Exploring reading and mathematics growth through psychometric innovations applied to longitudinal data

Gary L. Williamson

Cogent Education (2018), 5: 1464424
Exploring reading and mathematics growth through psychometric innovations applied to longitudinal data

Gary L. Williamson*

Abstract: Individual growth curves yield insights about growth that are not available from any other methodology; and developmental scales based on conjoint measurement models provide unique interpretive advantages for investigations of academic growth. The advantages are apparent when: (1) 15 consecutive statewide reading growth curves are annotated with historical policy actions; and (2) mathematics growth, mathematics achievement standards, and the difficulty of mathematical skills and concepts are conjointly quantified. Sequential examples demonstrate what can be gained when educational measurement scales are constructed to exhibit general objectivity (i.e. the absolute location of a person or an item on the scale is sample independent).

Subjects: Education; Education Policy; Research Methods in Development Studies; Educational Research; Education & Development; Research Methods in Education; Assessment

Keywords: developmental scale; conjoint measurement; specific objectivity; general objectivity; growth; reading; mathematics; longitudinal

ABOUT THE AUTHORS
Gary L. Williamson is Senior Research Scientist at MetaMetrics® in Durham, North Carolina. Williamson received his PhD from Stanford University in Mathematical Methods in Educational Research. He holds additional degrees in Statistics (MS, Stanford University), Educational Research & Evaluation (MED, University of North Carolina at Greensboro), and Mathematics (BS, University of North Carolina at Chapel Hill). Since 1974, his research has focused primarily on educational achievement and growth. The research reported in this paper explores reading and mathematics growth in North Carolina (1995–2014) and simultaneously pertains to methodological issues in the study of growth. Dr Williamson has three times received the Distinguished Paper Award from the North Carolina Association for Research in Education (NCARE) and the Consortium of State and Regional Educational Research Associations (SRERA). In October 2010, Governor Bev Purdue awarded him the Order of the Long Leaf Pine for extraordinary service to the state of North Carolina.

PUBLIC INTEREST STATEMENT
In the twentieth century, educational research was scarce with respect to student reading growth or mathematics growth, largely because of insufficient data, inadequate statistical methods and lack of a common scale for measuring student achievement over the long term. With the dawning of the twenty-first century, the situation improved. The goal of this research was to take a fresh look at student growth applying the newest methods of analysis, the best measurement scales and the most comprehensive data comprised of serial measures of achievement for the same individuals over time. Among the results, the crowning success is an analysis of reading growth for 15 successive groups of students, whose reading achievement is documented across multiple years with a common scale of measurement, making growth results more interpretable. The analysis constitutes a pivotal advance in educational research and measurement and provides an enlightening exemplar for future studies of academic growth.
1. Introduction
The approach to academic growth (e.g. reading growth, mathematics growth) adopted for this paper is based on a research tradition in the physical sciences that reaches back over 240 years (Tanner, 2010). In educational practice, that tradition frequently has been ignored (Koretz & Hamilton, 2006) or deliberately eschewed (Castellano & Ho, 2013). Yet, some educational research methodologists (e.g. Singer & Willett, 2003) advocate quantitative methods for the study of academic growth that respect the traditions established in the study of physical growth. This paper follows the latter path.

There are at least two important advantages of viewing academic growth from the perspective of individual growth curves. First and foremost, individual growth curves yield understandings and insights about growth that are not available from any other methodology. In this paradigm, the magnitude, velocity, and acceleration of growth are readily explored and a long-term view of development can be maintained quantitatively. Secondly, individual growth curves accrue additional unique interpretive advantages when they are based on conjoint measurement. In that case, individual growth can be seen in light of the difficulty or challenge that students face as they grow academically. In particular, interpretations of academic growth benefit from Rasch-based measurement scales that have been anchored in a quantified construct continuum by means of a construct specification theory. Over two decades ago, one state had the foresight to begin using such a scale with longitudinal data; their commitment perseveres today.

Measurement innovations adopted in North Carolina (NC) provide interesting, insightful interpretations of academic growth, which have become progressively more useful over time as the state’s measurement scales attained increased measurement generality. Three successive examples illustrate how the measurement of growth has improved as the state progressed from (1) a traditional item response theory (IRT) vertical scale to (2) a Rasch measurement scale (possessing specific objectivity and providing conjoint measurement) for mathematics ability, to (3) a Rasch measurement scale for reading ability that is integrated with a construct specification theory. In the first instance, although interpretations of growth are interesting, it must be acknowledged that they are not generalizable beyond the particular collection of growth curves. In the second example, because of specific objectivity and conjoint measurement, a single scale attains increased generality and can simultaneously address multiple interpretive perspectives, including for instance (1) student mathematics growth, (2) mathematics achievement level standards, (3) the difficulty of mathematics textbook lessons, and (4) the mathematics ability a student needs to be ready for college or career. In a final example, average reading growth curves for 15 successive panels can be cogently annotated with historical policy actions because all measurements are expressed on a Rasch measurement scale with general objectivity (Stenner, 1990, 1994).

2. Two core tenets for academic growth
There are two core tenets guiding the approach to academic growth used in this paper. First, growth must be formulated in terms of serial measures collected on the same individuals over time. Second, measurement scales used for the study of academic growth must have certain psychometric properties to support the scientific study of academic growth.

2.1. Serial measures on the same individual
In a series of papers published from 1841 to 1844, Lehmann became the first person to elucidate and emphasize the importance of collecting and analyzing serial measures on the same individual for the study of growth in stature. From 1892 to 1941, Boas persistently insisted on the importance of following the same individuals over time in studies of physical growth. Tanner (2010) traced this theme throughout A History of the Study of Human Growth, initially published in 1981. Similarly, Rogosa, Brandt, and Zimowski (1982) advocated the use of expanded longitudinal data collection designs in education, utilizing more than two waves of serial measures on the same individuals, accompanied by an analytical methodology focused on the individual growth curve. This approach is consonant with a developmental perspective advocated by educational psychologists (e.g. Baltes &
Nesselroade, 1979; Guire & Kowalski, 1979; Wohlwill, 1970). Though he is perhaps better known for his creation of the measurement model that bears his name, Georg Rasch also acknowledged the fundamental importance of focusing on the individual in his studies of growth (Olsen, 2003).

The term longitudinal has sometimes been used to connote a study in which information is collected on multiple occasions, but the same individuals are not involved on each occasion. In this paper, a narrowly construed definition of purely longitudinal is used. The fundamental characteristic of a purely longitudinal research design is the collection of serial measures on the same individuals on multiple occasions of measurement. In such a study, the central purpose is to understand the temporal nature of the developing construct. The dual aims are: (1) first to understand individual growth over time and then (2) to generalize about the growth of groups of individuals. To obtain maximal information about individual growth, one must adhere to this narrow definition of a longitudinal study, which Goldstein (1979) called a panel study.

Moreover, the analytical framework used for the study of individual growth must be developmentally grounded. It is certainly possible to generate group-level, cross-sectional analyses of panel data. However, this approach wastes important information contained in the data and may give rise to misleading results (Loken, 2010; Molenaar & Newell, 2010; Robinson, 1950; Tanner, 2010). Longitudinal analyses should incorporate the temporal nature of the data into the statistical model at the individual level. Thus the researcher focuses first on the individual growth curve (i.e. a model of intra-individual variation) when approaching the measurement of academic growth. These requirements can be accommodated by multilevel (hierarchically structured) statistical analyses (Bock, 1989; Goldstein, 1987; Singer & Willett, 2003; Raudenbush & Bryk, 2002).

The person-centered, developmental approach to growth presumes an individual growth process, which can be modeled via a suitable mathematical function. In that paradigm, selecting a growth function is fundamental to the parametric modeling of serial measures on an individual. There are many mathematical functions that may plausibly describe growth (Goldstein, 1979; Hauspie & Molinari, 2010; Willett, 1988). Depending on its intended use, a growth model may be expected to possess any of a number of characteristics such as developmental verisimilitude, adequate statistical fit, parsimony, ease of interpretation, dynamic consistency (Keats, 1983; Merrell, 1931), or a compelling parameterization of key features of growth (e.g. magnitude, velocity, and acceleration). Singer and Willett (2003) detailed a systematic approach to choosing an appropriate growth model. Their disciplined approach was used for this paper.

Collecting serial measures on the same individual is the first core tenet of the study of academic growth. However, the success of the endeavor also depends crucially on the psychometric properties that can be ascribed to the measures that are collected.

### 2.2. Measurement scales that support the study of growth

#### 2.2.1. Essential scale properties

In this paper, five specific properties are deemed essential for the optimal measurement of student growth. The measurement scale must be unidimensional, continuous, developmental, and equal-interval; and, it must have invariant scale location and unit size (Williamson, 2015). Invariance of location and unit size are regarded as a unitary property because an invariant unit size has traditionally been established in the physical sciences by instantiating an invariant location for the scale (e.g. as in temperature and length). These scale properties are necessary to attain generally objective measures analogous to measures used in the physical sciences—e.g. the meter (length). These five properties must be addressed in the design and development of a measurement scale.

A unidimensional scale for measuring academic growth is designed to target a persistent, measurable construct (e.g. reading ability, mathematics ability) that changes quantitatively (but not qualitatively) within person over time. This assures that the same construct is measured on every
occasion and eliminates the possibility of confusing alternative constructs that develop simultaneously but perhaps differently. Measurement specialists use extensive protocols (e.g. North Carolina Department of Public Instruction, 2009, 2016) to create tests designed to match particular content specifications, to develop test items to measure the targeted construct and to detect departures from unidimensionality.

To support inferences about growth velocity and acceleration, a measurement scale must be continuous (because calculus is applied to the growth function to derive velocity and acceleration curves). Stage development theories (e.g. Chall, 1996) need not be fundamentally inconsistent with assuming an underlying continuum of development. Any continuous scale can be discretized to represent stages of development (e.g. the adolescent growth spurt in height) and discontinuous growth can be modeled statistically (Singer & Willett, 2003). In contrast, categorical scales are not adequate to represent continuous development and it is impossible to capture accurately the velocity or acceleration of growth with a merely ordinal scale. Typically, scale continuity as a measurement design feature is accommodated by adopting the real number line as the means of denoting quantity of the target construct.

In addition to being unidimensional and continuous, scales for measuring academic growth must be designed to be developmental. A developmental scale is constructed specifically for measuring growth across ages or grade levels (Williams, Pommerich, & Thissen, 1998). Consequently, a developmental scale is often called a vertical scale. Among other things, being vertically scaled implies that measurement is invariant across time (e.g. as measured by grade or age) and across instruments/items. In education, person ability constructs are usually aligned with a content curriculum that spans the grades in school. A well-designed curriculum itself should be articulated so as to reveal how person ability manifests itself in the development of the individual over time. Most curricula are based on a developmental ordering of content to accomplish this goal, as exemplified by the Common Core State Standards (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). Vertical scales can be created with a variety of scaling designs (Kolen, 2006). However, the functionality of a developmental scale can also be attained by integrating construct theory with a Rasch measurement model (Williamson, 2015).

Scales used for the study of growth also must be equal-interval. That is, a unit-change should have the same quantitative meaning anywhere along the scale. Said another way, a change of one unit should represent the same amount of the underlying construct anywhere along the scale. Without equal intervals one cannot assess the velocity or acceleration of growth because changes in the construct per unit of time would have different meanings at different points on the scale. The ability scale produced by item response theory (IRT) models is widely taken to have the equal interval property because it is determined up to a linear transformation (Yen & Fitzpatrick, 2006). In contrast, the Rasch measurement model achieves quantification by analogy with the Standard Model in the physical sciences, advocated by Maxwell and others (Fisher, 2010). For example, Rasch (1960) formulated a measurement model relating reading accuracy, reading ability and item difficulty using the logic of Newton’s Second Law of Motion relating force, mass, and acceleration and he noted that this approach “provides a principle of measurement on a ratio scale of both stimulus [item] parameters and object [person] parameters, the conceptual status of which is comparable to that of measuring mass and force” (p. 115).

When a scale is unidimensional, continuous, developmental, and equal-interval, the researcher is using the same ruler in one grade as in another. It is important to understand that these properties must be psychometrically designed into the measurement; they cannot be statistically imposed after the fact. The widely used practice of standardizing scores on successive occasions does not produce an equal-interval, developmental scale, and is, in fact, counter-productive for the study of growth (Willett, 1988).
Nevertheless, even when equal-interval, developmental scales are successfully constructed, they are particular to a given test series and edition. New scales have to be created whenever new tests are designed or new editions of existing tests are published. Thus, comparability is generally unsupported across different tests or even across different editions of the same test unless a linking study is performed to specifically address the issue. This short-coming leads naturally to the final requirement for measuring growth—a measurement scale that is defined in an absolute frame of reference and has an invariant unit.

The magnitude of the scale unit—that is, the distance between two consecutive points on the scale—is the fundamental increment used to express distance. For example, the meter (m) is the base unit for length in the International System of Units (SI). Without an invariant scale unit, measurements have no general objectivity (as for example, the height of a person does) and repeated measurements lack sufficient comparability from one occasion to another. An invariant scale unit is clearly defined, reproducible, constant over time for any specific individual, and retains the same absolute magnitude regardless of which particular individual is measured. Invariance of location and unit size have been obtained for the measurement of reading ability by integrating a Rasch measurement model with a construct specification theory and anchoring the scale at two substantively meaningful points on the construct continuum (Williamson, 2015). The Lexile® Framework for Reading (Stenner, Burdick, Sanford, & Burdick, 2007), was designed to satisfy these measurement properties. It was adopted in North Carolina around two decades ago as a supplemental scale for assessing reading ability as manifested through student reading achievement.

2.2.2. Putative properties of the scales used in subsequent examples

In univariate studies of academic growth, the individual growth curve is a mathematical function that expresses how a latent individual construct (e.g. reading ability) changes within person over time. Consequently, there are two essential dimensions for a bi-variate plot of the growth curve. The person-ability dimension is usually plotted on the vertical axis; the time dimension is plotted on the horizontal axis. The scales of both dimensions must possess the five essential properties discussed earlier.

Table 1 summarizes the properties of measurement scales used in subsequent examples. Each column in the table corresponds to a different scale. Three different scales are used to measure

| Table 1. Putative properties of selected measurement scales |
|----------------------------------------------------------|
| **Name of scale (construct measured)** | **NC End-of-Grade Mathematics (Ed. 1) Scale** | **Quantile Scale** | **Lexile Scale** | **Time** |
| Scale property | (Mathematics ability) | (Mathematics ability) | (Reading ability) | (Temporal length) |
| Unidimensional | ✓ | ✓ | ✓ | ✓ |
| Continuous | ✓ | ✓ | ✓ | ✓ |
| Developmental | ✓ | ✓ | ✓ | ✓ |
| Equal interval | ✓ | ✓ | ✓ | ✓ |
| Specific objectivity | ✓ | ✓ | ✓ | ✓ |
| Conjoint measurement | ✓ | ✓ | |
| General objectivity | ✓ | ✓ | |

Notes: Specific objectivity requires that differences between person-measures (and differences between item calibrations) are sample-independent. General objectivity requires that the absolute measure of an individual (or an item calibration) is sample independent. Conjoint measurement makes it possible to place person measures and item calibrations on a common scale. For the optimal measurement of student growth, measurement scales for both construct and time must be unidimensional, continuous, developmental, and equal-interval; and, must possess invariant scale location and unit size (general objectivity).
person abilities (e.g. either mathematics ability or reading ability) in subsequent examples. These three scales are used to illustrate the consequences of having scales with different measurement properties applied to the study of academic growth. The fourth scale, \textit{Time}, is paired with one of the person construct scales for each of the subsequent growth examples. The rows in Table 1 correspond to different measurement properties. In addition to the five essential properties for studies of student growth, two additional measurement properties, specific objectivity (Rasch, 1977) and conjoint measurement (Luce & Tukey, 1964), are also listed because one person-construct scale lacks specific objectivity and cannot sustain conjoint interpretations of growth.

Historically, North Carolina used a variety of scales for measuring student performance on academic subjects. Three specific scales were selected for the examples featured in this paper. The NC End-of-Grade (EOG) Test of Mathematics features a mathematics scale score as its primary measurement scale. The scale was developed using the three-parameter logistic item response theory model (3PL IRT). As displayed in Table 1, the first four essential properties for the measurement of student growth characterize the EOG mathematics scale; however, it does not possess specific objectivity, conjoint measurement or general objectivity. Specific objectivity is not essential for characterizing academic growth, but it enables conjoint interpretations of student growth (Bond & Fox, 2001; Karabatsos, 2001). In 2008, the Quantile scale was incorporated as a secondary score scale for reporting mathematics achievement in NC. This scale possesses the first four essential properties for the study of growth; additionally, the Quantile scale has specific objectivity. Consequently, conjoint interpretations of student growth are possible using the Quantile scale. However, the Quantile scale does not possess general objectivity. The Lexile scale possesses all seven of the measurement properties listed in Table 1. Thus, it not only possesses the properties essential for growth, but it also has specific objectivity enabling conjoint interpretations of student reading growth and text complexity. As will be demonstrated, these differences have implications for the measurement and interpretation of academic growth.

The measurement of time is rigorously standardized and attains all five measurement properties essential for studies of academic growth. Time is operationalized as unidimensional, continuous and developmental by its very nature. The scale unit (i.e. second) attains its equal-interval property and general objectivity through standardization in terms of the frequency associated with specific atomic transitions of Cesium 133. The modern measurement of time is so precise that “the magnitude of discrepancies among timekeeping standards is far smaller than is required by almost all practical applications” (Tal, 2016, p. 308). The time scale possesses specific objectivity by virtue of its general objectivity; however, the time scale is not an example of conjoint measurement. This does not adversely affect its utility for the study of growth.

3. Method

3.1. Research strategy
The overarching thesis of this paper is that the study of academic growth is improved when growth is formulated in terms of individual growth curves and the measurement of growth is based on conjoint developmental scales that possess general objectivity. Three sequential examples will showcase what may be gained with better measurement and what is lost when measurement falls short. To that end, I present growth under three different measurement scenarios. In the first scenario, growth curves are quantified with data that possess merely interval measurement properties. In the second example, growth is quantified with an interval level, conjoint measurement scale that possesses specific objectivity. In the third example, growth is measured with a conjoint, developmental scale that possesses general objectivity.
Two research questions underlie all empirical growth analyses:

(1) What is the shape of the developmental trajectory (e.g. straight-line or curvilinear)?
(2) Are parameter estimates for specific features of growth (e.g. magnitude, velocity, acceleration) different from zero?

In the subsequent expositions, a third research question is also addressed.

(3) Do empirically observed features of growth correspond to external contexts or educational policy initiatives?

3.2. Data

Two types of data were used for the examples presented in this paper. The underlying longitudinal data for student growth consisted of individual student achievement measures. These data make it possible to characterize aggregate student growth curves for reading and mathematics achievement. Then, measures of the difficulty of mathematics skills and concepts make it possible to depict the difficulty of mathematics lessons appearing in textbooks associated with public schooling in the United States.

3.2.1. Student reading and mathematics assessment data

The aggregate student growth curves depicted in this paper were derived from analyses of individual reading and mathematics measures collected from administrations of the North Carolina End-of-Grade (EOG) Tests of Reading Comprehension and Mathematics, which are multiple choice tests administered in paper-and-pencil format. The NC EOG Testing Program was designed to satisfy statutory requirements related to the implementation of the North Carolina Standard Course of Study (NC SCS) and the state’s accountability program. To that end, EOG tests are closely aligned with the NC SCS during development. The EOG tests were designed to accurately measure the knowledge and skills of individual students (as well as groups of students) and to enable the state to monitor the academic growth of its students (Sanford, 1996). The tests are administered in Grades 3–8 near the end of each school year.

There have been four editions of the NC EOG Tests of Reading Comprehension and Mathematics. The first edition was made operational in 1992–1993. Periodically, the tests are revised to reflect curricular revisions. Accordingly, in 2003, a second edition of the reading test (Bazemore & Van Dyk, 2004) was made operational; and in 2008, a third edition (North Carolina Department of Public Instruction, 2009) was made operational. Similarly, in 2001, a second edition mathematics test (Bazemore, Van Dyk, Kramer, Yelton, & Brown, 2004) became operational; and, a third edition (Bazemore, Kramer, Gallagher, Englehart, & Brown, 2008) became operational in 2006. The current (fourth edition) reading and mathematics tests, based on the Common Core State Standards, became operational in 2013. Mbella, Zhu, Karkee, and Lung (2016a, 2016b) documented the technical characteristics of the fourth edition.

In the first edition assessments, internal consistency (coefficient alpha) ranged from 0.90 to 0.94 for the reading comprehension tests and from 0.91 to 0.94 for mathematics tests (Sanford, 1996). Criterion and construct validity were also supported by several studies reported by Sanford. Additional technical reports (cited in the previous paragraph) confirmed that sound psychometric characteristics have been maintained in subsequent editions of tests. The NC EOG tests have undergone several audits and technical reviews since their initial development, including the assessment peer reviews required under the Elementary and Secondary Education Act (ESEA). They consistently have been found technically sufficient for the purposes for which they were designed.

To facilitate the measurement of growth, the first edition of the NC EOG Tests was designed with a developmental scale for reading and a developmental scale for mathematics. Both scales were
made to range from 100 to 200 points, approximately. The vertical scales were constructed by means of a common-items equating design. The second edition tests reflected a revision of the Standard Course of Study and so new developmental scales were designed for reading and mathematics to range from approximately 200–300 points to differentiate them from the first edition scales. A linking study was conducted to allow transformation of the second edition scores onto the first edition scales for accountability purposes (i.e. to facilitate gain calculations for one transition year until two years of data were available on the new scale). Similarly, the third edition reflected a new revision of the NC SCS and so a third pair of developmental scales was designed to range from approximately 300–400 points to differentiate them from the second and first edition scales. A new linking study allowed the transformation of the third edition scores onto the second edition scale for accountability purposes. Similarly, a fourth edition\(^2\) (NC READY EOG) was created for reading and mathematics and linked with the third edition scales.

Table 2 summarizes the mathematics test editions that contributed data for the mathematics growth examples presented in this paper. The mathematics data layout for ten panels is illustrated. The panels are represented by the diagonals in the table; in each cell is a designation of which edition (1, 2, or 3) of EOG tests produced the mathematics measures in any given year.

In this longitudinal data collection design, grade, and year are confounded. The rows in the table correspond to the grades in which data were collected; the columns correspond to the years when data were collected. The diagonals in the table represent serial measures for the same individuals. Each diagonal represents Grades 3–8 for a given panel of students who progressed across those grades in different years. So, the layout of the table clarifies what grade students were in at any given time. For example, the first diagonal (moving from left to right) corresponds to students who were completing Grade 3 in spring 1995 and progressed from grade to grade (without repeating a grade) until they completed Grade 8 in spring 2000. The table also reveals that all of the scores for this panel of students came from NC EOG Mathematics, Edition 1. Later panels obtained data from multiple editions of tests. It should be understood that all tests were administered in the spring at the end of the grade.

Table 2. North Carolina mathematics assessment panels with mathematics test editions

| 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
|----|----|----|---|---|---|---|---|---|---|---|---|
|    |    | 1  | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 |
|    |    | 1  | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 |
|    |    | 1  | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 |
|    | 1  | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 3 |
|    | 1  | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 |
|    | 1  | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
|    | 1  | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 |
|    | 1  | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 |
|    | 1  | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |

Notes: The NC EOG Mathematics Edition 2 was linked to Edition 1 by equipercentile method in 2001. Edition 3 was linked to Edition 2 in 2006, ostensibly by equipercentile method. The EOG Mathematics scale for Edition 3 was linked to The Quantile® Framework for Mathematics in 2009. EOG Mathematics Editions 1 and 2 were not linked directly to the Quantile Framework. The first five panels were the subject of growth curves expressed in the text through the Mathematics scale score; the more recent five panels were the basis for growth curves reported on the Quantile scale.
Similarly, Table 3 summarizes the reading test editions that contributed scores for each longitudinal panel of students. From Table 3 we see the layout of grades and years that constitute the 15 panels used for the example of reading growth presented later. It should be clear that the reading growth example will utilize scores from all four editions of reading tests. This was accomplished by linking all of the scores to a common scale.

NC EOG Reading scales for Editions 1, 3, and 4 were linked with The Lexile® Framework for Reading. Using the between-edition links, EOG Reading Edition 2 scores were brought onto the scale for one of the adjacent reading editions and then converted to the Lexile scale for growth analyses. Only the NC EOG Mathematics scales for Edition 3 and Edition 4 were linked directly to The Quantile® Framework for Mathematics. Consequently, where necessary Edition 1 and Edition 2 mathematics scores first were brought onto the Edition 3 scale by applying the between-edition links and then converted from Edition 3 directly to the Quantile scale.

3.2.2. Difficulty of mathematical concepts and skills
Because the Quantile Framework is based on the Rasch measurement model (Rasch, 1960; Wright & Stone, 1979), it provides conjoint measurement for student mathematics ability and the difficulty of mathematics concepts and skills. Sanford-Moore et al. (2014) quantified the mathematical demand of textbook lessons based on the mathematics concepts and skills contained in them. She and her colleagues depicted the lesson continuum by a series of box plots representing the within-grade distributions of lesson difficulties for grades and courses extending from Kindergarten through Algebra 2. Williamson, Sanford-Moore, and Bickel (2016) extended the lesson continuum with the distribution of lesson difficulties associated with pre-calculus and trigonometry and furthermore provided an interpretive framework for inferring the range of student mathematical ability (1220Q-1440Q) needed to be ready for college and career. Their results provide a context for visualizing mathematics growth in relation to the demands of mathematics concepts and skills that characterize public schooling up to the boundary of postsecondary entry. This context will be used to provide conjoint interpretations for one of the growth examples presented below.

3.3. Analysis
EOG reading scores for 15 successive panels (Table 3) of eighth graders (2000–2014) and mathematics scores for ten successive panels (Table 2) of eighth graders (2000–2009) were assembled

| Table 3. North Carolina reading assessment panels with reading test editions |
|---|---|---|
| 12 | | |
| 11 | | |
| 10 | ACT | ACT |
| 9 | EI | EI |
| 8 | EI | EI |
| 7 | EI | EI |
| 6 | EI | EI |
| 5 | EI | EI |
| 4 | EI | EI |
| 3 | EI | EI |
| 2 | EI | EI |
| 1 | EI | EI |
| K | 1995 | '96 | '97 | '98 | '99 | 2000 | '01 | '02 | '03 | '04 | '05 | '06 | '07 | '08 | '09 | '10 | '11 | '12 | '13 | '14 |

Notes: NC End-of-Grade (EOG) Reading Edition 2 was linked to Edition 1 in 2003 by an equipercentile method. Edition 3 was linked to Edition 2 in 2008 by the Stocking-Lord method. NC READY EOG Reading Edition 4 was linked to NC EOG Reading Edition 3 by the Stocking-Lord method in 2013. Various editions of tests were linked to The Lexile® Framework for Reading as follows: EOG Edition 1 (1996); EOG Edition 3 (2008); End-of-Course (EOC) English I (EI) (2008); ACT (2013); NC READY EOG (2013); and, EOC English II (EII) (2013).
retrospectively for longitudinal data analysis. Each panel consisted of scores from Grade 3 through Grade 8, but in different calendar years. For the examples in this paper, I based the growth analyses on students who had all six measurements and progressed from grade to grade without repeating a grade. This was done to (1) emphasize the growth of students who received maximal exposure to the Public Schools of North Carolina; (2) avoid the complication and potential bias of incorporating some students whose growth occurred over a longer timeframe (because of repeating one or more grades); and (3) to facilitate comparability across panels.

The data in each panel examined for this study spanned at least six years, a long enough time that a curvilinear model may be an appropriate choice for modeling growth. Lacking a substantive reason to pick a more complicated model, Willett’s (1988) advice disposes one to consider a quadratic model for growth. This choice has been made by others as far back as Wishart (1938), but more recently by Bryk and Raudenbush (1989) and Raudenbush and Bryk (2002) to model growth in infants’ vocabulary sizes and by Singer and Willett (2003) to model change in children’s externalizing behavior.

Williamson, Appelbaum, and Epanchin (1991) found that a straight-line growth model worked reasonably well with eight waves of vertically scaled longitudinal data collected from 1978 to 1985 for a panel of students progressing through Grades 1–8 in Durham County, NC. However, a quadratic model provided a significantly better fit for some students in that panel. This suggests that a quadratic model with modest curvature could work well for more recent data collected in North Carolina from Grades 3–8. Schulte, Stevens, Elliott, Tindal, and Nese (2016) confirmed this when they used the quadratic growth model in a recent analysis of NC data. A quadratic polynomial has the advantage of being parsimonious, yet flexible enough to capture magnitude, velocity, and acceleration.

Following Singer and Willett (2003), a first-stage analysis strategy involved constructing exploratory plots of the individual longitudinal data, non-parametrically smoothing the trend data with splines, fitting several potential OLS regression models to the individual data, and looking at the distribution of fit statistics in the population. Aggregate trends in performance across grades 3–8 were also examined for successive cohorts of students.

Initial exploratory analyses were followed by formal statistical modeling of the data by fitting a sequence of multilevel models for change devised to confirm the expected functional form and yield specific mathematical characterizations of growth. Thus, a parametric model was utilized to represent the developmental nature of individual growth from the end of third grade to the end of eighth grade mathematically. When the hypothesized growth model was fit to the panel data for each cohort, the results included: (1) a fitted model to summarize growth; (2) associated parameter estimates and their standard errors; and (3) estimates of the variability of individual student growth in the state. (The statistical model and empirical results are summarized in Appendix 1). The parameters are directly interpretable in terms of the features of growth—magnitude, velocity and acceleration. By examining the parametric form across successive cohort panels, one can see patterns in salient features of growth in North Carolina over time.

4. Mathematics growth on an equal-interval developmental scale
The five growth curves displayed in Figure 1 represent the average mathematics growth for five successive panels of North Carolina students as each group progressed from Grade 3 to Grade 8 in different years (denoted in the legend). The labels on the time axis represent the occasions when EOG tests were administered each spring. Thus, the unit size on the time scale approximates the time lapse between successive spring test administrations, ignoring small differences due to leap years or variations among school calendars due to weather or other factors. The vertical axis in Figure 1 represents the IRT scale score adopted for the NC EOG Test of Mathematics, Edition 1. The original individual scores for students comprising the five panels were gathered from the first two editions of the NC EOG Mathematics tests as shown in Table 2. Scores from Edition 2 were translated onto the scale
The growth curves are naturally ordered chronologically from left to right. The fixed effects parameter estimates (Table A1) are statistically significant. So we retain the hypothesis that growth is curvilinear, characterized by positive initial velocity and negative acceleration. The estimated achievement at the end of Grade 3 increased across the first three panels, and then leveled off. Performance at the end of Grade 8 appears to have declined with the second panel (compared to the first), but Grade 8 status increased monotonically across the last four panels. Visually, the last four panels exhibited greater curvature than the first panel. This is corroborated by larger estimates of curvature for the last four panels.

The state initiated a new accountability program in 1996–1997, which set specific growth expectations for schools. The apparent systemic improvement indicated by the general rise of the growth curves would be consistent with a hypothesis that improvement in student performance was influenced by the accountability system.

4.1. Measurement implications

Figure 1 provides an interesting summary of mathematics growth for five panels of students early in the history of North Carolina’s EOG Testing Program. The analysis of growth was made possible by the vertical scale that the state created for its tests and by the between-edition linking study completed to allow comparisons across the two test editions. This is a notable achievement. However, the results must stand alone. Although links to the mathematics scales for later editions (Edition 3, Edition 4) are available, applying the linking studies requires a choice be made about which edition would provide the single scale for growth. Unfortunately, the growth results are not invariant to the choice of base scale. The growth curves may differ if Edition 3 were used instead of Edition 1, for example. Consequently, even though we retained the hypothesis of curvilinear growth, the empirically estimated velocities and curvatures manifested by these five panels can provide no insight about velocity or curvature in any general sense. Ultimately, the growth curves for these five panels for Edition 1 using the results of an equipercenitle linking study. The unconditional quadratic growth model was fit to the resulting Edition 1 scale scores.

Figure 1. Average mathematics growth curves for five panels of North Carolina students who were tested at the end of each school year in Grades 3–8 with the North Carolina End-of-Grade Test of Mathematics.

Notes: Growth was quantified with the mathematics developmental score scale for Edition 1. The analysis reflects the growth of \( n = 299,635 \) students, distributed as follows across panels: 48,887 (1995–2000); 53,576 (1996–2001); 63,006 (1997–2002); 65,760 (1998–2003); and 68,406 (1999–2004).
cannot be compared definitively to any other analyses of growth with any other mathematics scale because the EOG Mathematics vertical scales do not possess general objectivity. Additionally, student growth in terms of scale scores cannot be interpreted in the context of mathematics skills or concepts either, because the IRT vertical scale does not provide conjoint measurement.

5. Mathematics growth on an equal-interval developmental scale possessing conjoint measurement properties

The mathematics growth curves shown in Figure 2 represent five new panels of students whose mathematics achievement was measured from Grade 3 to Grade 8. Chronologically, these five panels succeeded the panels shown in Figure 1. As in Figure 1, the time axis for Figure 2 is calibrated in years. However, the vertical axis in Figure 2 represents student mathematics achievement expressed as a Quantile measure (Q).

For the growth curves shown in Figure 2, the original individual student mathematics scores came from multiple editions of the NC EOG Test of Mathematics as shown in Table 2. For example, in the earliest panel (2000–2005), the Grade 3 scores came from Edition 1 while scores for Grades 4–8 came from Edition 2. Scores for the next four panels came from Editions 2 and 3, with increasing numbers of Edition 3 scores being present in more recent panels.

However, Edition 3 was directly linked to the Quantile Framework for Mathematics and taking advantage of that fact leads to increased measurement generality and facilitates conjoint interpretations of growth. Specifically, scores from Edition 2 and Edition 1 were first brought onto the Edition 3 scale by means of the between-edition linking studies routinely carried out by the state. Once on the Edition 3 scale, all scores were expressed as Quantile measures, thus enabling conjoint interpretations of student ability and the difficulty of mathematics concepts and skills.

Figure 2 provides another example of mathematics growth in North Carolina that is interesting in its own right. Fixed effects parameter estimates for the growth curves in Figure 2 were all statistically significant (Table A1), leading us to retain the hypothesis that growth is curvilinear with positive initial velocity and negative acceleration across Grades 3–8. It appears that there were upticks in

![Figure 2. Average mathematics growth curves for five panels of North Carolina students who were tested at the end of each school year in Grades 3–8 with the North Carolina End-of-Grade Test of Mathematics.](image)

Notes: Growth was quantified with The Quantile® Framework for Mathematics based on a Rasch measurement model. The analysis reflects the growth of n = 378,262 students distributed as follows across panels: 74,076 (2000–2005); 74,658 (2001–2006); 75,601 (2002–2007); 77,055 (2003–2008); and, 76,872 (2004–2009).
Grade 3 performance in 2002 and 2003. Interestingly, student accountability standards were implemented in Grade 3 in 2002. Thus, the increase in achievement is consistent with a hypothesis that the adoption of higher standards may have influenced student performance in Grade 3. Average achievement in Grade 8 monotonically increased with each subsequent panel. The first three panels manifested greater curvature than the last two panels. Overall, the growth curves give the impression of systemic improvement during the time frame, which was a period of increasing accountability for schools and students in the state.

5.1. Measurement implications
One may be tempted to compare the five curves in Figure 2 to the five curves in Figure 1. Unfortunately, direct comparisons are not justified. We cannot infer whether the magnitude of student achievement in the first five panels was higher, lower, or about the same as the magnitude of achievement in the more recent five panels. Neither can we infer more or less curvature in Figure 2 than in Figure 1, although cosmetically it may appear to be so. Differences in appearance of the two sets of curves are purely cosmetic and due to arbitrary choices made in the construction of the two graphs for aesthetic reasons. Although the EOG mathematics scale and the Quantile scale both have equal intervals, they are not the same equal intervals.

The Quantile scale is attractive to mathematics educators because it is calibrated in relation to mathematics skills and concepts contained in widely adopted mathematics content standards and in mathematics textbooks commonly used in the United States. Also, the Quantile scale is linked to numerous mathematics tests. Consequently, the NC mathematics growth displayed in Figure 2 may be reasonably compared to growth in other contexts if that growth was also measured in Quantile scale units. Consequently, using the Quantile scale has the benefit of increased generality because of the scale’s ubiquity.

However, the Quantile scale is anchored in the context of a specific quantified content continuum. If curriculum were to change substantially, there could be implications for the magnitude of the Quantile scale unit. In short, neither the EOG mathematics scale nor the Quantile scale possesses general objectivity. However, the Quantile scale does possess a valuable property not possessed by the EOG mathematics scale. Because the Quantile scale is based on Rasch measurement, it has specific objectivity and can sustain conjoint interpretations.

5.2. Conjoint measurement perspectives for mathematics growth
Multiple perspectives of mathematics growth are shown in Figure 3, using conjoint measurement. First, note that the horizontal axis used in Figure 3 is Grade rather than Year and the code for grade indicates the time during the grade when the EOG tests were administered. So for example, when grade is equal to 3 on the time scale, the 3 indicates the end of Grade 3; grade is equal to 4 at the end of fourth grade, and so on. Because grade and year are confounded in a longitudinal data collection, growth curves can be displayed using either grade or year on the horizontal axis. The difference is that the curves appear chronologically from left to right when year is used to code the time scale; however, the growth curves appear vertically when grade is used to code the time scale. Again, the vertical axis represents the Quantile scale.

Notice that the earliest (2000–2005) and latest (2004–2009) growth curves from Figure 2 are displayed in Figure 3. Now they are vertically arranged with the lowest curve in the plot corresponding to the earlier panel and the highest curve representing the more recent panel. This gives a different visual effect and allows one to immediately see the increased magnitude (elevation) of aggregate (average) achievement status across the time frame.

In addition to mathematics growth, Figure 3 contains multiple additional interpretive contexts made possible because of conjoint measurement. The additional interpretive contexts include (1) student mathematics achievement level standards; (2) the difficulty of mathematics textbook lessons; and, (3) the mathematics ability necessary for college and career readiness.
In conjunction with each edition of the NC EOG tests, student achievement level standards were adopted for the Public Schools of North Carolina by the State Board of Education. Achievement level standards for different test editions are not directly comparable because each test edition has a different vertical scale. However, achievement levels are imbued with greater generality and comparability when expressed on the Quantile scale. There are two reasons. First, Edition 3 and Edition 4 were each linked directly to the Quantile Framework. So achievement level standards for these two editions are brought onto a common scale by expressing both on the Quantile scale. Second, and more importantly, the Quantile scale was developed in conjunction with a mathematics content continuum that has remained relatively stable across multiple test editions. By expressing them on the Quantile scale, the achievement level standards are brought into a conjoint relationship with the difficulty of mathematics skills and concepts reflected in the curriculum. Thereby, the challenge of the mathematics achievement level standards is directly and quantitatively interpretable in the context of the content demand of the curriculum. In Figure 3, horizontal hash marks denote the Quantile measures of the mathematics achievement level proficiency standards in Grades 3–8 for the NC EOG Tests of Mathematics, Edition 3 and Edition 4.

In addition to student growth curves and grade-level achievement standards, Figure 3 represents the difficulty of mathematics text book lessons in each grade, from Grade 3 to Grade 11 (Algebra 2). The dashed line segments provide the upper and lower boundaries of the inter-quartile range of within-grade distributions of mathematics lesson-difficulty measures, expressed on the Quantile scale. Thus, the range between the dashed line segments represents the middle 50% of mathematics lessons in terms of lesson difficulty. Finally, the gray rectangle in the upper right portion of the graph represents the range of mathematics ability required for college and career readiness (Williamson et al., 2016).

Several observations emerge from an examination of Figure 3. First, NC EOG Mathematics Edition 4 had higher achievement level standards than Edition 3. This was the result of a deliberate policy action by the State Board of Education. Notice also that the difficulty of the achievement level standards is directly and quantitatively interpretable in the context of the content demand of the curriculum.
standards was raised in relation to the difficulty of mathematics text book lessons to obtain a more ambitious challenge for students. This should be expected from the fact that the text book lesson difficulties correspond to the student mathematics abilities needed to be ready for instruction while the achievement level standards represent the mathematics abilities students should have after instruction. Thus, both student growth curves and student achievement level standards have increased over time and now align better with the demands of mathematics lessons in US text books. However, the range of abilities students may ultimately need as they finish high school are higher still, as denoted by the increasing mathematics demand during Grades 9–11 and the college and career readiness range depicted at Grade 12. It is important to note that the growth curves in Figure 3 represent the growth of historical panels of students who matriculated before the Edition 4 achievement level standards were adopted and before the studies of mathematics text books or college and career demands which are represented in Figure 3. Nevertheless, Figure 3 does suggest that the NC SBE has raised standards to be more consistent with the most recent content demands and historical growth seems to be rising toward these ambitious new standards.

6. Expressing reading growth on a conjoint developmental scale with general objectivity

Aggregate reading growth curves for 15 successive panels of students are featured in Figure 4 using a common scale for reading achievement. The growth curves are ordered chronologically from left (earlier) to right (more recent). Each growth curve spans Grades 3–8 in different years, noted along the horizontal axis (Year) and in the legend. Each panel consists of students who progressed from grade to grade without repeating a grade and who had complete data across all six occasions.

The vertical axis represents the Lexile (L) scale, a conjoint measurement scale for reading ability which also possesses general objectivity by virtue of its construction. Specifically, the Lexile Framework for Reading integrates a Rasch measurement model with a construct specification equation and anchors the resulting scale at two substantively meaningful points in the construct continuum. This strategy is analogous to the way the temperature scale was established (Chang, 2004; Sherry, 2011).
The fixed effects parameter estimates for the 15 reading panels appear in Table A1. Based on the results, we can retain the research hypothesis that reading growth is curvilinear for these 15 panels in North Carolina. We also can retain the hypothesis that features of growth are non-zero; in particular growth is characterized by positive initial velocity and negative acceleration (i.e. deceleration) over Grades 3–8. The third research hypothesis is informed by annotating Figure 4 with policy actions that occurred over the time frame, as explained next.

6.1. Historical policy annotations

In addition to the displayed growth curves, Figure 4 is annotated schematically with historical policy actions that took place during the time frame (1993–2014). There are three categories of operational policy initiatives: (1) assessment editions; (2) accountability programs; and (3) early intervention reading initiatives. A few comments about the three categories will facilitate understanding the subsequent discussion of results.

6.1.1. Assessment editions

Over time, four consecutive editions of NC EOG Tests of Reading Comprehension have been administered statewide in Grades 3–8—Edition 1 (1993–2002); Edition 2 (2003–2007); Edition 3 (2008–2012); and a fourth edition, called NC READY EOG (2013–present). These editions are annotated across the top of Figure 4. Each edition had its own unique vertical scale. Three of the four scales were directly linked to the Lexile scale via symmetric linking functions. However, Edition 2 was not directly linked to the Lexile scale, though it was linked to Edition 1 by equi-percentile method and to Edition 3 by the Stocking-Lord method (Stocking & Lord, 1983). These linking studies provided the means to translate all scores to the Lexile scale.

6.1.2. Accountability programs

Six waves of accountability initiatives occurred during the time frame and are annotated from left to right in the mid-section of Figure 4. From 1993 to 1996 institutional accountability was focused on cross-sectional analyses of school district average performance (Senate Bill 2). In school year 1995–1996, the state piloted a new accountability system in ten local education agencies using the school building as the unit of analysis for institutional accountability. The program, dubbed the ABCs, was first implemented statewide in elementary schools in 1996–1997 and additionally in high schools the following year. The signal feature of the ABCs was that unique growth expectations were set for each school and growth was based on the performance of the same students over time. There were two generations of the ABCs, distinguished by different methods (gain score vs. status projection) for calculating growth. The current generation of school-level accountability, NC READY, was initiated in 2012–2013. Under NC READY, the Education Value-Added Assessment System (EVAAS) provides school value-added results (SAS Institute, Inc, 2016). Finally, individual accountability became effective for students in Grade 3 and Grade 8 in 2002. Student accountability standards required that individual students must score at or above a grade-level cut score in order to be promoted to the subsequent grade.

6.1.3. Early intervention initiatives

Two early-intervention initiatives were implemented during the time frame for Figure 4. They are annotated just above the horizontal axis. In 1993, SMART START was enacted to provide educational resources for four-year-olds to better prepare them to enter Kindergarten at age five. SMART START was gradually implemented. By 1998, it had been deployed in 55 of the state’s 100 counties. By 2001, SMART START was fully deployed statewide. Also in 2001, a second early intervention program called More at Four was added to provide greater support for four-year-olds. In 2011, the latter program was subsumed and has since been known as the North Carolina Pre-Kindergarten Program.

6.2. Observations about growth and policy initiatives

Each growth curve in Figure 4 visually manifests three features of growth—magnitude, velocity, and acceleration. The overall elevation of each curve in relation to the vertical axis reflects achievement status on each occasion (i.e. the magnitude of growth). The slope of a tangent to a curve quantifies the instantaneous velocity of growth at a particular time. The curvature of each trajectory connotes...
the acceleration of growth. The general impression is that growth was fairly consistent from panel to panel, characterized by positive initial velocity and deceleration over time. Furthermore, there was systemic improvement in the magnitude of growth over time as evidenced in Figure 4 by the general rise of the curves moving from left to right in chronological sequence. Other insights are revealed by considering the three categories of policy initiatives.

In relation to changing assessment editions, an important insight is that the general shape of the growth curves is robust and consistent across time, even though four successive editions of assessments were used and each edition was aligned with a different curricular framework. Capturing this consistency of growth is possible because all four editions of reading tests were calibrated to the same (Lexile) scale.

When growth curves and policy initiatives are viewed together in Figure 4, there are interesting correspondences that suggest how policy actions may have influenced changes in growth patterns over time. Although, I offer some thoughts on possible causal hypotheses, they are limited to NC accountability and early intervention initiatives. For this reason, I must emphasize that what I present is not a demonstration of cause and effect. Rather, I am merely illustrating how one might identify potential causal hypotheses that could be relevant to interpreting growth and performance.

Let the reader scan Figure 4 from left to right and notice the beginning points of the growth curves. There seems to be an uptick in Grade 3 performance in 1997 and initial performance continues to rise for several years before leveling off around 2000. This uptick in the starting point for growth is interesting because 1997 was the first year that ABCs growth expectations applied to all elementary schools and it is also the year when the first students to experience SMART START reached Grade 3.

There seems to be another uptick in initial status of growth in spring 2002 and once again initial status continued to rise for several subsequent years. This is interesting because student accountability standards first applied to Grade 3 in 2002. Also, the panel that began in 2002 would have contained larger numbers of students who benefited from SMART START because the four-year-olds from 1998 (when 55 of 100 counties had the program) would have reached the end of Grade 3 in spring 2002.

Now, observe the end-points of the growth curves (i.e. student performance at the end of Grade 8) for the first six panels. Eighth grade performance was gently rising for the first three panels, following the introduction of the ABCs. Then eighth grade performance increased more rapidly across the next three panels, coincident with the introduction of student accountability standards for eighth graders in 2002. Of course, temporal correspondence is a necessary but not sufficient condition for a causal claim.

Eighth grade achievement peaked in 2004 and 2005. The subsequent four panels exhibited lower Grade 8 performance due to increased deceleration of growth. This coincided with slowly declining reading growth in middle schools (particularly sixth grade) and the subsequent adoption of revised growth standards for the ABCs (2nd generation) effective in 2006.

Other interesting scenarios can be constructed for Figure 4. I mention only one more—the fact that the last five growth curves tend to be longer than the first ten growth curves. This can be traced, in part, to a methodological artifact. For the first ten panels, Edition 2 scores were translated to the Lexile scale after applying the equi-percentile link between Edition 2 and Edition 1. For the last five panels, Edition 2 scores were translated to the Lexile scale after applying the Stocking-Lord link between Edition 2 and Edition 3. The change in linking strategy could account for some of the observed difference in the lengths of the growth curves. Although there may be other possible influences on the length of the final five growth curves, the methodological influence is a possibility that cannot be ruled out. In that light, placing all scores on a common scale not only revealed systemic patterns in
the features of growth across the 15 panels, it may have helped flag an anomalous methodological influence, as well.

Finally, it should be noted that conjoint interpretations are also possible for reading growth. Smith and Williamson (2016) provided an example with detailed discussion in *The Education Standard*. In their explication, they featured five interpretive perspectives using a common scale. They depicted student reading growth near the beginning and end of a twenty year period. Then, they juxtaposed reading achievement level standards from three editions of reading tests used during the time frame. They showed three text complexity perspectives: (a) a K-12 text-complexity continuum, (b) postsecondary text complexity, and (c) median text complexity associated with job entry for 59 selected occupations. The authors used the example to illustrate how states can assure that student growth is well-aligned with achievement standards and postsecondary expectations associated with college and career readiness.

7. Summary
I hold two tenets for the measurement of academic growth. Studies of growth must be based on serial measures of the same individuals over time; and, those measurements must possess certain psychometric properties. The measurement of academic growth is optimal when the Rasch measurement model is integrated with construct theory to provide an operational construct specification equation, leading to a scale that can be anchored in such a manner as to provide invariance of location and unit size. The result is a scale, which by its design attains general objectivity. Furthermore, general objectivity is preserved because the construct specification equation can be used to maintain the unit across instruments (e.g. editions of tests) and populations (e.g. successive longitudinal panels).

In this paper, three scales were used to present three examples showing how measurement properties influence the detection and interpretation of academic growth. In the first example, a traditional IRT vertical scale was used to present five mathematics growth curves. By its construction, the IRT scale had a single anchor point in the developmental continuum, and an arbitrary within-grade standard deviation. The scale distances between grades were empirically determined from the sample of students who participated in a vertical equating study. The scale does not have general objectivity. The five growth curves derived with this scale are interesting but they cannot be compared or generalized to anything else.

In the second example, the developmental scale associated with the Quantile Framework for Mathematics was used to model mathematics growth. The Quantile scale was developed using a Rasch measurement model (MetaMetrics, 2011). To establish the Quantile scale, an anchor point and a scaling constant were chosen to transform logits to a scale similar to the Lexile scale in terms of location and unit size. The identification and ordering of mathematics skills and concepts was guided by mathematics subject matter experts; calibration of mathematics skills and concepts was based on examinee performance during an empirical field study, taking advantage of conjoint measurement. Consequently, mathematics growth can be interpreted relative to the difficulty of mathematics lessons. However, there is as yet no construct specification equation for the Quantile Framework; and although the scale has been meticulously maintained, its location and unit size are tied to normative information about student performance at a specific point in time. In addition, the Quantile Framework is situated in a particular instantiation of K-12 mathematics content characterized by a particular ordering of mathematics skills and concepts. Thus, the Quantile Framework does not yet possess general objectivity.

In contrast with the IRT mathematics scale and the Quantile scale, the Lexile scale does exhibit general objectivity. This was accomplished by integrating Rasch measurement with a construct specification equation (Burdick & Stenner, 1996; Stenner, Smith, & Burdick, 1983) and anchoring the scale at two substantively important points, thereby establishing an invariant unit magnitude (Stenner et al., 2007). With the benefit of general objectivity it was possible to assemble 15 panels of data on a common scale over a span of two decades across four editions of assessments. A key aspect of this accomplishment is the realization that general objectivity was a design objective for the
Lexile Framework and it was approached by analogy with the standard model used in the physical sciences. The contrast between the three growth examples provides a concrete illustration that scales must be constructed to have general objectivity.

There remains some debate about the prospects of producing quantitative measures in the social sciences. Indeed, Briggs (2013) and Domingue (2014) concluded that the Lexile Framework does not provide quantification of reading comprehension and does not possess equal intervals because it did not satisfy the cancellation axioms of additive conjoint measurement in an empirical application. Karabatsos (2001) discussed the fact that even though the Rasch measurement model satisfies the axioms of conjoint measurement, it can be difficult to demonstrate this with data because the axioms are formulated deterministically. In contrast, empirical data always contain some degree of measurement error and item-level data typically contain considerable error. Stenner, Fisher, Stone, and Burdick (2013) formulated a more robust test of the quantification hypothesis based on the principle of a trade-off property that is subject to experimental manipulation. In their explication, the Lexile Framework yielded examples of quantification that supported a causal interpretation of Rasch measurement.

Without demonstration, I have claimed that the Lexile scale possesses general objectivity, which presumes quantification, equal intervals, and sample invariance. My claim is based on the manner in which the Lexile scale was designed and constructed. The integration of a construct specification equation with a Rasch measurement model allowed the Rasch scale to be calibrated in a manner that was sample independent. Furthermore, establishing location and unit size by setting two substantive anchor points on the scale and defining the unit as a fraction of the distance between the two anchor points is analogous to the way that temperature was scaled. It is also analogous to the way that the meter was standardized.

There are some additional aspects of the 15-panel reading growth example that I regard as evidence of general objectivity. First, I note that the results were robust and consistent across successive panels of data spanning 20 years, in spite of evolving differences in curricular standards, a succession of test editions, and changes in psychometric methodology. The fact that growth is detectable and consistently manifests positive velocity and deceleration are prima facie evidence for the internal validity of the assessment program (e.g. alignment between assessments and curricula, the success of the various linking studies that were necessary to maintain a common scale, etc.).

Second, consider the many observed correspondences between changing features of growth and various policy initiatives over successive panels of students. I submit that the observed fidelity between growth and policy actions may be regarded as external validity evidence that the measurement process has captured the intended reading construct and is sufficient to detect plausible external influences on academic growth. This consistency between the features of growth and external events would be less likely (if not impossible) without generally objective measurement. In addition to being consonant with various policy initiatives, the growth is also generally consistent with NC performance trends for fourth graders and eighth graders on the National Assessment of Educational Progress (NAEP) during the time frame of the 15 panels.

Third, as alluded to earlier, an apparent methodological anomaly was detected as a result of using the Lexile scale for reading ability. By calibrating each reading edition scale (except for Edition 2) to the Lexile scale, the NC assessment program maintained the capacity to report the vast majority of its reading data on a common scale regardless of test edition. Edition 2 was brought onto the Lexile scale by using one of the two between-edition links to translate Edition 2 onto an adjacent scale that was linked to the Lexile scale. However, the choice of which between-edition link to use was of significance. The resulting growth curves differed depending on which link was used. Further exploration revealed that the difference in the growth curves was manifested in the original scale scores before they were translated onto the Lexile scale. As it turned out, a different linking methodology was used for the two alternative between-edition linking studies accounting (at least in part) for the
observed difference in the growth curves resulting from the two linking approaches. It was the fact that all scores could be expressed on the Lexile scale that first drew attention to this anomaly. Had Edition 2 been linked directly to the Lexile Scale, I would hypothesize that the observed difference between the growth of the last five reading panels and the growth of the first ten reading panels would be moderated or disappear.

Fourth, the estimation of growth velocity and acceleration is achieved by an application of calculus to a growth function. Student academic ability is manifested through student achievement on tests that have been calibrated to a developmental scale. As a consequence, academic growth also is calibrated in terms of the same developmental scale. If the scale is unidimensional, continuous, developmental, and equal-interval, it can sustain the application of calculus to growth functions expressed with the scale. However, estimates of growth velocity and acceleration have limited generalizability and interpretability unless the scale also possesses general objectivity. Although I could estimate velocity and acceleration in all three examples, the interpretation of these features attained much greater generality (over panels, across time, in different policy contexts) when I used the Lexile scale. I regard this also as prima facie evidence for the general objectivity of the Lexile scale.

Lastly, in retrospect, it is striking how much actually seems to go right in the 15-panel reading growth example. I am rarely surprised when something goes wrong in a complex educational accountability system involving hundreds of thousands of individuals, millions of data points, and hundreds of millions of calculations. On the other hand, I am surprised when everything (or nearly everything) appears to go right. Can such coherence be possible without generally objective measurement?

In consideration of the 15-panel reading growth example, I propose that the ability to model individual growth in the manner demonstrated here (relating time to an attribute that develops within the individual) provides a powerful test of the quantity hypothesis. Furthermore, sustaining such model-based interpretations over multiple instruments and populations is a test of the general objectivity hypothesis.

There are some additional implications of the reading example. Perhaps most important is that it is empirically possible to investigate academic growth as a long-term developmental phenomenon. Not only can the magnitude of student achievement be characterized over time, it is possible to quantify the velocity of growth and the acceleration of growth for individuals and for groups of individuals. Furthermore, quantifying growth over the long term informs about growth over the short term. It does not work the other way around—a short-term approach to growth cannot inform adequately about the long-term growth process.

The 15-panel example of reading growth and the example of conjoint interpretations of mathematical growth offered in this paper, along with the example of conjoint interpretations of reading growth provided by Smith and Williamson (2016), can serve as exemplars for future investigations of academic growth. Attention to the quality of measurement and the consistent use of the same scale over time are precursors in any studies of academic growth. With this foundation, developmental analyses of growth become possible, growth norms are improved and relationships between features of growth and policy contexts become more apparent. Additionally, with conjoint measurement, student reading growth is easily interpreted in relation to the complexity of texts that students read; and student mathematics growth is easily interpreted in relation to the difficulty of mathematics skills and concepts. Additionally, the analysis of individual growth curves is the best way to describe growth and must be supported by longitudinal research designs and data collections. The study of academic growth can closely parallel the study of growth in the physical sciences if these research methods are used.
I hope the results featured in this paper can inspire a vision for improvements that might be accomplished in future studies of academic growth. One aspect needing further attention is our ability to measure educational constructs over longer portions of the life-span. Currently, the measurement of reading ability and mathematics ability in the United States is predominantly conducted in Grades 3–11, a relatively small portion of the entire developmental continuum. Our understanding of reading growth and mathematics growth would be greatly enhanced if we could extend our measurement to span Pre-K through postsecondary education, and ultimately the entire life course. Similarly, it is important to maintain our measurement systems and data collections over longer time frames so that successive panels of individuals may be followed. In conjunction, attention must be given to the adoption of standard units of measurement for each educational construct of interest. A consistent commitment to constructing generally objective scales and to expanding measurement within person and across time could establish a generational view of academic growth over the next 100 years and more.

In a sense, educators can now gaze across a new frontier similar to the one that was crossed in the study of physical growth when the meter was introduced in 1795. By designing, constructing, and maintaining generally objective scales for the measurement of life-span, developmental academic constructs, future studies of academic growth can begin to take their place alongside the best studies of physical growth that have been conducted over the last 200 years. When we take the long-term view, measuring academic growth requires the best measurement scales we can construct because the learners of the world, when their academic growth is measured, deserve nothing less.

Funding
The author received no direct funding for this research.

Author details
Gary L. Williamson1
E-mails: gwilliamson@lexile.com, glwk22@att.net

1 Department of Research and Development, MetaMetrics, 1000 Park Forty Plaza Drive, Suite 120, Durham, NC 27713, USA.

Citation information
Cite this article as: Exploring reading and mathematics growth through psychometric innovations applied to longitudinal data. Gary L. Williamson, Cogent Education (2018), 5: 1464424.

Notes
1. Although this paper focuses on reading growth and mathematics growth as two particular types of academic growth, the principles and methods applied can be tested with other academic constructs (e.g. writing) as well.

2. The fourth edition mathematics tests were not vertically scaled. However, they are not used for the mathematics growth analyses reported here.

3. Three reading panels (i.e. the panels based on students who were in Grade 8 in spring 2010, 2011, and 2012, respectively) incorporated additional Lexile measures collected after Grade 8. These additional waves of data were used in the estimation of the average growth curve. However, only the portion of the growth curve spanning Grades 3–8 is shown for the example in the paper.

4. In fact, several functional forms fit the NC data well within the empirical time frame. For example, quadratic, cubic, negative exponential, and time (square-root) transformed growth models produced strikingly similar estimates of reading ability at the aggregate level. Ultimately, the quadratic growth model was chosen for explication in this study because of its relatively good fit with both reading and mathematics constructs, its parsimony and its heuristic advantages.

5. Curvature connotes, but does not equal, acceleration. The acceleration rate is calculated by multiplying the curvature parameter estimate by 2.

Cover image
Source: MetaMetrics.

References
Bazemore, M., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade & P. B. Baltes (Eds.), Longitudinal research in the study of behavior and development (pp. 1–39). New York, NY: Academic Press. Bazemore, M., Kramer, L. M. B., Gallagher, M. P., Englehart, T. P., & Brown, R. A. (2008). The North Carolina mathematics tests (3rd ed.). Raleigh, NC: Department of Public Instruction. Bazemore, M., & Van Dyk, P. B. (2004). North Carolina reading comprehension tests: Technical report (2nd ed., Citable Draft). Raleigh, NC: Department of Public Instruction. Bazemore, M., Van Dyk, P. B., Kramer, L. M. B., Yelton, A., & Brown, R. A. (2004). The North Carolina mathematics tests: Technical report (2nd ed.). Raleigh, NC: Department of Public Instruction.

Bock, R. D. (Ed.). (1989). Multilevel analysis of educational data. San Diego, CA: Academic Press.

Bond, T. G., & Fox, C. M. (2001). Applying the Rasch Model: Fundamental measurement in the human sciences. Mahwah, NJ: Lawrence Erlbaum Associates.

Briggs, D. C. (2013). Measuring growth with vertical scales. .Journal of Educational Measurement, 50(2), 204–226. https://doi.org/10.1111/jedm.2013.50.issue-2

Bryk, A. S., & Raudenbush, S. W. (1989). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. In R. D. Bock (Ed.), Multilevel analysis of educational data (pp. 159–204). San Diego, CA: Academic Press.

Burdick, H., & Sterne, A. J. (1996). Theoretical prediction of test items. Rasch Measurement Transactions, 10(1), 475.

Castellano, K. E., & Ho, A. D. (2013). A practitioner’s guide to growth models. Washington, DC: Council of Chief State School Officers (CCSSO).

Chall, J. S. (1996). Stages of reading development (2nd ed.). Fort Worth, TX: Harcourt Brace.
Chang, H. (2004). Inventing temperature: Measurement and scientific progress. New York, NY: Oxford University Press. https://doi.org/10.1093/01915171276.001.0001

Domigne, B. (2014). Evaluating the equal-interval hypothesis with test score scales. Psychometrika, 79(1), 1–19. doi:10.1007/s11336-013-9342-4

Fisher, Jr., W. P. (2010). The standard model in the history of the natural sciences, econometrics, and the social sciences. Journal of Physics: Conference Series, 238, 012016. doi:10.1088/1742-6596/238/1/012016

Goldstein, H. (1979). The design and analysis of longitudinal studies. New York, NY: Academic Press.

Goldstein, H. (1987). Multilevel models in educational and social research. New York, NY: Oxford University Press.

Guire, K. E., & Kowalski, C. J. (1979). Mathematical description and representation of developmental change functions on the intra- and interindividual levels. In J. R. Hauspia & P. B. Boltes (Eds.), Longitudinal research in the study of behavior and development (pp. 89–110). New York, NY: Academic Press.

Hauspia, R. C., & Molinari, L. (2010). Parametric models for postnatal growth. In R. C. Hauspia, N. Cameron, & L. Molinari (Eds.), Methods in human growth research (pp. 205–233). New York, NY: Cambridge University Press.

Jöreskog, K. G. (1979). Statistical estimation of structural models in longitudinal-developmental investigations. In J. R. Nesselroade & P. B. Boltes (Eds.), Longitudinal research in the study of behavior and development (pp. 303–351). New York, NY: Academic Press.

Karabatsos, G. (2001). The Rasch Model, additive conjoint measurement, and new models of probabilistic measurement theory. Journal of Applied Measurement, 2(4), 389–423.

Keats, J. A. (1983). Ability measures and theories of cognitive development. In H. Wainer and S. Messick (Eds.), Principles of modern psychological measurement: A festschrift for Frederick M. Lord (pp. 81–101). Hillsdale, NJ: Lawrence Erlbaum.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 155–186). Westport, CT: American Council on Education and Praeger Publishers.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 531–578). Westport, CT: American Council on Education and Praeger Publishers.

Loken, E. (2010). Person-specific analysis of the dynamics of weight change. In P. C. M. Molenaar & K. M. Newell (Eds.), Individual pathways of change: Statistical models for analyzing learning and development (pp. 9–22). Washington, DC: American Psychological Association. https://doi.org/10.1037/12140-000

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. Journal of Mathematical Psychology, 1, 1–27. https://doi.org/10.1016/0022-2496(64)90015-X

Mebello, K., Zhu, M., Karkee, T., & Lung, H. (2016a). The North Carolina testing program technical report, 2012–2015: English language arts/reading assessments (ELA) end-of-grade 3–8 and end-of-course English II. Raleigh, NC: Department of Public Instruction.

Mebello, K., Zhu, M., Karkee, T., & Lung, H. (2016b). The North Carolina testing program technical report, 2012–2015: Mathematics assessments end-of-grade 3–8 and end-of-course Math I. Raleigh, NC: Department of Public Instruction.

Merrell, M. (1931). The relationship of individual growth to average growth. Human Biology, 3, 37–70.

MetaMetrics. (2011). The Quantile Framework® for Mathematics technical guide. Durham, NC: Author.

Molenaar, P. C. M., & Newell, K. M. (Eds.). (2010). Individual pathways of change: Statistical models for analyzing learning and development. Washington, DC: American Psychological Association.

Molinar, J. J., & Bressler, T. (2010). The human growth curve: Distance, velocity and acceleration. In R. C. Hauspia, N. Camerson, & L. Molinari (Eds.), Methods in human growth research (pp. 27–54). New York, NY: Cambridge University Press.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). Common core state standards. Washington, DC: Author. Retrieved from http://www.corestandards.org/

North Carolina Department of Public Instruction. (2009). North Carolina reading comprehension tests: Technical report (3rd ed.). Raleigh, NC: Author. Retrieved from http://www.ncpublicschools.org/docs/accountability/testing/reports/ncreadingtechman3.pdf

North Carolina Department of Public Instruction. (2016). North Carolina testing program test development process: Item, selection and form development. Retrieved from http://www.ncpublicschools.org/docs/accountability/testing/technote/testdevprocess16.pdf

Olsen, L. W. (2003). The Rasch model and its applications: Methodology in the study of behavior and development (pp. 89–110). New York, NY: Cambridge University Press.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedogogiske Institut. (Reprint 1980). Chicago, IL: University of Chicago Press.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for a general and valid definition of objectivity. In M. Blegvad (Ed.), The Danish yearbook of philosophy, (pp. 58–94). Copenhagen: Munksgaard.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. American Sociological Review, 15(3), 351–357. https://doi.org/10.2307/2087176

Rogosa, D. R., Brandt, D., & Zimowski, M. (1983). A growth curve approach to the measurement of change. Psychological Bulletin, 92, 726–748. https://doi.org/10.1037/0033-2909.92.3.726

Sanford, E. E. (1996). North Carolina end-of-grade tests technical report # 7: Reading comprehension, mathematics. Raleigh, NC: Department of Public Instruction.

Sanford-Moore, E. E., Williamson, G. L., Bickel, L., Koons, H., Baker, R. F., & Price, R. (2014). A quantitative task continuum for K-12 mathematics (MetaMetrics Research Brief). Durham, NC: MetaMetrics.

SAS Institute, Inc. (2016). North Carolina Department of Public Instruction technical documentation for 2016 EVAAS analyses. Cary, NC: Author.

Schulte, A. C., Stevens, J. J., Elliott, S. N., Tindal, G., & Nese, J. F. (2016, January 25). Achievement gaps for students with disabilities: Stable, widening, or narrowing on a state-wide reading comprehension test? Journal of Educational Psychology. Advance online publication. doi:10.1037/edu0000107

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. Studies in History and Philosophy of Science Part A, 42(4), 509–524. https://doi.org/10.1016/j.shpsa.2011.07.001

Singer, J. D., & Willett, J. B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence. New York, NY: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001
Appendix 1

The unconditional quadratic growth model

A multilevel model is posited in successive hierarchical levels (hence the name multilevel or hierarchical linear model). With a growth analysis, the first-level model expresses the individual growth curve (i.e., the change within person over time) as a function of individual growth parameters and time. For example, in my growth analyses I hypothesized that the longitudinal student ability measures for individual i were a quadratic function of time, T, as follows:

\[ Y_{ti} = \beta_0 + \beta_1 T_{ti} + \beta_2 T_{ti}^2 + e_{ti} \]  

(A1)
where

- \( Y_{ti} \) is the person ability (e.g. either the Lexile measure of reading ability or the Quantile measure of mathematics ability) as manifested through the reading or mathematics achievement on occasion \( t = 1, \ldots, T_i \) for individual \( i = 1, \ldots, N \);
- \( \pi_0i \) is the intercept of the growth curve for individual \( i \), or the expected value of \( Y_{ti} \) when \( T_{ti} = 0 \);
- \( \pi_1i \) represents the true instantaneous rate of growth for individual \( i \) when \( T_{ti} = 0 \) (see the discussion of Equation A7);
- \( \pi_2i \) represents the curvature, or acceleration, in the growth curve for individual \( i \) (see the discussion of Equation A8);
- \( T_{ti} \) is the value of the time metric on occasion \( t \) for individual \( i \); and
- \( e_{ti} \) is a random disturbance indicating the deviation of the observed achievement measure, \( Y_{ti} \), from the individual’s true growth curve. It was assumed that \( e_{ti} \sim N(0, \sigma^2_e) \) for each \( t \) and \( i \) (Other assumptions about the within-person error structure are possible. However, independence is commonly assumed (Thissen & Bock, 1990) and is often the first error structure investigated).

For the hypothesized model, although time is continuous, longitudinal data were gathered on specific occasions approximately equally spaced in time. That is, students were tested within the last 2–3 weeks of schooling in Grades 3–8. Consequently, the time scale was coded in terms of grade and was centered on the first occasion of measurement (Grade 3). Thus, time was coded as \( T_i = (\text{GRADE}_i - 3) \) for analysis. Because the longitudinal data span Grades 3–8, \( \text{GRADE}_i \) also ranges from 3 to 8. Consequently, \( T_i \) ranges from 0 to 5, with \( T_i = 0 \) corresponding to the end of third grade. With these coding conventions, \( \pi_0i \) represents the true status of individual \( i \) at the end of third grade (In the past, some researchers have recommended the use of orthogonal polynomial coding for quadratic growth models (e.g. Jöreskog, 1979; Wishart, 1938) because of certain computational advantages. However, other researchers have used ordinary quadratic coefficients because they yield more substantively meaningful interpretations for the model parameters (Goldstein, 1979; Raudenbush & Bryk, 2002; Singer & Willett, 2003). The latter approach is adopted here for the same reason).

The choice of centering point for the time scale is methodologically arbitrary, except that it justifies a substantive interpretation of the individual intercept parameter. Other points in time could have been chosen—e.g. the final occasion of measurement or a point near the center of the time frame. I chose the initial occasion of measurement as the centering point for several reasons. First, an individual’s status at the beginning of the data collection window is of substantive interest. Because students first take the North Carolina EOG Test of Reading Comprehension at the end of Grade 3, centering the time scale at the end of third grade allows us to interpret \( \pi_0i \) as the true status of individual \( i \) when he or she first entered the data collection. Second, even if one wishes to explore other centering points for the model, the initial occasion of measurement is a good place to begin (Singer & Willett, 2003). Finally, the end of Grade 3 is a substantively meaningful time in the context of the North Carolina testing and accountability program—it is the earliest grade in which students are held to individual student accountability standards.

In a multilevel model, the individual growth parameters are specified in the first level and then in the second level are permitted to vary across individuals to represent how individual growth trajectories differ among (or between) persons. For the growth analyses presented in this paper, the individual parameters of Equation (A1) were expressed as the sum of a grand mean in the population and a random disturbance for individual \( i \). Hence,

\[
\begin{align*}
\pi_{0i} &= \gamma_{00} + r_{0i} \\
\pi_{1i} &= \gamma_{10} + r_{1i} \\
\pi_{2i} &= \gamma_{20} + r_{2i}
\end{align*}
\]
where the gamma-terms, respectively, represent the average intercept, slope, and curvature parameters in the population of students; and the associated \( r \)-terms are errors to capture the degree to which features of individual growth curves deviate from the corresponding features of the grand population growth curve. The variances of these error terms express the variation of the individual growth parameters in the population. It is assumed that \( r \) is distributed as:

\[
\begin{pmatrix}
    r_0 \\
    r_1 \\
    r_2
\end{pmatrix} 
\sim N\left(\begin{pmatrix}
    0 \\
    0 \\
    0
\end{pmatrix}, \begin{pmatrix}
    \tau_{00} & \tau_{01} & \tau_{02} \\
    \tau_{10} & \tau_{11} & \tau_{12} \\
    \tau_{20} & \tau_{21} & \tau_{22}
\end{pmatrix}\right) \tag{A3}
\]

where the \( \tau \)-matrix is symmetric. It is further assumed that the error terms in the level-2 model are independent of the error terms (\( e_i \)) in the level-1 model.

By substituting Equation (A2) into Equation (A1), a single composite model equation is derived that expresses the population growth curve, the deviations of the individual growth curves from the population trajectory, and the within-person variation of scores in a single expression.

\[
Y_{ti} = \gamma_{00} + \gamma_{10}T_{ti} + \gamma_{20}T_{ti}^2 + (r_{0i} + r_{1i}T_{ti} + r_{2i}T_{ti}^2 + e_{ti}) \tag{A4}
\]

Table A1 provides estimates of the fixed and random effects for the unconditional quadratic growth model fit to the North Carolina reading and mathematics longitudinal data.

**Features of the growth trajectory**

Once a particular hypothetical growth model has been specified, it has important consequences for the way that the growth trajectory is summarized and for what its important features are hypothesized to be. Magnitude (status), velocity, and acceleration are features of growth that receive central attention in studies of physical growth (Molinari & Gasser, 2010) and merit more attention in studies of academic growth. Every parametric model has a unique parametrization. In that light, it is imperative to evaluate a hypothesized growth function and its derivatives in order to understand how that particular mathematical model summarizes the magnitude, velocity, and acceleration of growth. To illustrate, the quadratic growth model is examined in this regard.

**Achievement status**

Based on the composite model in Equation (A4), true expected achievement status for individual \( i \) at any given time is provided by the structural (i.e. non-stochastic) part of the model. The individual quadratic growth curves for individuals \( i = 1, ..., N \) are:

\[
E(Y_{ti}|T_{ti}) = \gamma_{00} + \gamma_{10}T_{ti} + \gamma_{20}T_{ti}^2 \tag{A5}
\]

Notice that the functional form is the same for all individuals. What distinguishes one individual growth curve from another is the values of its parameters, \( \gamma_{00}, \gamma_{10}, \) and \( \gamma_{20} \). Also note that the values of the time variable on each occasion are allowed to differ for each individual, although this was not necessary for the complete-case analyses of this paper. There is also a subtle distinction between the lower case \( t \) used as a subscript and the capital \( T_t \). The lower case \( t \) indexes the occasion of measurement (enumerated by the positive integers) whereas the capital \( T_t \) denotes the value of the time variable on any specific occasion, \( t \). So for example, in the North Carolina data, \( t = 1, 2, 3, 4, 5, 6; \) yet \( T_{ti} = 0, 1, 2, 3, 4, 5 \) (because the time scale was centered at the end of third grade, as explained earlier).

Similarly, the true average growth curve in the population has the same functional form and is given by the structural part of the composite model (Equation A4).

\[
E(Y_{t}|T_{t}) = \gamma_{00} + \gamma_{10}T_{t} + \gamma_{20}T_{t}^2 \tag{A6}
\]
Once the multilevel model has been fit, the estimated values of the parameters in Equations (A5) and (A6) can be used to calculate either the estimated status for an individual or the estimated average status for the population, respectively, at any particular time. Just replace the parameters in the equations with their respective estimates; then substitute the value for time on the desired occasion into the formula for the individual growth curve (Equation A5) or the population average growth curve (Equation A6), respectively. The latter strategy was used to produce the growth curve plots presented in the paper.

**Velocity of growth**

The velocity of growth is given by the first derivative of the growth curve. For instance, the first derivative of the individual growth curve (Equation A5) with respect to time provides the instantaneous velocity of individual growth, which is expressed as a function of time, as follows:

\[
\frac{dE(Y_{i}|T_{i})}{dT} = \pi_{1i} + 2\pi_{2i} T_{i} \tag{A7}
\]

This shows, for the quadratic growth model, that the instantaneous velocity is not constant but is a linear function of the time of measurement. By substituting \( T_{i} = 0 \) into Equation (A7), it is observed that the parameter \( \pi_{1i} \) is the true instantaneous velocity for person \( i \) at the end of Grade 3. The instantaneous velocity on other occasions may be calculated by substituting the appropriate value of \( T_{i} \) into Equation (A7). Similarly, the first derivative of Equation (A6) gives the average instantaneous velocity of growth in the population. Alternatively, simply replace the \( \pi \)-terms in Equation (A7) by the corresponding \( \gamma \)-terms.

The fact that the velocity curve is a linear function of time is specific to the quadratic growth model. It is critical to examine the first derivative for any hypothesized growth model to confirm the nature of velocity in the context of that particular model for growth.

**Acceleration of growth**

The second derivative of the growth curve provides the acceleration rate. For the quadratic growth model, the acceleration rate for individual \( i \) is given by:

\[
\frac{d^2E(Y_{i}|T_{i})}{dT^2} = 2\pi_{2i} \tag{A8}
\]

Thus the quadratic growth curve has a constant rate of acceleration equal to twice the value of the curvature parameter \( \pi_{2} \). When the value is negative, it connotes deceleration. The acceleration of the population average growth curve is analogous—simply, substitute \( \gamma_{20} \) into Equation (A8).

In summary, the \( \pi_{0i} \) represent the intercepts of the individual student growth curves; the \( \pi_{1i} \) represent the instantaneous individual rates of growth at the end of Grade 3; and the \( \pi_{2i} \) represent the individual curvature parameters (which give individual constant acceleration rates, when multiplied by 2). The corresponding level-2 parameters, \( \gamma_{00}, \gamma_{10}, \gamma_{10} \) and \( \gamma_{20} \) represent the respective features of the average growth curve in the population. These parameters capture the salient developmental features of growth that are of interest. Stated more succinctly, they represent magnitude, velocity, and acceleration, respectively. These interpretations are corroborated by Raudenbush and Bryk (2002) and by Singer and Willett (2003).

The fixed effects estimates of the population growth parameters are provided in Table A1, along with estimates of the random effects for the fitted growth models.
### Table A1: Fixed and random effects parameter estimates for North Carolina growth curves by subject/scale and panel

| Panel | n   | $\gamma_{00}$ | $\gamma_{10}$ | $\gamma_{20}$ | $\tau_{00}$ | $\tau_{01}$ | $\tau_{02}$ | $\tau_{11}$ | $\tau_{12}$ | $\tau_{22}$ | $\sigma_e^2$ |
|-------|-----|----------------|---------------|---------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| **Mathematics scale score (edition 1)** |     |                |               |               |             |             |             |             |             |             |             |
| 1995-2000 | 48,887 | 142.6 | 8.4 | -0.275 | 95.2 | -9.6 | 1.7 | 5.2 | -0.7 | 0.1 | 15.3 |
| 1996-2001 | 53,576 | 143.2 | 9.3 | -0.524 | 88.7 | -6.7 | 1.1 | 4.5 | -0.6 | 0.1 | 16.3 |
| 1997-2002 | 63,006 | 144.7 | 9.2 | -0.538 | 86.4 | -6.0 | 0.9 | 4.5 | -0.7 | 0.1 | 15.9 |
| 1998-2003 | 65,760 | 144.5 | 9.8 | -0.605 | 83.0 | -4.6 | 0.6 | 4.6 | -0.6 | 0.1 | 14.8 |
| 1999-2004 | 68,406 | 144.9 | 9.9 | -0.617 | 82.0 | -7.5 | 1.2 | 4.1 | -0.5 | 0.1 | 14.8 |
| **Quantile measure** |     |                |               |               |             |             |             |             |             |             |             |
| 2000-2005 | 74,076 | 541.9 | 139.3 | -10.914 | 42,839.0 | -5,757.5 | 724.6 | 2,619.1 | -356.2 | 54.8 | 5,849.1 |
| 2001-2006 | 74,658 | 543.1 | 152.0 | -13.425 | 43,194.0 | -5,994.6 | 759.4 | 2,085.5 | -270.1 | 41.7 | 5,929.6 |
| 2002-2007 | 75,601 | 583.2 | 136.9 | -11.271 | 39,546.0 | -5,171.4 | 721.6 | 2,085.9 | -279.0 | 43.8 | 6,155.1 |
| 2003-2008 | 77,055 | 628.2 | 101.5 | -5.144 | 42,715.0 | -1,137.6 | 721.6 | 2,085.9 | -279.0 | 43.8 | 6,155.1 |
| 2004-2009 | 76,872 | 623.2 | 114.1 | -6.337 | 27,330.0 | -2,277.5 | 721.6 | 2,085.9 | -279.0 | 43.8 | 6,155.1 |
| **Lexile measure** |     |                |               |               |             |             |             |             |             |             |             |
| 1995-2000 | 48,790 | 625.8 | 127.4 | -6.273 | 48,923.0 | -4,134.6 | 238.6 | 2,306.5 | -304.1 | 45.6 | 7,560.0 |
| 1996-2001 | 53,375 | 633.6 | 139.5 | -8.645 | 44,682.0 | -2,061.5 | -101.4 | 1,554.1 | -187.0 | 26.6 | 8,030.8 |
| 1997-2002 | 62,692 | 653.9 | 126.9 | -6.740 | 40,046.1 | -425.6 | -680.6 | 1,646.9 | -214.4 | 33.9 | 8,306.1 |
| 1998-2003 | 65,395 | 687.5 | 115.0 | -5.423 | 40,046.1 | -1,137.6 | -268.9 | 1,503.3 | -176.4 | 23.6 | 8,049.5 |
| 1999-2004 | 67,908 | 700.0 | 118.7 | -6.093 | 42,157.3 | -2,160.0 | -328.0 | 1,492.7 | -153.6 | 19.0 | 8,371.1 |
| 2000-2005 | 73,475 | 697.5 | 118.3 | -6.049 | 43,640.0 | -5,669.2 | 374.8 | 1,588.5 | -133.8 | 8.2 | 7,935.6 |
| 2001-2006 | 74,224 | 704.8 | 126.2 | -8.500 | 37,182.0 | -2,887.1 | 0.0 | 374.1 | 0.0 | 0.0 | 8,225.1 |
| 2002-2007 | 75,236 | 730.1 | 119.7 | -8.152 | 33,074.0 | -2,459.7 | 0.0 | 346.3 | 0.0 | 0.0 | 7,689.0 |
| 2003-2008 | 77,281 | 740.9 | 111.4 | -7.003 | 32,365.0 | -2,257.7 | 0.0 | 305.0 | 0.0 | 0.0 | 7,540.4 |
| 2004-2005 | 76,523 | 748.8 | 102.3 | -5.183 | 31,696.0 | -1,970.6 | 0.0 | 283.2 | 0.0 | 0.0 | 7,486.6 |
| 2005-2013 | 54,657 | 757.1 | 136.6 | -7.824 | 23,892.1 | -518.3 | 102.0 | 958.1 | -136.1 | 23.2 | 8,430.8 |
| 2006-2014 | 51,221 | 763.8 | 138.6 | -8.26 | 26,135.5 | -62.0 | 611.1 | 19.6 | -81.6 | 13.9 | 9,189.8 |
| 2007-2014 | 66,541 | 752.1 | 143.3 | -8.793 | 36,384.1 | -1,237.7 | 91.5 | 547.4 | -65.0 | 12.0 | 8,891.4 |
| 2008-2014 | 81,688 | 725.5 | 168.7 | -14.408 | 66,733.8 | -13,564.9 | 1,967.0 | 4,483.2 | -680.6 | 119.0 | 9,029.1 |
| 2009-2014 | 83,112 | 747.1 | 154.7 | -12.664 | 62,479.9 | -11,480.6 | 1,759.6 | 3,467.0 | -512.8 | 89.9 | 10,016.8 |

Note: Analyses were run with SAS PROC MIXED, full maximum likelihood estimation (METHOD = ML) and unstructured covariance matrix (TYPE = UN). Unless noted otherwise, all parameter estimates were statistically significant at $p < 0.0001$.

*p not available. °variance fixed to zero. °$p = 0.1677$. °$p = 0.0006$. 
