PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2015, 52(2), 98-118

Pontificia Universidad Católica de Chile
Facultad de Educación

Measuring Academic Growth Contextualizes Text Complexity

La medición del crecimiento académico contextualiza la complejidad del texto

Gary L. Williamson

MetaMetrics®, North Carolina, USA

Abstract

Optimally measuring academic growth requires a scale that is unidimensional, continuous, equal-interval, developmental, and invariant with respect to location and unit size. These characteristics can be attained by combining Rasch measurement with an operationalized reading construct theory and anchoring the resulting scale at two points on a text-complexity continuum. Hence, persons and texts are brought onto a common developmental scale. Coincidentally, an educational policy initiative in the United States of America (USA) recommended increasing students' exposure to complex texts. The text-exposure standard can be productively examined in light of historical reading growth when both are measured with a common theory-based scale. Parametric modeling of alternative growth curves motivates a conversation about how students may attain a college and career readiness goal.

Keywords: growth, reading, text complexity, measurement

Post to:
Gary L. Williamson
MetaMetrics®, North Carolina, USA.
1000 Park Forty Plaza Drive, Suite 120
Durham, NC 27713, USA.
Email: gwilliamson@lexile.com

ISSN: 0719-0409 DDI: 203.262, Santiago, Chile
doi: 10.7764/PEL.52.2.2015.6

Resumen

Para una medición óptima del crecimiento académico se requiere una escala unidimensional, continua, a iguales intervalos, de desarrollo e invariable con respecto a la ubicación y el tamaño de la unidad. Estas características pueden obtenerse combinando la medición Rasch con una teoría de constructos de lectura operacionalizada y el anclaje de la escala resultante en dos puntos en un continuo de complejidad de texto. De esta manera, personas y textos se juntan en una escala de desarrollo común. Así, una iniciativa de política educativa de Estados Unidos de América (EE.UU.) recomendó incrementar la exposición de los estudiantes a textos complejos. El estándar de exposición a textos puede examinarse en forma productiva a la luz del crecimiento histórico de la lectura si ambos se miden con una escala común de base teórica. El modelado paramétrico de curvas alternativas de crecimiento promueve un debate sobre cómo los estudiantes pueden alcanzar el objetivo de estar preparados para una universidad y una carrera.

Palabras clave: crecimiento, lectura, complejidad de texto, medición

In spring 2014, I made a presentation to the Metrology and Outcomes section of the 17th International Objective Measurement Workshop (IOMW). During that session, I described several related threads of work that I think portray the power of measurement to inform two recent issues of educational practice: growth and text complexity. In this paper, I summarize those ideas and describe supporting examples that illustrate the practical potential for applying Rasch measurement to educational problems. However, it may be useful to first provide some context to motivate the research and to define some terms that may be unfamiliar to the uninitiated reader.

Context for the reader

Reading research and pedagogy

In a retrospective paper about reading research and reading pedagogy in the United States of America (USA), Pearson (2004) argued that reading instruction and reading research had for decades been subject to repeated «pendulum swings» between alternate pedagogical practices (whole language vs. skill-oriented instruction) and corresponding alternate research paradigms (randomized experiments vs. naturalistic, qualitative investigations). Furthermore, he asserted that reading instruction and reading research had been shaped by competing political forces, which favored one approach or the other. He called this scenario The Reading Wars. In his paper, Pearson called for more balance and observed, «It is our moral and ethical obligation to use the best evidence we can muster for making policy decisions of consequence» (p. 21).

Over the last decade, perhaps some of the balance that Pearson (2004) summoned has found its way into core reading texts and instructional emphases used in the early grades (Personal communication, J. Fitzgerald, March 25, 2015). Nevertheless, perspectives about reading still bifurcate along some lines; the poles now seem to correspond to cognitive (e.g., Kendeou, van den Broek, Helder, & Karlsson, 2014) and sociocultural viewpoints (e.g., Alvermann, Hinchman, Moore, Phelps, & Waff, 2006). Readers can peruse Alvermann, Unrau, and Ruddell (2013) to observe the two ends of the spectrum. Consequently, the measurement of reading achievement is still guided by different views of what reading is, how it is best taught, and how (or whether) reading achievement should be quantified.

Relevant educational policy actions in the USA

During the first half of the 20th century, there was no definitive way to ascertain the average reading ability of students in the USA, because there was no commonly-used educational measurement of reading achievement. Many reading tests were available, but they yielded results that were not comparable. To alleviate the situation, the USA created a National Assessment of Educational Progress (NAEP) during the 1960s to collect data on educational performance. The new assessment program was first administered in 1969; the first reading achievement data were collected during the 1970–71 school year. Today, the NAEP continues to periodically collect reading achievement data. However, the NAEP data collection is designed to produce national or state aggregate summaries of student reading performance; it does not produce scores for individual students.

In the USA, educational governance predominantly resides at the state and local community levels. The historically limited role of the federal government in education increased somewhat when the United States Congress passed the Elementary and Secondary Education Act (ESEA) of 1965. From its beginning, the core purpose of the ESEA was to improve educational opportunities for all students. The law has been periodically updated or reauthorized. The most recent reauthorization of the ESEA was the No Child Left Behind Act of 2001 (No Child Left Behind [NCLB], 2002), which was signed into law by President George W. Bush in January 2002. Since that time, the ESEA has been called NCLB. NCLB requires all public schools to measure student achievement in reading and mathematics during specified grades. The federal requirement to measure and report the results of individual student reading achievement represented a major shift in educational measurement practice in the USA.

The federal requirement to measure student reading achievement did not specify which reading test(s) states should use. States still retained the authority to create their own state-specific reading curricula and to select or create their own specific measures of reading achievement. As a result, the individual states collectively use a plethora of different reading measures and each state sets its own standards for acceptable student performance on its adopted measure of reading achievement. This has gradually resulted in widespread cognitive dissonance because the student performance results derived from state assessments are not comparable; they are also not comparable with the results from NAEP.

In the USA, each state has a chief executive officer, called a governor, who is elected by the people. Since 1908, the National Governors Association (NGA) has served as a bipartisan organization that supports various leadership activities for the nation's governors. In particular, one of its initiatives is the Center for Best Practices, which promotes the identification and dissemination of information to improve educational practice throughout the country. The members of the NGA are the governors of the states, territories and commonwealths of the USA.

Each state also has a chief state school officer, who is in charge of the public schools. In some states, the chief state school officer is an elected official; in other states, the chief state school officer may be appointed, perhaps by the governor or the governor's designee(s). The Council of Chief State School Officers (CCSSO) is a national organization made up of the educational leaders of the states.

Following the enactment of NCLB, public accountability for school performance became a central theme in education. The lack of comparability of state results and the apparent disconnect between state achievement reports and the NAEP caused increasing concerns among the governors and chief state school officers in the USA. Accordingly, the two organizations began a collaboration to improve educational practice and accountability.

The CCSSO and the NGA Center for Best Practices directed their energies toward engaging educators in developing a common framework of consistent student learning goals aimed at enabling students to be prepared for college and careers by the time they graduate high school. Following extensive development, input, and revision, the NGA and CCSSO released the Common Core State Standards (CCSS) in June 2010. States individually reviewed the standards and by the end of 2013, the CCSS were widely adopted. As of June 2014, 43 states, the District of Columbia, four territories, and the Department of Defense Education Activity (DoDEA) have adopted the CCSS and are implementing the standards according to their own timelines.

In the process of implementing the CCSS, states began to realize that they needed to give some attention to the assessment of student performance in relation to the standards. Two consortia of states evolved with slightly different assessment philosophies: the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC). Though their guiding principles vary, the two consortia nevertheless aim to improve assessment practices and enhance the comparability of student performance results.

Even before the so-called reading wars, an important feature of reading instruction and reading research in the USA was a lack of common agreement about what reading is; understandably, a multiplicity of reading measures are available for educational use. Although the CCSS represents a widely-adopted common agreement about student learning goals, and SBAC and PARCC have made progress toward designing tests aligned with the CCSS, there is still no common measure of reading comprehension for individual students in the USA and no common agreement regarding the level of reading achievement students must attain.

Measurement context

A general review of educational measurement is well beyond the scope of this paper. For the most recent encyclopedic coverage of the topic, I encourage the interested reader to consult Brennan (2006), a 779-page volume jointly sponsored by the National Council on Measurement in Education and the American Council on Education. Instead, I will comment on three particular aspects of educational measurement —specific objectivity, general objectivity, and conjoint measurement— that are pertinent to the examples presented later in this paper. Accordingly, I trace (only broadly) the developments that gave rise to these three ideas.

Educational measurement emerged in the USA during the early part of the 20th century. By the middle of the century, existing knowledge about the theory of testing was being canonized as *classical test theory* (e.g., Gullicksen, 1950). This approach continued to be the primary one taught, even as new *item response theory* (IRT) models (e.g., Lord, 1952; Lord & Novick, 1968; Rasch, 1960) began to supplement and expand the field of psychometrics. Gradually, IRT found widespread use (Lord, 1980) and today is considered essential for developing educational tests. Its central advantage over classical test theory is that IRT provides an explicit mathematical model to characterize an individual's response to an item on a test.

Mathematically, the simplest IRT model is the one attributed to Rasch (1960). Increasingly, Rasch (1966, 1977) realized that his mathematical formulation resulted in interesting properties, which he regarded as epistemologically important. For example, when calculating the ratio between the probabilities of correct response (to an item) for two persons, the item parameters cancel from the expression. «Thus the model makes *objective* or *item-free* statements about the relative likelihood that two persons will respond correctly to an item or a set of items, without reference to the items themselves» (Thissen & Wainer, 2001, p. 75, italics in the original). Another consequence of the model is that calculating the log-odds ratio of a person's correct response to an item yields the arithmetic difference between the person's ability and the item's difficulty. Bond and Fox (2001) suggested that the ability to compare persons and items directly, without reference to other items or persons, yields the type of measurement «we have come to expect in the physical sciences» (p. 203). The ability to produce person-free item calibrations and item-free person measures is unique to the Rasch model. This mathematical property of the Rasch model has been called *parameter separation*; Rasch (1977) called it *specific objectivity*.

IRT provided a new framework for developing educational tests by modeling what happens when individuals respond to items, taking into account the characteristics of the items. The Rasch model goes further by ensuring that measurement scales do not depend on the particular individuals or items used to construct them. However, even though IRT models posit a person construct (e.g., reader ability) and a task dimension (e.g., text complexity), IRT models themselves do not provide explicit details about what those constructs are. Furthermore, IRT models do not use an absolute frame of reference to calibrate their scales. Thus, one reading test constructed with IRT is not directly comparable to another reading test also constructed with IRT.

Is it possible to make an absolute scale for reading ability —one that can be universally interpreted in terms of the construct itself? How does one advance from specific objectivity to general objectivity in educational measurement? It would require a clear operational definition of the construct continuum and a commitment to use a consistent frame of reference for the construct measurement scale.

There was a breakthrough in the 1980s when two measurement companies independently attempted to construct generally objective measurements of reading ability by combining the Rasch measurement model with an explicit operational specification of the reading task continuum. As with other IRT models, the Rasch model makes it possible to place both persons and tasks (items) on a common scale, but the location of the scale is indeterminate. The key innovation involved two steps that anchored the scale and defined its unit size in terms of a well-specified reading task continuum.

First, the researchers quantified the reading difficulty of texts in terms of their semantic and syntactic features, which are effective proxies for the cognitive demand experienced by individuals when they read. This step essentially creates a quantitative construct specification (or model) for the reading task continuum. Secondly, the researchers demonstrated that the empirical difficulties of specific, well-defined, text-based item types could be nearly perfectly predicted by the complexities of the texts associated with the items. Once students and items were measured via the Rasch model, the construct theory specification was used to calibrate the items to the text difficulty continuum. This effected a direct correspondence between the person-measures and the text-measures in terms of a real-world, text-complexity continuum.

The company now known as Questar Assessment, Inc. was the first to use this type of approach. They developed the Degrees of Reading Power® (DRP®) Program, which reports student reading measures from criterion-referenced tests on a proprietary DRP Scale of Text Complexity, which it uses to measure the reading difficulty of printed material (Questar Assessment, Inc., 2012). Nelson, Perfetti, D. Liben, and M. Liben (2012) described the scale as follows:

> DRP text difficulty is expressed in DRP units on a continuous scale with a theoretical range from 0 to 100. In practice, commonly encountered English text ranges from about 25 to 85 DRP units, with higher values representing more difficult text (p. 11).

Questar defined a DRP «prose comprehension model» based on the application of the Bormuth (1969) readability formula to measure text complexity. Their reference item type was a text-embedded cloze item administered according to a specific protocol. The unit size of the DRP scale was specified in terms of a transformation of the Bormuth text complexity measure, R. Research has shown that the DRP scale places both student reading ability and text complexity on a common well-defined, unidimensional scale that remains invariant over time. Thus, research supports the claim that the DRP tests «are like measures in the natural sciences» (Koslin, Zeno, & Koslin, 1987, p. 171).

At nearly the same time, a second company pursued the same fundamental idea. MetaMetrics® developed The Lexile® Framework for Reading to measure both readers and texts on a common scale. The Lexile Framework also rests fundamentally on an application of the Rasch model. However, again, the Rasch model was expanded to achieve general objectivity by explicitly quantifying the task continuum. The process involved three steps. First, a specific text-based, reference item type was designed to measure reader ability. In order to respond to the item, readers had to read and comprehend a segment of prose text. Second, a *construct specification equation* (Stenner, Smith, & Burdick, 1983) was developed to quantify the text complexity of prose reading material based on features of the textual material. Third, it was shown empirically that the text-complexity measures produced by the specification equation could predict the Rasch-based difficulties of the reference items with great accuracy (Burdick & Stenner, 1996). Consequently, item difficulty and person ability could both be expressed via the Rasch model on a common invariant scale defined by the text-complexity continuum.

In order to define a logical scale unit for the Lexile scale, MetaMetrics chose to explicitly anchor its scale at two points on the text complexity continuum. Based on its anchoring, a Lexile scale unit equals 1/1000 of the difference between the readability of certain specific basal primers and the readability of an online adult encyclopedia (Stenner, H. Burdick, Sanford, & D. S. Burdick, 2007). Note that

anchoring the scale at two well-defined points is sufficient to unambiguously define the scale's unit size. This approach provided a well-defined unit of measurement that retains its absolute size across different applications of measurement. It may be noted that this method is directly analogous to the way the meter was standardized based on the length of the meridian quadrant through Paris (Legendre, 1805). It is also precisely analogous to the way that temperature scales are anchored. Ironically, some years later, this technique was recommended by the outgoing President of the National Council for Measurement in Education as a desirable feature for academic achievement scales and he challenged the measurement community to develop such scales for education (Reckase, 2009), apparently unaware that it had already been done for reading ability.

The Rasch model expresses the probability of correct response to an item as the exponentiated difference between the person's ability and the item's difficulty. Earlier, it was noted that, because of parameter separation, the log-odds of a correct response is the arithmetic difference between the person-ability and item-difficulty. These facts are manifestations of another important property of the Rasch measurement model —namely, that it produces conjoint measurement. Conjoint measurement makes it possible to simultaneously scale two variables that jointly predict an outcome. For example, reader ability and text difficulty jointly predict reading comprehension; thus, both the reader measure and the text difficulty measure can be placed on a common scale through conjoint measurement.

Through conjoint measurement with the Rasch model, both the Lexile Framework and the DRP can be utilized to generate student reading-ability scores that are reported on a text-complexity continuum, giving the scores supplemental interpretability in terms of a real-world reading task continuum. The primary use of both of these systems to date has been matching students with texts of appropriate difficulty. The examples presented later in this paper illustrate that the methodology of anchoring student reading measures to a well-defined continuum of text complexity holds promise for the measurement of student growth in reading.

Measurement technologies used for the examples

The examples featured in this paper rely on multiple threads of previous work completed by a variety of individuals and organizational entities. One category of work consisted of measuring text-complexity for collections of texts, which produced distributions of text-complexity measures for reading materials characteristic of the public schools and postsecondary domains of endeavor in the USA. Another category of work produced longitudinal reading achievement measures for individuals who attended public schools in the state of North Carolina. The third category of work consisted of multilevel growth models applied to the longitudinal data.

About 20 years ago, the Department of Public Instruction in North Carolina made the decision to use the Lexile Framework for Reading as a supplemental scale for its End-of-Grade Tests of Reading Comprehension in Grades 3–8 in order to provide increased instructional utility for its test results (specifically, helping teachers match students with texts of an appropriate difficulty). The examples in this paper illustrate the additional utility of the scale for modeling student reading growth and for making conjoint interpretations of student reading growth relative to text-complexity requirements. Consequently, the genesis of these data sources is of contextual interest and is briefly outlined next.

Text measurement. The Lexile Framework for Reading bases its text-complexity measures on the semantic and syntactic features of texts. Specifically, text complexity is quantified as a linear combination of two variables —word frequency and sentence length (Stenner, H. Burdick, Sanford, & D. S. Burdick, 2007). Lexile text-complexity measures are calculated for professionally edited prose text. Texts that do not have conventional punctuation (e.g., poetry, song lyrics, recipes) are not measured using the Framework.

Texts must be prepared for measurement according to specific conventions. For example, before analyzing a text, all headings and section titles are removed (because they are not complete sentences). Texts are digitized to facilitate the analysis. The production of Lexile text measures is automated in a software program called the Lexile® Analyzer, which is available free to educators or can be licensed for professional use.

Over the years, numerous studies have been conducted to measure the reading difficulty, or text complexity, of textbooks that are used in the public schools in the USA. It is beyond the scope of this paper to provide a thorough review or critique of that literature. Instead, I briefly describe three recent papers that provide relevant background for the specific text collections used for the examples in this paper.

First, Williamson, Koons, Sandvik, and Sanford-Moore (2012) described a text-complexity continuum for Grades 1–12. The authors identified texts by consulting textbook adoption lists from seven states (Florida, Georgia, Indiana, North Carolina, Oregon, Texas and Virginia) where statewide textbook adoptions were common practice. They focused on student editions of textbooks, organized into six content categories —Health, Language Arts, Literature, Mathematics, Science and Social Studies— to provide diverse curricular coverage. Textbooks that appeared on adoption lists in multiple states were selected for the study, on the presumption that such texts were likely to be used by a large number of students. When textbooks occurred in a series, all books in the series were included in the study. Texts designated for a single grade were selected for the analyses so as to unequivocally characterize within-grade text complexity distributions. A total of 487 textbooks in Grades 1 through 12 were included in the study. The authors found that median text complexity increased with grade, nearly monotonically and that text complexity varied within grade.

Williamson (2008) considered reading materials found in various contexts of postsecondary life (e.g., university, community college, workplace, military, and citizenship). He presented a continuum of text difficulty for the postsecondary domains of endeavor and quantified the text-complexity gaps between high school texts (used in Grades 11 and 12) and texts used in the postsecondary world. Notably, he found a 265L gap between the medians of high school texts and university texts.

Stenner, Sanford-Moore, and Williamson (2012) extended the scientific community's knowledge of postsecondary reading materials. Their work expanded upon Williamson (2008) by enhancing the text collections for postsecondary educational institutions and adding two additional categories of texts. Focusing on three southern states (Georgia, Tennessee and Texas), they added additional university and community college texts; they also included technical college texts. In addition, the authors included text samples drawn from international English-language newspapers and featured articles selected from Wikipedia. Pooling all postsecondary text sources (n = 2,990), they reported a median postsecondary text complexity measure of 1300L and an interquartile range extending from 1200L to 1380L.

Student measures of reading ability. Lexile measures of student reading ability are derived from reading assessments that have been psychometrically linked to the Lexile scale. Given a target test, which is to be linked, the process involves creating a Lexile Linking Test (LLT) that is rigorously designed to: (a) measure the same construct as the target test, (b) have the same reliability as the target test, and (c) have the same test specifications in all respects except one —the LLT consists only of the standard item type associated with the Lexile Framework. The LLT and the target test are both administered to a sample of students and a symmetric linking function is derived. Thus, scores from the target test are expressed on the Lexile scale. Lexile linking studies satisfy at least three of the five requirements specified by Holland and Dorans (2006) to qualify as an equating, which is the strongest kind of psychometric link.

Some of the examples in this paper depict student reading growth and use results from previous research described by Williamson (2014). His reading data derived from administrations of the North Carolina End-of-Grade (NC EOG) Reading Comprehension Tests (Sanford, 1996), which the state linked to the Lexile Framework for Reading. The characteristics of the student data are briefly summarized here for the interested reader's convenience.

The NC EOG Tests of Reading Comprehension (1st Edition) were multiple choice tests administered in paper-and-pencil format. They were designed to satisfy state and federal statutory requirements related to the implementation of the state's curriculum and accountability program. To that end, EOG tests were closely aligned with the state's curriculum during development. They 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 growth. EOG reading tests were designed for administration at the end of each grade in Grades 3–8, during the last three weeks of school.

Item development and test construction were based on IRT using the three-parameter logistic (3PL) model. Internal consistency of the reading comprehension tests (coefficient alpha) ranged from 0.90 to 0.94. Test-retest reliability for the reading comprehension test was reported to be 0.86; criterion and construct validity were also supported by several studies reported by Sanford (1996).

The NC EOG reading tests have been periodically modified to reflect curricular revisions. Accordingly, in 2003, a second edition was made operational, and in 2008, a third edition was made operational. (The current edition, based on the Common Core State Standards, was implemented in 2012–13, but is not reflected in the data examples used for this paper). Bazemore and Van Dyk (2004) described the second edition reading tests, while the third edition was described by the North Carolina Department of Public Instruction (North Carolina Department of Public Instruction [NCDPI], 2009). These technical reports confirm that subsequent editions of EOG tests have similar reliability and validity characteristics as the first edition.

The NC EOG tests have undergone several audits and technical reviews since their initial development, including the assessment peer reviews required by NCLB. They have always been found technically sufficient for the purposes for which they were designed. On at least one occasion, they were ranked the best in the country (Princeton Review, 2002).

For this paper, the EOG reading scores for 10 successive cohorts of eighth-graders (2000–2009) were each retrospectively examined over six years from the end of their third-grade year through the end of their eighth-grade year. Longitudinal panels of the students' reading measures were constructed, spanning the years 1995–2009. This compilation of longitudinally matched data yielded 10 cohort panels. The 10 successive panels represent growth in reading ability from Grades 3 to 8, but across different years: 1995–2000, 1996–2001, 1997–2002, 1998–2003, 1999–2004, 2000–2005, 2001–2006, 2002–2007, 2003–2008, and 2004–2009, respectively.

To facilitate the measurement of growth, the first edition of the North Carolina EOG Reading Comprehension Tests was designed with a developmental scale. It ranged from 100 to 200 points, approximately. The vertical scale was constructed by means of a common-items equating design.

When the second edition was made operational in 2003, it reflected a revision of the Standard Course of Study for reading, and so a new developmental scale was designed to range from approximately 200 to 300 points to differentiate it from the first edition scale. A linking study was completed to allow translation of the second edition scores onto the first edition scale 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, when the third edition of the reading tests was made operational in 2008, it reflected a new revision of the Standard Course of Study for reading and language arts, and so a third reading developmental scale was designed, this one to range from approximately 300 to 400 points to differentiate it from the second and first edition scales. A new linking study allowed the translation of the third edition scores onto the second edition scale for accountability purposes.

In his research, Williamson (2014) took advantage of the between-edition linking studies to place all scores for a cohort panel on the first edition scale. Then, because the NCDPI had linked the first edition of the NC EOG Reading Comprehension Tests to the Lexile Framework for Reading, all of the first edition scores (both the original first edition scores and the converted scale scores from the second and third editions) were expressed as Lexile measures for his growth analyses. In retrospect, we can observe that the State of North Carolina took advantage of traditional IRT to construct its tests and vertical scales. Then, by virtue of the Lexile linking study, the state's vertical scale was calibrated to the Lexile scale, which utilized a Rasch model (fit to the LLT) and the construct specification equation to calibrate the person-ability and task-complexity continuum. Thus, the NC EOG reading scale was placed into a conjoint interpretive framework that possesses general objectivity.

Modeling student growth. There are five properties that I regard as essential for the optimal measurement of student growth. First, the measurement scale should be *unidimensional*. That is, the scale should measure one and only one construct consistently across time. Otherwise, researchers

cannot state clearly what is growing from one occasion to another. So for example, researchers can measure the progression of individual reading ability across time if they have a measure of reading ability that is valid during the life-course of an individual.

Secondly, the measurement scale must be *continuous* in the usual mathematical sense. I am disposed toward using a continuous, differentiable function to describe the growth process because I wish to model not only how status changes with time, but also the velocity and acceleration of growth. Note that even if growth itself is not continuous, it is still advantageous for the measurement scale to be continuous. A discontinuous growth process can be modeled with a continuous scale, but a continuous growth process cannot be adequately modeled with a categorical scale.

Third, the measurement scale should be *developmental* in nature. That is, the scale should span grades or chronological age. This makes it possible to collect comparable, repeated measurements of an individual as he or she progresses through school or matures.

Fourth, the measuring scale must be an *equal-interval* scale. This means that the scale unit must have the same quantification (i.e., must represent the same amount of the construct) anywhere on the scale. For IRT scales, the equal-interval claim is sometimes taken to be equivalent to the claim that the scale is invariant to linear transformation (Yen & Fitzpatrick, 2006). Most researchers would maintain that these four properties are sufficient for measuring growth. However, there is a fifth property that I desire.

The fifth important measurement property is *invariance of scale unit*. Another way of saying this is that the measurement possesses general objectivity (Stenner & Stone, 2010). In simple terms, this means that the size of the scale unit is absolute. It does not change merely because a different reading test is used. It does not require reference to any particular individual or individuals. General objectivity is obtained through theory or by anchoring the scale in a real-world context (e.g., analogous to the way the meter was anchored in terms of the length of the meridian arc through Paris).

As described in the previous section, the NC EOG vertical scale was linked to the Lexile Framework for Reading, which utilizes a Rasch measurement model with a construct specification equation to place both persons and texts on a common scale, which is anchored at two points on a well-defined text-complexity continuum (Stenner et al., 2007). Consequently, for the examples I will present below, I can assert that the scale has the five properties outlined above, in addition to providing conjoint measurement of both persons and texts.

The subsequent student growth examples utilize the results of previous research, described in detail by Williamson (2014). Based on recommendations provided by Singer and Willett (2003), the first phase of Williamson's analysis strategy involved constructing exploratory plots of the individual longitudinal data, non-parametrically smoothing the trend data with splines, fitting OLS regression models (e.g., linear, quadratic, and log-time transformation models) to the individual data, and looking at the distribution of fit statistics in the population. In addition, Williamson examined aggregate trends in performance across Grades 3–8 for each successive cohort of students. He thus confirmed that the same developmental function seemed appropriate for the aggregate as well as the individual data. The aggregate trends yielded a preliminary view of how growth had changed over time.

In the second phase of his analysis strategy, Williamson's exploratory analyses were followed by formal statistical modeling of the data. He fit a sequence of multilevel models for change, which were devised to confirm the expected functional form and yield specific mathematical characterizations of growth. Thus, a parametric model was utilized to characterize 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, Williamson (2014) obtained: (a) a fitted model to summarize growth, (b) associated parameter estimates and their standard errors, and (c) estimates of the variability of individual student growth in the state. The parameter estimates he obtained are directly interpretable in terms of the features of growth (magnitude, velocity, and acceleration). By examining the parametric form across successive cohort panels, he explicated patterns in the features of growth in North Carolina over time. The fixed effects parameter estimates reported by Williamson (2014) were the basic results used for the student growth depictions provided in this paper.

Purpose and methodology

I wish to demonstrate, by a series of examples, how a conjoint understanding of student reading growth in relation to text-complexity standards provides insights for policy discussions concerned with improving student growth in the future. This narrative proceeds around the following four focal points.

First, a new policy foundation for measuring student growth has emerged in the USA. Since the passage of NCLB, it is now possible to document historical student reading growth for some states by virtue of the testing requirements in the law. NCLB required states to test annually in Grades 3–8 and one high school grade. Although the law did not require states to use developmental scales that would provide better measurement of student growth, a few states did employ such scales. In states that maintained longitudinal data and implemented sufficient measurement, it became possible to measure growth in reading and mathematics from the end of Grade 3 to the end of Grade 8 (Williamson, Thompson, & Baker, 2007). At least one state (NC) also employed a supplemental, conjoint measurement scale so that both student reading ability and text complexity could be reported with a common scale. In such a situation, historical student reading growth provides a context for understanding how challenging the emerging CCSS text-complexity standards might prove to be in practice.

Secondly, the National Governors Association Center for Best Practices and the Council of Chief State School Officers (National Governors Association [NGA] & Council of Chief State School Officers [CCSSO], 2012) proposed quantitative text-complexity ranges for specified grade bands to encourage educators to require students to read more complex texts. Their Common Core State Standards (CCSS) recommended increasing text complexity across Grades K–12 as one strategy to ameliorate a gap between high school reading ability and the complexity of postsecondary reading materials (Williamson, 2008). Williamson, Fitzgerald, and Stenner (2013) discussed some of the implications of the text complexity standards.

Third, knowledge of historical growth inspires and motivates alternative hypothetical growth trajectories that are commensurate with attaining college and career readiness by the end of twelfth grade. With an appropriate parametric growth model, alternative trajectories can be analytically examined in terms of three basic features of growth: magnitude, velocity, and acceleration.

Fourth, this strategy of considering alternative trajectories provides the basis for a richer conversation among educators about how educational policies and practices can lead to greater student preparedness for the postsecondary world. Williamson, Fitzgerald, and Stenner (2014) explored that conversation in depth.

The examples highlighted in this paper illustrate the connections between these four focal ideas. To that end, a series of figures will be presented and a helpful thought-experiment will be described. I begin with student growth.

Student growth example

In Figure 1, we see an example of measuring student growth. The horizontal axis uses grade-in-school as a time-scale. Because students were annually tested in Grades 3–8 during the final three weeks of each school year, the numerals on the horizontal axis represent times that are spaced approximately one year apart corresponding to the designated grades on the axis. The vertical axis quantifies the estimated reading ability associated with the growth curves displayed in the graph. Although the vertical scale actually ranges from below 0L to above 2000L, I have displayed only the portion of the scale from 600L to 1200L so as to provide better discrimination for reading the graph.

The 10 curves in the figure represent the estimated average growth for 10 different groups of students who were each followed and measured across six years of schooling (Grades 3–8). Each curve represents student achievement during the same span of grades (3–8), but in different years. For example, the bottom curve in the chart comes from the earliest years. Students associated with the

bottom growth curve were third-graders in spring 1995 and progressed without repeating a grade until they were eighth-graders in spring 2000. The next curve represents students who were in Grades 3–8 during the years 1996–2001, and so on, to the students associated with the highest curve, which corresponds to Grades 3–8 in the years 2004–2009. Each group of students was measured on six occasions. Cumulatively, there were 674,899 students who made up these 10 groups.

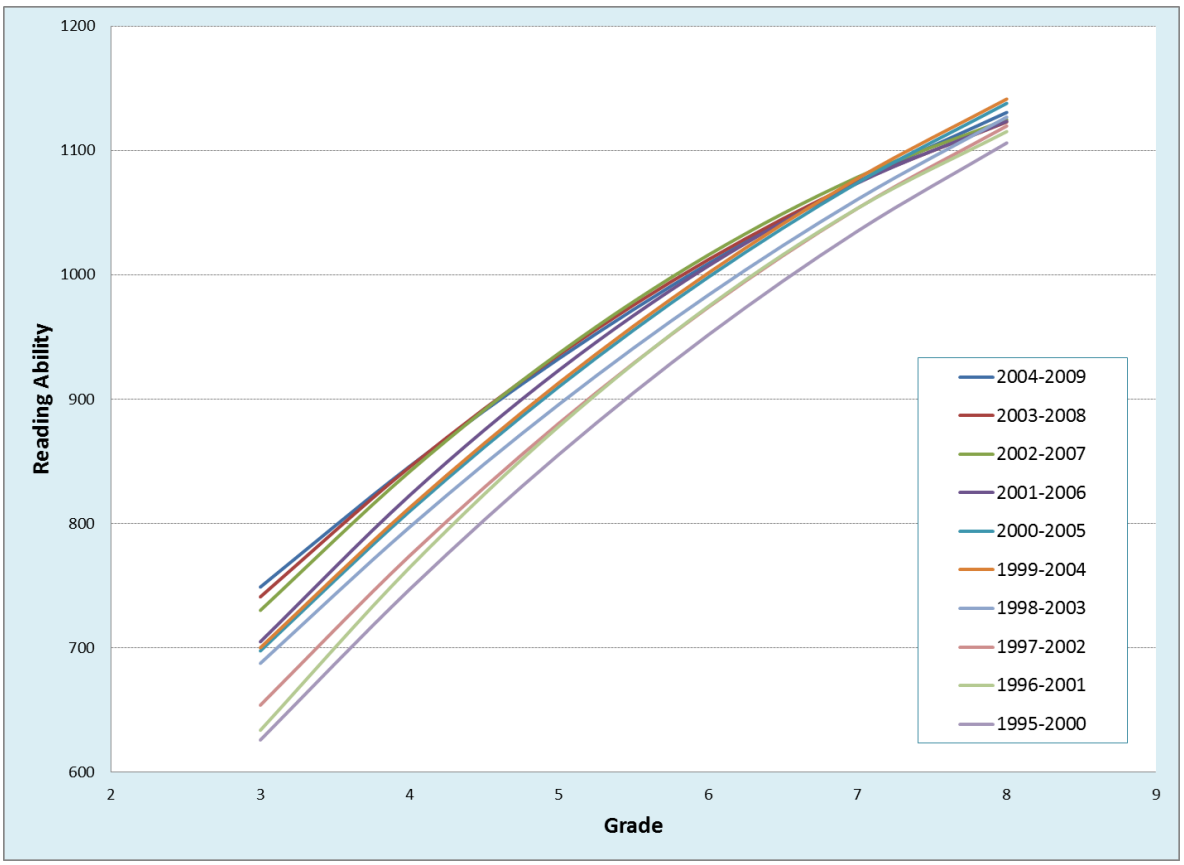

Figure 1. Average North Carolina reading growth for 10 panels of students (N = 674,899). Each curve is based on six waves of data spanning Grades 3–8 in different years, as given in the legend.

The growth curves in Figure 1 represent the average growth of students in NC, a southern (USA) state, during the specified years and timeframes. Because the curves stack up vertically one above the other in chronological order along the vertical axis, it is obvious from Figure 1 that NC experienced systemic improvement in its average reading achievement across multiple groups of students whose education spanned 1.5 decades (1995–2009). It also appears that velocity and acceleration were pretty consistent from one growth curve to another, except for slightly higher deceleration in some of the more recent curves, which was associated with slightly lower average performance at the end of eighth grade. In spite of the evidence of slightly higher deceleration for some groups, overall growth appears to be relatively strong and consistent for these students.

Text complexity examples

In Figure 2, box-and-whisker plots are used to depict within-grade distributions of text complexity. Once again, the horizontal axis is calibrated in terms of grade, spanning Grades 1–12. An additional point, labeled CCR, was added after Grade 12 on the horizontal axis to denote the postsecondary world of *college and career readiness*. The vertical axis is a scale for text complexity, ranging from 0L to 1600L. Increasing values on this scale indicate increasing difficulty or higher text complexity.

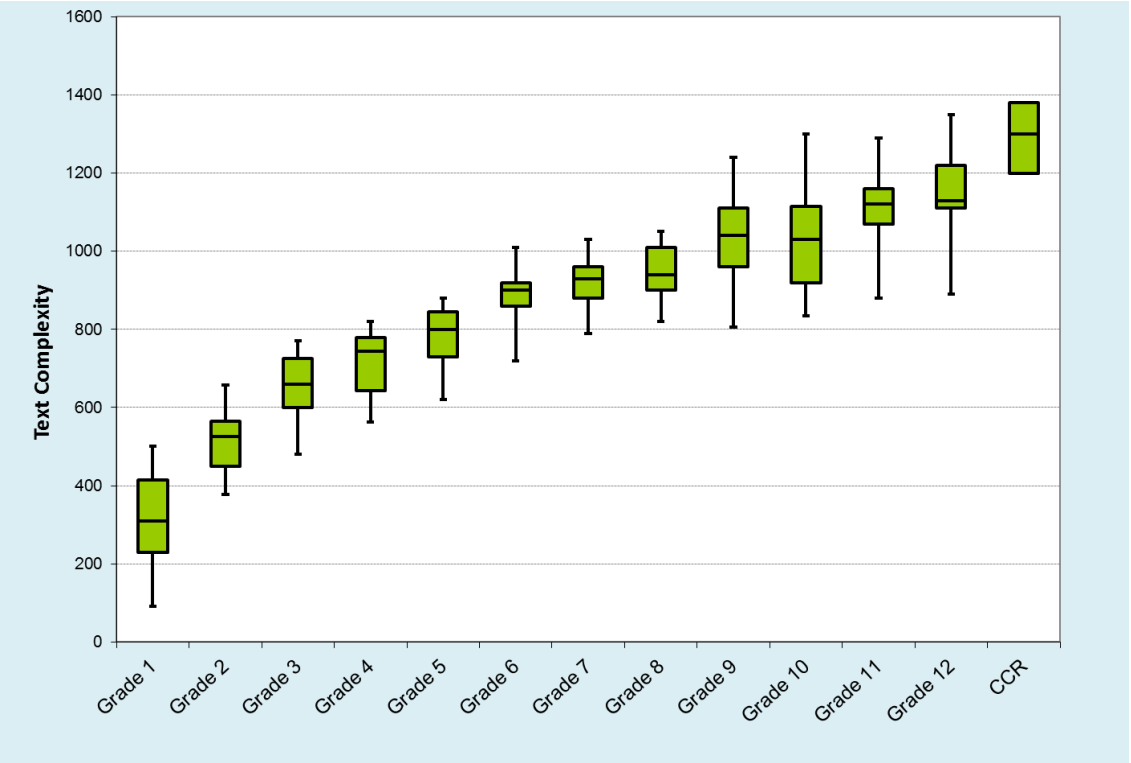

Figure 2. Empirical text-complexity distributions by grade and for college and career readiness (CCR). Box-and-whiskers represent the 5th, 25th, 50th, 75th, and 95th percentiles of the text distributions.

In the plot, the individual boxes represent the interquartile range (middle 50%) of text complexity for textbooks used in that particular grade or in the postsecondary world. The hash-mark within the box represents the median (middle) text complexity for the grade. The whiskers extend downward to the 5th percentile of texts and upward to the 95th percentile of texts. So the distance between the endpoints of the whiskers represents the text complexity of 90% of the texts associated with that grade. The box and whiskers taken together represent the within-grade variability in text complexity for textbooks that were commonly used in the public schools in the USA.

The plot reveals an increasing trend in median text complexity from earlier grades up to and including the postsecondary reading material. Median text complexity increases more rapidly during the early grades (e.g., Grades 1–6) and slows during the middle and high school years. Although there is considerable within-grade variation in text complexity and differing degrees of overlap in the text-complexity distributions for adjacent grades, the graph indicates a dramatic increase in the magnitude of text complexity for nearly all texts over the developmental experience of students as they progress to young adulthood. However, the increase in median text complexity between the end of Grade 12 and the postsecondary occasion is of particular interest, as it quantifies the gap in reading ability needed for postsecondary texts relative to high school texts.

Because choice of textbooks for schooling is an alterable educational policy and pedagogical decision, we might imagine a text continuum where the textbooks for Grades 1–12 are more aligned with the text complexity encountered in postsecondary reading materials. By shifting the text complexity distributions for Grades 1–12 upward until the Grade 12 median text complexity aligns with the median postsecondary text complexity, we arrive at an *aspirational* text continuum, which is displayed in Figure 3. In fact, this is the approach underlying the text complexity standards of the CCSS. The relationship is obvious in Figure 3 by noting the upper and lower boundaries of the CCSS text complexity ranges (the two gray lines spanning from Grade 3 to Grade 12). Sanford-Moore and Williamson (2012) summarized a strategy for shifting an empirical text-complexity continuum to create an aspirational text-complexity continuum. The text-complexity ranges actually adopted by the CCSS were described by NGA and CCSSO (2012).

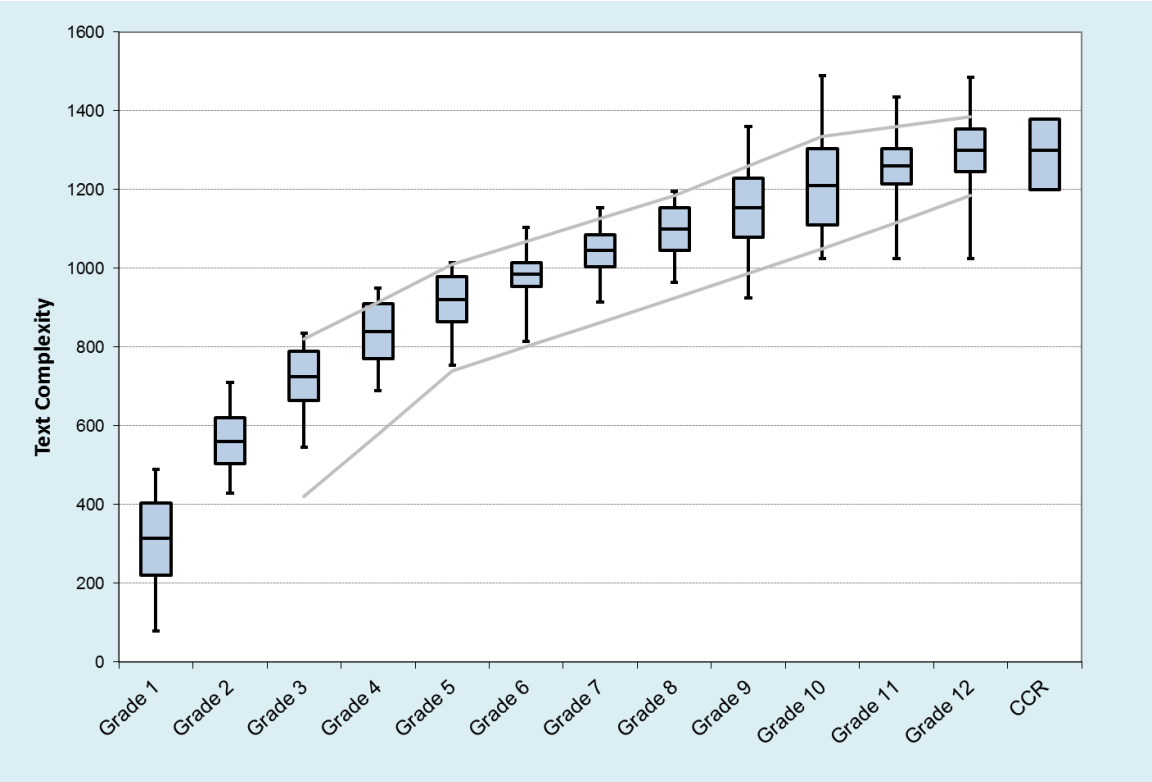

Figure 3. Aspirational text distributions, Common Core State Standards (CCSS) text-complexity ranges (gray) and interquartile range for postsecondary (CCR) text complexity. Box-and-whiskers represent the 5th, 25th, 50th, 75th, and 95th percentiles of the text distributions.

The conjoint perspective

The power of combining Rasch measurement with a theory-based model of task complexity becomes evident in Figure 4. Once again, the horizontal axis denotes time (i.e., end of consecutive grades). The vertical axis is precisely the same scale used in Figures 1–3. The striking feature of Figure 4 is that quantitative measures of both persons and texts are manifested in one graph using a single scale.

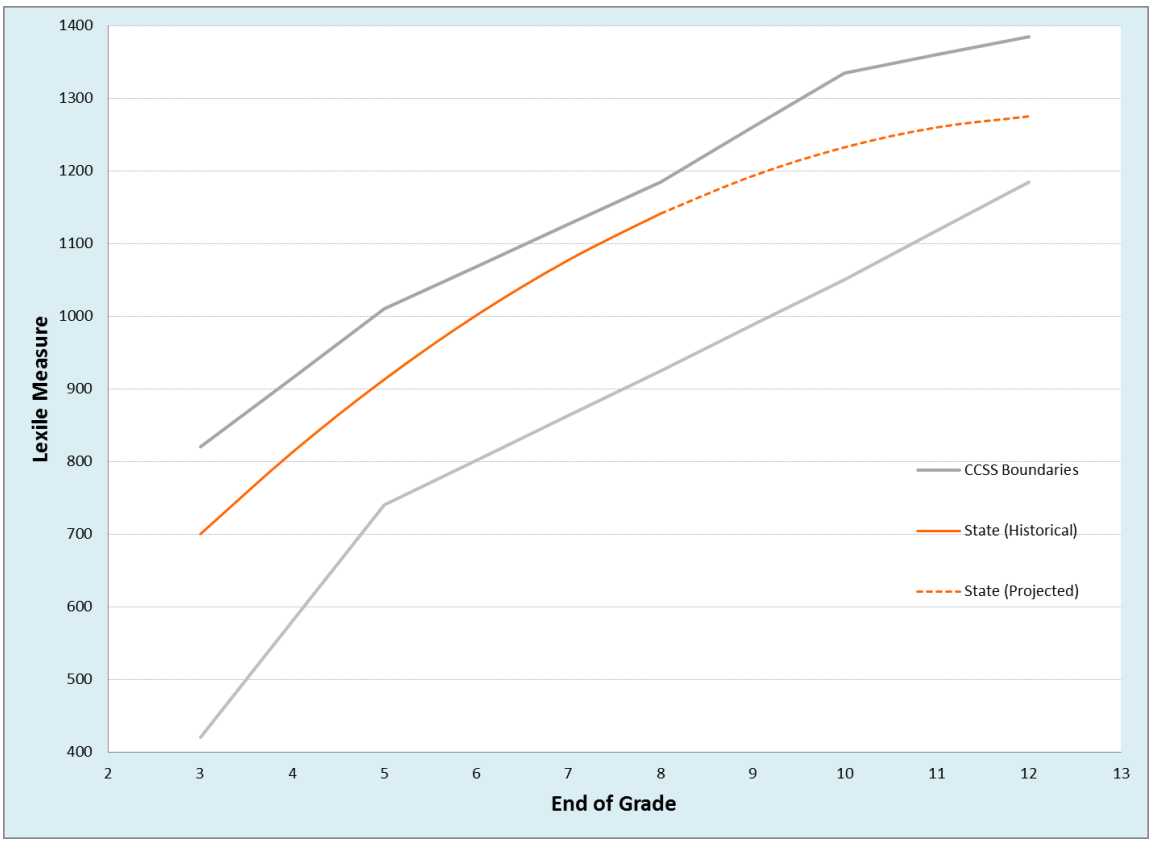

Figure 4. Historical average student growth and projected student growth relative to Common Core State Standards (CCSS) text complexity ranges. Student reading ability and text complexity are both quantified with a common (Lexile) scale.

Using this framework, two seemingly disparate perspectives (person and task) can be contrasted within a common frame of reference. Referring to Figure 4, readers will notice the CCSS text-complexity boundaries displayed from Grade 3 to Grade 12. These are the same text complexity boundaries previously depicted in Figure 3. The second thing to notice about Figure 4 is the aggregate student growth curve. Part of the curve is shown by a solid pattern, whereas the other part is shown in a dashed pattern. The solid portion of the curve represents the empirically fitted growth curve; in fact, this curve is one of the 10 curves shown in Figure 1 (the curve for the years 1999–2004). The dashed portion of the curve is a mathematical extrapolation based on the fitted growth model.

In Figure 4, it is now very easy to see the relationship between the text-complexity standards and the average growth of students —both the actual historical growth in Grades 3–8 and the forecasted growth in Grades 9–12 under the condition that students follow the same historical path into the future without any change in their growth curve. Two outcomes are readily apparent from the picture: (a) During Grades 3–8, students on average appear to be performing well in relation to the CCSS text-complexity standards, climbing near to the upper boundary of the text-complexity range by the end of Grade 8; and (b) the deceleration of the historical trajectory implies that students could actually fall farther below the text-complexity requirements by the end of Grade 12 if nothing occurs to attenuate their deceleration. Of course, the growth curve in Figure 4 represents average growth. Individual growth curves would likely vary greatly around the average growth curve, with some students doing even better in relation to the text-complexity standards but other students doing far worse than average, perhaps. This raises a potentially important question. How can student growth be altered so that students may reach higher levels of performance in the future? That is, how would the features of growth (magnitude, velocity, acceleration) have to change in order for students to reach a different aspirational outcome?

Implications for educators

If educators are to be able to contemplate the long-term implications of growth, they have to think beyond the immediate status of an individual at a particular point in time. In fact they have to go beyond short-term notions of growth, such as the amount of gain made during a given school year. Truly, educators must think about the entire growth trajectory as a dynamic entity with specific characteristics that determine growth over the lifespan. More concretely, it is useful to focus on specific features of growth (e.g., magnitude, velocity, and acceleration) that are inherent in the growth process. With this perspective, it is possible to refine pedagogical questions and strategies to focus on specific features of growth. For example, what feature of growth must change —and by how much— to result in a different terminal outcome? Below is an example of such a thought experiment, depicted in Figure 5.

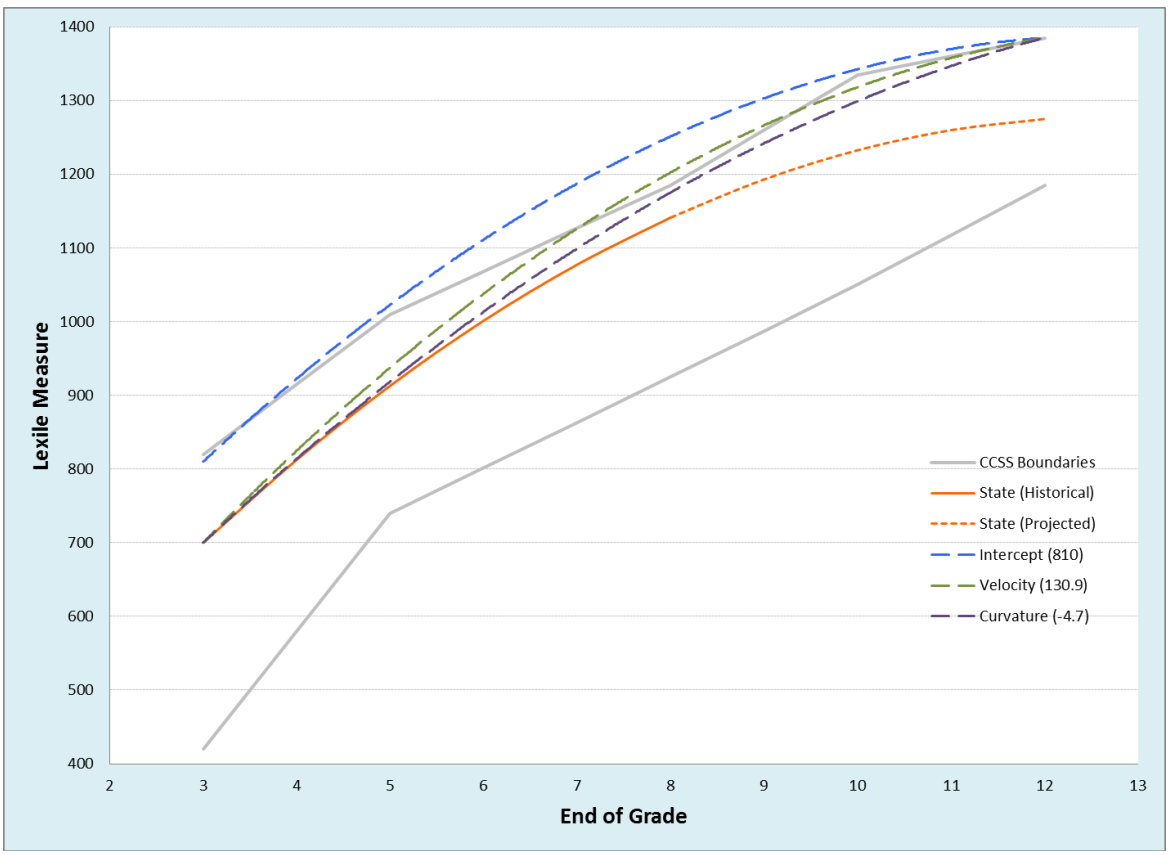

Figure 5. Historical average student growth relative to Common Core State Standards (CCSS) Text Complexity ranges, with three alternative paths to 1385L. Student reading ability and text complexity are both quantified with a common (Lexile) scale. Adapted from «Student Reading Growth Illuminates the Common Core Text-Complexity Standard: Raising Both Bars,» by G. L. Williamson, J. Fitzgerald, and A. J. Stenner, 2014, Elementary School Journal, 115(2), p. 245. © 2014 by The University of Chicago.

Note that Figure 5 has the same axes as Figure 4 and the two featured elements in Figure 4 —the CCSS text complexity boundaries and the average growth curve— are replicated in Figure 5. However, three additional curves appear in Figure 5. These curves provide examples of three archetypal paths to an aspirational outcome (1385L), which is the CCR target established by the CCSS text-complexity standards.

The upper-most curve depicted in Figure 5 represents a path to 1385L that is achieved by raising the elevation of the entire historical growth curve across all grades. Specifically, the upper-most curve has a higher initial starting point (810L) than the historical curve (700L), yet they are parallel in every other respect. The consequence of this one change is that the new curve reaches the CCR goal, whereas the forecast based on the historical curve does not. The next-lower curve was fashioned by changing the

initial velocity of the historical curve so that the new curve reached 1385L. And finally, the third curve from the top was fashioned by changing the deceleration rate of the historical curve just enough to reach 1385L by the end of Grade 12.

It is important to understand that this thought experiment is an oversimplification. Often, there is a correlation structure among the parameters of empirical growth curves; that is, the status, velocity, and acceleration parameters are not completely independent of each other. So, researchers cannot change just one of these three features without expecting a concomitant change in the others. Nevertheless, I am setting that reality aside for heuristic purposes. The three new trajectories are not paramount in and of themselves. Of importance are the questions the three new trajectories raise about educational practice and for that discussion, it is useful to think about each possibility in turn.

Contemplating the highest of the new curves, the pertinent question for educational policy and practice is how to impact initial status for a group of students. This is a common question with which educators have much experience. The typical response is to explore options for early intervention initiatives. The next-lower curve invokes a consideration of how to change the velocity of student learning. Responses might revolve around the pacing of instruction during the timeframe. A consideration of the third curve raises the question of how to impact deceleration. This is a question rarely contemplated, except perhaps in the context of programs designed to combat summer loss in student learning.

Using the scenarios depicted in Figure 5, Williamson et al. (2014) provided an in-depth discussion of the pedagogical and policy implications of attempting to modify the magnitude, velocity, or acceleration of student growth; they also discussed alternatives and implications in the context of extant reading research. As they discussed each alternative student growth scenario, the authors addressed the following three issues, along with evidence bases:

> (*a*) What implications are there for the challenge levels of texts students would need to read at different grade levels, and how do those challenge levels compare to the CCSS grade-by-grade target boundaries? (*b*) In comparison to the historical student reading-growth curve, which students would be most affected by the shift, and what is known from developmental reading research that might inform consideration of that shift? (*c*) What curricular, instructional, or policy implications are there? (p. 246).

The analysis provided by Williamson et al. (2014) is an example of how educators can intellectually engage with the substantive implications of student reading growth in the context of text-complexity standards.

Discussion

There are several additional benefits which derive from the methodological perspective described in this paper. First, this perspective motivates a conversation about academic growth that is fundamentally anchored in a sound measurement process. Second, this perspective encourages educators to adopt a long-term, lifespan approach to student growth based on collecting and analyzing longitudinal data. Third, this approach encourages educators to conceptualize growth as a dynamic developmental process. These benefits lead to new insights about student growth, which in turn can inform educational policy and practice.

The most important lessons of this paper may be for educational measurement professionals. Optimally measuring academic growth requires a scale that is unidimensional, continuous, equal-interval, developmental, and invariant with respect to location and unit size. For the reading ability construct, these characteristics can be attained by combining Rasch measurement with an operational specification for the reading construct (in terms of the task continuum) and anchoring the resulting scale at two points on the text-complexity continuum. Hence, persons and texts are brought onto a common, developmental scale, which possesses general objectivity.

It is worth noting that this approach need not be restricted to the reading construct. For example, if the difficulty of mathematical (or other) skills and concepts can be quantified and shown to be highly related to the item difficulties of a particular reference item type, then student mathematics ability and the difficulty of mathematical tasks can be conjointly scaled using the Rasch model. This would provide theory-driven, conjoint measurement for mathematics (or other constructs) similar to that demonstrated here for reading ability.

This paper focused on how student growth provides a context for considering the educational implications of text complexity. In fact, with the approach highlighted here, text complexity could also provide a context for understanding student growth. This is the dual benefit of conjoint measurement. Each facet (persons, tasks) can provide a context for understanding the other.

The original article was received on November 10th, 2014
The revised article was received on April 17th, 2015
The article was accepted on May 13th, 2015

References

Alvermann, D. E., Hinchman, K. A., Moore, D. W., Phelps, S. F., & Waff, D. R. (Eds.). (2006). *Reconceptualizing the literacies in adolescents' lives.* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Alvermann, D. E., Unrau, N. J., & Ruddell, R. B. (Eds.). (2013). *Theoretical models and processes of reading.* (6th ed.). Newark, DE: International Reading Association.

Bazemore, M., & Van Dyk, P. B. (2004). *North Carolina reading comprehension tests: Technical report.* (2nd ed.). (Citable draft). Raleigh, NC: North Carolina Department of Public Instruction.

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

Bormuth, J. R. (1969). *Development of readability analyses.* (Final Report, Project No. 7-0052, Contract No. OEG-3-7-070052-0326). Washington DC: US Office of Education, Bureau of Research, US Department of Health, Education and Welfare.

Brennan, R. L. (Ed.). (2006). *Educational measurement.* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

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

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley. Reprinted by Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.

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

Kendeou, P., van den Broek, P., Helder, A., & Karlsson, J. (2014). A cognitive view of reading comprehension: Implications for reading difficulties. *Learning Disabilities Research & Practice, 29*(1), 10-16. doi: 10.1111/ldrp.12025

Koslin, B. L., Zeno, S., & Koslin, S. (1987). *The DRP: An effectiveness measure in reading.* New York: The College Entrance Examination Board.

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes.* Paris: Courcier.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph, 7.* Richmond, VA: Psychometric Corporation.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

National Governors Association (NGA), & Council of Chief State School Officers (CCSSO) (2012). *Supplemental information for Appendix A of the Common Core State Standards for English language arts and literacy: New research on text complexity*. Washington, DC: Author. Retrieved from http://www.corestandards.org/resources

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance.* Washington, DC: Council of Chief State School Officers (CCSSO). Retrieved from http://www.ccsso.org/Documents/2012/Measures%20ofText%20Difficulty_final.2012.pdf

No Child Left Behind (NCLB) Act of 2001. *Public Law N° 107-110, § 1, 115 Stat. 1425* (2002).

North Carolina Department of Public Instruction (NCDPI) (2009). *North Carolina Reading Comprehension Tests: Technical report.* (3rd ed.). Raleigh, NC: Author. Retrieved from http://www.ncpublicschools.org/docs/accountability/testing/reports/eogreadingtechman3.pdf

Pearson, P. D. (2004). The reading wars. *Educational Policy, 18*(1), 216-252. Retrieved from http://www.corwin.com/upm-data/31886_Article1.pdf

Princeton Review (2002). *Testing the testers 2002: An annual ranking of state accountability systems.* New York: Author.

Questar Asssessment, Inc. (2012). *Degrees of Reading Power (DRP) Program.* Brewster, NY: Author. Retrieved from http://www.questarai.com/Products/ DRPProgram/Pages/default.aspx

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielson and Lydiche (for Danmarks Paedagogiske Institut).

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19,* 49-57. doi: 10.1111/j.2044-8317.1966.tb00354.x

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In M. Blegvad (Ed.), *The Danish yearbook of philosophy* (pp. 58-94). Copenhagen: Munksgaard. Retrieved from http://www.rasch.org/memo18.htm

Reckase, M. D. (2009). Logical units for achievement test scores. *NCME Newsletter, 17*(1), 1-15. Retrieved from http://ncme.org/default/assets/File/pdf/newsletter/vol_17_num_1.pdf

Sanford, E. E. (1996). *North Carolina end-of-grade tests technical report # 1: Reading comprehension, mathematics.* Raleigh, NC: Department of Public Instruction.

Sanford-Moore, E. E., & Williamson, G. L. (2012). *Bending the text-complexity curve to close the gap* (MetaMetrics Research Brief). Durham, NC: MetaMetrics, Inc.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2007). *The Lexile framework for reading technical report.* Durham, NC: MetaMetrics, Inc.

Stenner, A. J., Sanford-Moore, E., & Williamson, G. L. (2012). *The Lexile® framework for reading quantifies the reading ability needed for «College & Career Readiness»* (MetaMetrics Research Brief). Durham, NC: MetaMetrics.

Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Education Measurement*, *20*, 305-316. doi: 10.1111/j.1745-3984.1983.tb00209.x

Stenner, A. J., & Stone, M. (2010). Generally objective measurement of human temperature and reading ability: Some corollaries. *Journal of Applied Measurement*, *11*(3), 244-252.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring.* Mahwah, NJ: Lawrence Erlbaum Associates.

Williamson, G. L. (2008). A text readability continuum for postsecondary readiness. *Journal of Advanced Academics, 19,* 602-632. doi: 10.4219/jaa-2008-832

Williamson, G. L. (2014). *Measuring and modeling individual academic growth: Methodological foundations for educational applications in the 21st century.* (Unpublished manuscript).

Williamson, G. L., Fitzgerald, J., & Stenner, A. J. (2013). The Common Core State Standards' quantitative text-complexity trajectory: Figuring out how much complexity is enough. *Educational Researcher, 42*(2), 59-69. doi: 10.3102/0013189X12466695

Williamson, G. L., Fitzgerald, J., & Stenner, A. J. (2014). Student reading growth illuminates the Common Core text-complexity standard: Raising both bars. *Elementary School Journal*, *115*(2), 230-254. doi: 10.1086/678295

Williamson, G. L., Koons, H., Sandvik, T., & Sanford-Moore, E. (2012). *The text complexity continuum in grades 1-12* (MetaMetrics Research Brief). Durham, NC: MetaMetrics.

Williamson, G. L., Thompson, C. L., & Baker, R. F. (2007). *North Carolina's growth in reading and mathematics.* Paper presented at the 2007 Annual Meeting of the American Educational Research Association (AERA), Chicago, IL., USA.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 64-110). Westport, CT: American Council on Education and Praeger Publishers.