic resonance imaging studies can inform us about cognitive and social-affective growth trajectories. Wiley Interdiscip. Rev. Cogn. Sci. 6 (1), 53–63. http://dx.doi.org/10.1002/wcs.1327.
Curran, P.J., Bauer, D.J., 2011. The disaggregation of within-person and between-person effects in longitudinal models of change. Annu. Rev. Psychol. 62, 583–619. http://dx.doi.org/10.1146/annurev.psych.093008.100356.
Curran, P.J., Hussong, A.M., 2003. The use of latent trajectory models in psychopathology research. J. Abnorm. Psychol. 112 (4), 526.
Curran, P.J., Howard, A.L., Bainter, S. a., Lane, S.T., McGinley, J.S., 2014. The separation of between-person and within-person components of individual change over time: a latent curve model with structured residuals. J. Consult. Clin. Psychol. 82 (5), 879–894. http://dx.doi.org/10.1037/a0035297.
Curran, P.J., 2003. Have multilevel models been structural equation models all along/. Multivariate Behav. Res. 38 (4), 529–569. http://dx.doi.org/10.1207/s15327906mbr3804_5.
Davis-Kean, P.E., Jager, J., Maslowsky, J., 2015. Answering developmental questions using secondary data. Child Dev. Perspect. 9 (4), 256–261. http://dx.doi.org/10.1111/cdep.12151.
Dishion, T.J., Patterson, G.R., Kavanach, K.A., 1992. An experimental test of the coercion model: linking theory measurement, and intervention. Preventing Antisocial Behavior: Interventions from Birth Through Adolescence. Guilford Press, pp. 253–282.
Enders, C.K., Tofighi, D., 2007. Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. Psychol. Methods 12 (2), 121–138. http://dx.doi.org/10.1037/1082-989X.12.2.121.
Ferrer, E., McArdle, J., 2003. Alternative structural models for multivariate longitudinal data analysis. Struct. Equ. Model. Multidiscip. J. 10 (4), 493–524. http://dx.doi.org/10.1207/S15328007SEM1004_1.
Ferrer, E., McArdle, J.J., 2010. Longitudinal modeling of developmental changes in psychological research. Curr. Direct. Psychol. Sci. 19 (3), 149–154. http://dx.doi.org/10.1177/0963721410370300.
Flora, D.B., 2008. Specifying piecewise latent trajectory models for longitudinal data. Struct. Equ. Model. Multidiscip. J. 15 (3), 513–533. http://dx.doi.org/10.1080/10705510802154349.
Fraley, R.C., Roberts, B.W., 2005. Patterns of continuity: a dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. Psychol. Rev. 112 (1), 60–74. http://dx.doi.org/10.1037/0033-295X.112.1.60.
Gibbons, R.D., Hedeker, D., DuToit, S., 2010. Advances in analysis of longitudinal data. Annu. Rev. Clin. Psychol. 6 (March), 79–107. http://dx.doi.org/10.1146/annurev.clinpsy.032408.153550.
Glenn, N.D., 1976. Cohort analysts' futile quest: statistical attempts to separate age, period and cohort effects. Am. Sociol. Rev. 41 (5), 900. http://dx.doi.org/10.2307/2094738.
Gollob, H.F., Reichardt, C.S., 1987. Taking account of time lags in causal models. Child Dev. 58 (1), 80. http://dx.doi.org/10.2307/1130293.
Granic, I., Patterson, G.R., 2006. Toward a comprehensive model of antisocial development: a dynamic systems approach. Psychol. Rev. 113 (1), 101–131. http://dx.doi.org/10.1037/0033-295X.113.1.101.
Greenhoot, A.F., Dowsett, C., 2012. Secondary data analysis: an important tool for addressing developmental questions. J. Cogn. Dev. 13 (1), 2–18. http://dx.doi.org/10.1080/15248372.2012.646613.
Grimm, K.J., Kuhl, A.P., Zhang, Z., 2013. Measurement models, estimation, and the study of change. Struct. Equ. Model. Multidiscip. J. 20 (February), 504–517. http://dx.doi.org/10.1080/10705511.2013.797837.
Grimm, K.J., Mazza, G.L., Mazzocco, M.M.M., 2016. Advances in methods for assessing longitudinal change. Educ. Psychol. 51 (3–4), 342–353. http://dx.doi.org/10.1080/00461520.2016.1208569.
Grimm, K.J., 2012. Intercept centering and time coding in latent difference score models. Struct. Equ. Model. Multidiscip. J. 19 (1), 137–151. http://dx.doi.org/10.1080/10705511.2012.634734.
Hamaker, E.L., Kuiper, R.M., Grasman, R.P.P.P., 2015. A critique of the cross-Lagged panel model. Psychol. Methods 20 (1), 102–116. http://dx.doi.org/10.1037/a0038889.
Hamilton, K.R., Littlefield, A.K., Anastasio, N.C., Cunningham, K.A., Fink, L.H.L., Wing, V.C., et al., 2015. Rapid-response impulsivity: definitions, measurement issues, and clinical implications. Personal. Disord. 6 (2), 168–181. http://dx.doi.org/10.1037/per0000100.
Harden, K.P., Tucker-Drob, E.M., 2011. Individual differences in the development of sensation seeking and impulsivity during adolescence: further evidence for a dual systems model. Dev. Psychol. 47 (3), 739–746. http://dx.doi.org/10.1037/a0023279.
Hedeker, D., Gibbons, R.D., Flay, B.R., 1994. Random-effects regression models for clustered data with an example from smoking prevention research. J. Consult. Clin. Psychol. 62 (4), 757–765.
Hedge, C., Powell, G., Sumner, P., 2017. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Methods. http://dx.doi.org/10.3758/s13428-017-0935-1. https://link.springer.com/article/10.3758/s13428-017-0935-1#citeas.
Hertzog, C., Nesselroade, J.R., 1987. Beyond autoregressive models: some implications of the trait-state distinction for the structural modeling of developmental change. Child Dev. 58 (1), 93–109. http://dx.doi.org/10.2307/1130294.
Hu, L., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct. Equ. Model. Multidiscip. J. 6 (1), 1–55. http://dx.doi.org/10.1080/10705519909540118.
Jackson, K.M., Sher, K.J., 2005. Similarities and differences of longitudinal phenotypes across alternate indices of alcohol involvement: a methodologic comparison of trajectory approaches. Psychol. Addict. Behav. 19 (4), 339–351. http://dx.doi.org/10.1037/0893-164X.19.4.339.
Jackson, K.M., Sher, K.J., 2006. Comparison of longitudinal phenotypes based on number and timing of assessments: a systematic comparison of trajectory approaches II. Psychol. Addict. Behav. 20 (4), 373–384. http://dx.doi.org/10.1037/0893-164X.20.4.373.
Jackson, K.M., Sher, K.J., 2008. Comparison of longitudinal phenotypes based on alternate heavy drinking cut scores: a systematic comparison of trajectory approaches III. Psychol. Addict. Behav. 22 (2), 198–209. http://dx.doi.org/10.1037/0893-164X.22.

2.198.
Jackson, K.M., O'Neill, S.E., Sher, K.J., 2006. Characterizing alcohol dependence: transitions during young and middle adulthood. Exp. Clin. Psychopharmacol. 14 (2), 228–244. http://dx.doi.org/10.1037/1064-1297.14.2.228.
Jackson, D.L., Gillaspy, J.A., Purc-Stephenson, R., 2009. Reporting practices in confirmatory factor analysis: an overview and some recommendations. Psychol. Methods 14 (1), 6–23. http://dx.doi.org/10.1037/a0014694.
Johnston, J., Dinardo, J., 1997. Econometric Methods, 4th ed. The McGraw-Hill Companies, Inc., New York, NY.
Jung, T., Wickrama, K.A.S., 2008. An introduction to latent class growth analysis and growth mixture modeling. Social Personal. Psychol. Compass 21 (10), 302–317. http://dx.doi.org/10.1111/j.1751-9004.2007.00054.x.
Kenny, D.A., Zautra, A., 2001. Trait–State Models for Longitudinal Data. In New Methods for the Analysis of Change. American Psychological Association, Washington, pp. 243–263. http://dx.doi.org/10.1037/10409-008.
Kievit, R.A., Brandmaier, A.M., Ziegler, G., van Harmelen, A.L., de Mooij, S.M., Moutoussis, M., ... Lindenberger, U., 2017. Developmental cognitive neuroscience using Latent Change Score models: a tutorial and applications. Dev. Cogn. Neurosci. http://dx.doi.org/10.1016/j.dcn.2017.11.007.
King, D.W., King, L.A., McArdle, J.J., Grimm, K., Jones, R.T., Ollendick, T.H., 2006. Characterizing time in longitudinal trauma research. J. Trauma Stress 19 (2), 205–215. http://dx.doi.org/10.1002/jts.20112.
King, K.M., Molina, B.S.G., Chassin, L., 2009. Prospective relations between growth in drinking and familial stressors across adolescence. J. Abnorm. Psychol. 118 (3), 610–622. http://dx.doi.org/10.1037/a0016315.
Kirby, K.N., 2009. One-year temporal stability of delay-discount rates. Psychon. Bull. Rev. 16 (3), 457–462.
Littlefield, A.K., Alquist, J.L., 2014. Greater clarity with consilience: testing causal models across methodological approaches. Eur. J. Personal. 28, 394–395.
Littlefield, A.K., Sher, K.J., Wood, P.K., 2009. Is maturing out of problematic alcohol involvement related to personality change? J. Abnorm. Psychol. 118 (2), 360–374. http://dx.doi.org/10.1037/a0015125.
Littlefield, A.K., Sher, K.J., Steinley, D., 2010a. Developmental trajectories of impulsivity and their association with alcohol use and related outcomes during emerging and young adulthood I. Alcohol.: Clin. Exp. Res. 34 (8), 1409–1416. http://dx.doi.org/10.1111/j.1530-0277.2010.01224.x.
Littlefield, A.K., Sher, K.J., Wood, P.K., 2010b. A personality-based description of maturing out of alcohol problems: extension with a five-factor model and robustness to modeling challenges. Addict. Behav. 35 (11), 948–954. http://dx.doi.org/10.1016/j.addbeh.2010.06.008.
Littlefield, A.K., Vergés, A., Wood, P.K., Sher, K.J., 2012. Transactional models between personality and alcohol involvement: a further examination. J. Abnorm. Psychol. 121 (3), 778–783. http://dx.doi.org/10.1037/a0026912.
Liu, S., Rovine, M.J., Molenaar, P.C.M., 2012. Selecting a linear mixed model for longitudinal data: repeated measures analysis of variance, covariance pattern model, and growth curve approaches. Psychol. Methods 17 (1), 15–30. http://dx.doi.org/10.1037/a0026971.
Marsh, H.W., Hau, K.-T., Wen, Z., 2004. In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Struct. Equ. Model. – Multidiscip. J. 11 (3), 320–341. http://dx.doi.org/10.1207/s15328007sem1103_2.
Masten, A.S., Cicchetti, D., 2010. Developmental cascades. Dev. Psychopathol. 22 (3), 491–495. http://dx.doi.org/10.1017/S0954579410000222.
McArdle, J.J., 2009. Latent variable modeling of differences and changes with longitudinal data. Annu. Rev. Psychol. 60, 577–605. http://dx.doi.org/10.1146/annurev.psych.60.110707.163612.
McCall, R.B., Appelbaum, M.I., 1991. Some issues of conducting secondary analyses. Dev. Psychol. 27 (6), 911–917. http://dx.doi.org/10.1037/0012-1649.27.6.911.
McCoach, D.B., Kaniskan, B., 2010. Using time-varying covariates in multilevel growth models. Front. Psychol. 1 (June), 17. http://dx.doi.org/10.3389/fpsyg.2010.00017.
McLaughlin, K.A., King, K.M., 2014. Developmental trajectories of anxiety and depression in early adolescence. J. Abnorm. Child Psychol. 43 (2), 311–323. http://dx.doi.org/10.1007/s10802-014-9898-1.
McNeish, D., Stapleton, L.M., Silverman, R.D., 2016. On the unnecessary ubiquity of hierarchical linear modeling. Psychol. Methods 58 (12), 7250–7257. http://dx.doi.org/10.1037/met0000078.
Mehta, P.D., West, S.G., 2000. Putting the individual back into individual growth curves. Psychol. Methods 5 (1), 23–43. http://dx.doi.org/10.1037/1082-989X.5.1.23.
Mills, K.L., Goddings, A.-L., Herting, M.M., Meuwese, R., Blakemore, S.-J., Crone, E.A., et al., 2016. Structural brain development between childhood and adulthood: convergence across four longitudinal samples. Neuroimage 141, 273–281. http://dx.doi.org/10.1016/j.neuroimage.2016.07.044.
Miyake, A., Friedman, N., 2012. The nature and organization of individual differences in executive functions four general conclusions. Current Directions in Psychological Sci. 21 (1), 8–14. http://dx.doi.org/10.1177/0963721411429458.
Miyake, A., Friedman, N.P., Emerson, M.J., Witzki, a H., Howerter, A., Wager, T.D., 2000. The unity and diversity of executive functions and their contributions to complex Frontal Lobe tasks: a latent variable analysis. Cognit. Psychol. 41 (1), 49–100. http://dx.doi.org/10.1006/cogp.1999.0734.
Moffitt, T.E., 1993. Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. Psychol. Rev. 100 (4), 674–701. http://dx.doi.org/10.1037/0033-295X.100.4.674.
Muthén, B., Shedden, K., 1999. Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics 55 (2), 463–469.
Nagin, D.S., Tremblay, R.E., 1999. Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency. Child Dev. 70 (5), 1181–1196. http://dx.doi.org/10.1111/1467-8624.00086.
Nagin, D.S., Tremblay, R.E., 2005. Developmental trajectory groups: fact or a useful fiction? Criminology 43 (4), 873–904. http://dx.doi.org/10.1111/j.1745-9125.2005.00026.x.
Nagin, D.S., 1999. Analyzing developmental trajectories: a semiparametric, group-based approach. Psychol. Methods 4 (2), 139–157. http://dx.doi.org/10.1037/1082-989X.4.2.139.
National Advisory Mental Health Council Workgroup on Tasks and Measures for Research Domain Criteria (RDoC), 2016. Behavioral Assessment Methods for RDoC Constructs Behavioral Assessment Methods for RDoC Constructs TABLE OF CONTENTS. Retrieved from. https://www.nimh.nih.gov/about/advisory-boards-and-groups/namhc/reports/rdoc_council_workgroup_report_153440.pdf.
Nylund, K.L., Asparouhov, T., Muthén, B.O., 2007. Deciding on the number of classes in latent class analysis and growth mixture modeling: a monte carlo simulation study. Struct. Equ. Model. 14 (4), 535–569. http://dx.doi.org/10.1080/10705510701575396.
Odum, A.L., 2011. Delay discounting: trait variable? Behav. Processes 87 (1), 1–9.
Ordaz, S.J., Foran, W., Velanova, K., Luna, B., 2013. Longitudinal growth curves of brain function underlying inhibitory control through adolescence. J. Neurosci. 33 (46), 18109–18124. http://dx.doi.org/10.1523/JNEUROSCI.1741-13.2013.
Ostby, Y., Tamnes, C.K., Fjell, A.M., Westlye, L.T., Due-Tønnessen, P., Walhovd, K.B., 2009. Heterogeneity in subcortical brain development: a structural magnetic resonance imaging study of brain maturation from 8 to 30 years. J. Neurosci. 29 (38), 11772–11782. http://dx.doi.org/10.1523/JNEUROSCI.1242–09.2009.
Pawitan, Y., 2001. In All Likelihood: Statistical Modelling and Inference Using Likelihood. University Press, Oxford: Oxford.
Pitts, S.C., West, S.G., Tein, J.-Y., 1996. Longitudinal measurement models in evaluation research: examining stability and change. Eval. Program Plann. 19 (4), 333–350. http://dx.doi.org/10.1016/S0149-7189(96)00027-4.
Ram, N., Grimm, K.J., 2009. Growth mixture modeling: a method for identifying differences in longitudinal change among unobserved groups. Int. J. Behav. Dev. 33 (6), 565–576. http://dx.doi.org/10.1177/0165025409343765.
Ram, N., Grimm, K., 2015. Growth curve modeling and longitudinal factor analysis. Handbook of Child Psychology and Developmental Science I: Theory 1:20. pp. 1–31.
Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical linear models: applications and data analysis methods. Advanced Quantitative Techniques in the Social Sciences Series, vol. 1 Sage, Los Angeles, CA.
Ritchie, S.J., Bates, T.C., Plomin, R., 2015. Does learning to read improve intelligence? A longitudinal Multivariate analysis in identical twins from age 7–16. Child Dev. 86 (1), 23–36. http://dx.doi.org/10.1111/cdev.12272.
Roberts, B.W., Mroczek, D., 2008. Personality trait change in adulthood. Curr. Direct. Psychol. Sci. 17 (1), 31–35. http://dx.doi.org/10.1111/j.1467-8721.2008.00543.x.
Rogosa, D.R., Willett, J.B., 1985. Understanding correlates of change by modeling individual differences in growth. Psychometrika 50 (2), 203–228. http://dx.doi.org/10.1007/BF02294247.
Rogosa, D., Brandt, D., Zimowski, M., 1982. A growth curve approach to the measurement of change. Psychol. Bull. 92 (3), 726–748. http://dx.doi.org/10.1037/0033-2909.92.3.726.
Rogosa, D., 1988. Myths about longitudinal research. In: Schaie, K.W., Campbell, R.T., Meredith, W., Rawlings, S.C. (Eds.), Methodological Issues in Aging Research. Springer, New York, pp. 171–209.
Romeo, R.D., 2013. The teenage brain: the stress response and the adolescent brain. Curr. Direct. Psychol. Sci. 22 (2), 140–145. http://dx.doi.org/10.1177/0963721413475445.
Rovine, M.J., Molenaar, P.C.M., 2005. Relating factor models for longitudinal data to quasi-simplex and NARMA models. Multivariate Behav. Res. 40 (1), 83–114. http://dx.doi.org/10.1207/s15327906mbr4001.
Salthouse, T.A., 2014. Why are there different age relations in cross-sectional and longitudinal comparisons of cognitive functioning? Curr. Direct. Psychol. Sci. 23 (4), 252–256. http://dx.doi.org/10.1177/0963721414535212.
Schuster, C., von Eye, A., 1998. Determining the meaning of parameters in multilevel models for longitudinal data. Int. J. Behav. Dev. 22 (3), 475–491.
Selig, J.P., Little, T.D., 2012. Autoregressive and cross-lagged panel analysis for longitudinal data. In: Laursen, B., Little, T.D., Card, N.A. (Eds.), Handbook of Developmental Research Methods. Guilford Press, pp. 265–278.
Sher, K.J., Jackson, K.M., Steinley, D., 2011. Alcohol use trajectories and the ubiquitous cat's cradle: cause for concern? J. Abnorm. Psychol. 120 (2), 322–335. http://dx.doi.org/10.1037/a0021813.Alcohol.
Somerville, L.H., Jones, R.M., Ruberry, E.J., Dyke, J.P., Glover, G., Casey, B.J., 2013. The medial prefrontal cortex and the emergence of self-conscious emotion in adolescence. Psychol. Sci. 24 (8), 1554–1562. http://dx.doi.org/10.1177/0956797613475633.
Steinberg, L., Icenogle, G., Shulman, E.P., Breiner, K., Chein, J., Bacchini, D., et al., 2017. Around the world, adolescence is a time of heightened sensation seeking and immature self-regulation. Dev. Sci. e12532. http://dx.doi.org/10.1111/desc.12532.
Stoolmiller, M., 1995. Using latent growth curve models to study developmental processes. The Analysis of Change. pp. 105–138.
Tamnes, C.K., Herting, M.M., Goddings, A.-L., Meuwese, R., Blakemore, S.-J., Dahl, R.E., et al., 2017. Development of the cerebral cortex across adolescence: a multisample study of inter-related longitudinal changes in cortical volume, surface area, and thickness. J. Neurosci. 37 (12), 3402–3412. http://dx.doi.org/10.1523/JNEUROSCI.3302-16.2017.
Tomarken, A.J., Waller, N.G., 2003. Potential problems with well fitting models. J. Abnorm. Psychol. 112 (4), 578–598. http://dx.doi.org/10.1037/0021-843X.112.4.578.
Usami, S., Hayes, T., McArdle, J.J., 2015. On the mathematical relationship between

latent change score and autoregressive cross-lagged factor approaches: cautions for inferring causal relationship between variables. Multivariate Behav. Res. 50 (6), 676–687. http://dx.doi.org/10.1080/00273171.2015.1079696.

Vandenberg, R.J., Lance, C.E., 2000. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organ. Res. Methods 3 (1), 4–70. http://dx.doi.org/10.1177/109442810031002.

Vijayakumar, N., Mills, K.L., Alexander-Bloch, A., Tamnes, C.K., Whittle, S., 2017. Structural brain development: a review of methodological approaches and best practices. Dev. Cogn. Neurosci (this issue).

Wöstmann, N.M., Aichert, D.S., Costa, A., Rubia, K., Möller, H.-J., Ettinger, U., 2013. Reliability and plasticity of response inhibition and interference control. Brain Cogn. 81 (1), 82–94. http://dx.doi.org/10.1016/j.bandc.2012.09.010.

Walhovd, K.B., Fjell, A.M., Giedd, J., Dale, A.M., Brown, T.T., 2016. Through thick and thin: a need to reconcile contradictory results on trajectories in human cortical development. Cereb. Cortex 1989http://dx.doi.org/10.1093/cercor/bhv301. bhv301.

Weafer, J., Baggott, M.J., Wit, H., De de Wit, H., 2013. Test-retest reliability of behavioral measures of impulsive choice, impulsive action, and inattention. Exp. Clin. Psychopharmacol. 21 (6), 475–481. http://dx.doi.org/10.1037/a0033659.

Wood, P.K., Steinley, D., Jackson, K.M., 2015. Right-sizing statistical models for longitudinal data. Psychol. Methods 20 (4), 470–488. http://dx.doi.org/10.1037/met0000037.

Zeger, S.L., Liang, K.Y., Albert, P.S., 1988. Models for longitudinal data: a generalized estimating equation approach. Biometrics 44 (4), 1049–1060. http://dx.doi.org/10.2307/2531734.