Technical Issues in the Use and Interpretation of Growth Models for Students With and Without Disabilities

Joseph Stevens
Keith Zvoch
and
Gina Biancarosa

Department of Educational Methodology, Policy, and Leadership
University of Oregon

Presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, BC, Canada, April, 2012.

Contact Information:

Joseph Stevens, Ph.D.
College of Education
5267 University of Oregon
Eugene, OR 97403
(541) 346-2445
stevensj@uoregon.edu

Presentation available at:
http://www.uoregon.edu/~stevensj/stevens2012.pdf
And on NCAASE web site soon: http://www.ncaase.com/

This work was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R32C110004 awarded to the University of Oregon. The opinions expressed are those of the author and do not necessarily represent views of the Institute or the U.S. Department of Education.

Presentation Purpose

- Presentation serves as an incomplete guide to some of our future directions in research conducted by NCAASE
- NCAASE will study formative, interim, and summative assessments
- Focus here is on brief introduction to six issues:
  - (a) measurement, scaling, and vertical linking
  - (b) number and timing of assessments
  - (c) missing data, attrition, and mobility
  - (d) functional form of growth
  - (e) uncertainty in estimation
  - (f) sensitivity and validity

http://www.uoregon.edu/~stevensj/stevens2012.pdf 3

Measurement, Scaling, and Vertical Linking

- How do current methods of test development, sampling of content domains, and the analysis of items for test construction relate to the estimation of growth?
- State summative assessments designed to provide a representative sample of state mandated content standards at each grade level tested
- Measuring the developmental trajectory of growth a different assessment goal
- Some question whether these traditionally useful and effective normative or standards based assessment methods are well suited to the measurement of growth and instructional effectiveness

http://www.uoregon.edu/~stevensj/stevens2012.pdf

Measurement Issues

- Molar versus molecular constructs
- Cognitive development versus status within a grade
- "Growth in performance does not occur in a uniform way. There are jumps and pauses in the patterns of growth" (Reckase & Martineau, 2004, p. 17); dimensionality, linearity
- Schmidt et al. (2005) identify curricular changes across grades that can undermine growth estimation through resulting changes in test dimensionality and in the measured constructs
- Construct equivalence and temporal stability
  - Across different spans
  - Across different content
  - For students with and without disabilities

Measurement, Scaling and Vertical Linking

- As the span of time covered by the growth model increases, the importance of construct temporal stability expands (Willett, et al., 1998)
- Without stability, uncertain whether growth is due to construct or due to concomitant changes in instrument or other construct irrelevant factors
- Degree of temporal invariance also depends on the type and kind of content and construct being assessed (Loyd & Plake, 1987; Yen & Burket,1997)
- General although not universal agreement that vertically linked, developmental scales needed to measure growth (Betebenner, 2008; Briggs & Weeks, 2009; Kolen & Brennan, 2004)

The Challenges of Vertical Linking

"When I was about 5 years old, I used to follow my father...[who] had a folding ruler [that] was yellow, with hinged 1-foot lengths that would unfold...to 6 feet. If I held the extended ruler at one end, it would curve gracefully through space. To my disappointment, if I leaned it too much to the side, one of the looser hinges would suddenly bend sharply" (Yen, 2007, pp. 274-5).

Vertical Linking

- Typical linking designs include the *Scaling Test* design and the *Common Item* design (Kolen & Brennan, 2004)
- *Scaling Test* design uses two types of tests, a scaling test covering content taught at all grade levels and level tests that are grade specific
  - Scaling test administered to students at all grade levels
  - Level tests administered only to the students at the relevant grade level
  - Scaling test locates scores on developmental scale; level test used to estimate the ability level of examinees within each grade level
- *Common Item* design, only level tests used
  - Linking accomplished through common item blocks shared across adjacent grade levels
  - Chained linking process used to put all grades on a single common developmental scale

Vertical Linking

- However, growth patterns differ depending on design used (Andrews, 1995; Hendrickson, Kolen, & Tong, 2004; Petersen, Kolen, & Hoover, 1989)
- Different linking methods and data collection designs appear to lead to somewhat different vertical scales (Camilli, Yamamoto, & Wang, 1993; Williams et al., 1998; Yen, 1986)
- As a result, interpretations of growth depend on the degree to which the vertical scale is "stretched" or compressed" by the choice of methods used in scale creation (Briggs & Weeks, 2009)
- More recently, however, some have argued that vertical scales are not necessary in order to measure growth

Measurement, Scaling and Vertical Linking

- Several models now in use track conditional change:
  - Change in student proficiency categories over time (Furgol & Helms, 2011; Hoffer et al., 2011)
  - Change in a student's normative position in an achievement distribution over time (e.g., Betebenner, 2009)
- A reconciliation of differing views on vertical scales aided by definitions of Briggs & Betebenner (2009):
  - Growth conditional on time is an absolute growth model
  - Growth conditional on prior achievement is a relative growth model
- Kolen & Brennan (2004) and Harris (2007): concept of growth has no empirical definition; can only be established externally through the linking of scales to an understanding of how students learn in the content area or ability of interest

Measurement, Scaling and Vertical Linking

- Issues further complicated for special education students
  - If assessed using the same instrument as regular education students, many of the psychometric and scaling issues may be same or similar
  - Some growth analyses have shown that, although there are large differences in intercept, there are no statistically significant differences in linear slope or quadratic change for students with disabilities or students receiving a modified test administration (e.g., Stevens, 2005)
- Alternate assessments bigger challenge for measurement of growth
  - A few state alternate assessments are scaled using the same methods as the regular education assessment (e.g., Oregon)
  - More commonly alternate assessments involve performance assessment methods or do not place student performance on a continuous score scale but in one of several proficiency categories.
- Will require different analytic methods

Number and Timing of Assessments

- Many growth models based on analysis of annual test results; may not be the optimal design
- Number and timing of occasions important design consideration
- Pre-post, two wave studies most common longitudinal design in social science research (Willett, et al., 1998)
- "Two waves of data are better than one, but maybe not much better" (Rogosa, 1995, p. 744)
- Most growth models assume and apply linear growth models; most data appear to be at least curvilinear
- Multiple measurement occasions allow for increased reliability of estimation of the growth function
- A few examples:

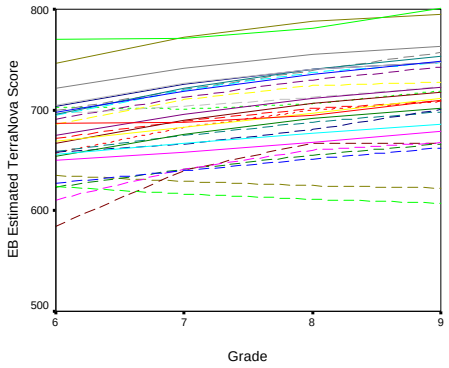


Figure 3. A Sample of Student Curvilinear Growth Trajectories over Four Waves of Annual State Mathematics Assessments.

13

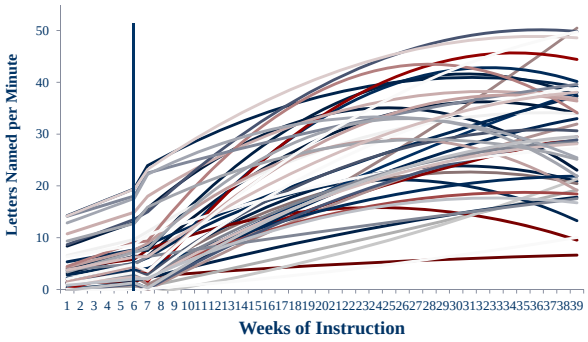


Figure 4. A Sample of Student Curvilinear Fluency Growth Trajectories Over Weeks of the School Year With Intervention at Week 6.

14

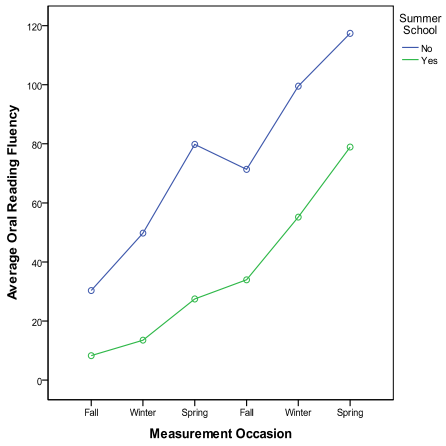


Figure 5. Seasonal Growth for At-risk Students Who Attended Summer School and Students who Were Not Invited to Summer School.

15

Missing Data, Attrition, and Mobility

- Missing data can undermine accuracy of growth estimates and may occur for many reasons (e.g., illness, drop out, mobility)
- Analysis of only the stable, non-mobile students in the same school for two or three years yields biased results (see Lockwood et al., 2006; Zvoch & Stevens, 2005)
- Percentage of students matched over 3-4 years ranges from less than 30% to about 85% (Lockwood, et al, 2006; McCall, Kingsbury, & Olson, 2004)
- Stevens (2005) study of middle school students (*N* ~ 24,000 per grade) matched 85% over two years, 81% over three years, and 75% over four years

Missing Data, Attrition, and Mobility

- Uncertain how attrition will bias results for school accountability
  - More temporary "attendance-type" attrition may be essentially random and not bias estimates (may inflate SE's and precision of estimation)
  - More permanent enrollment and mobility changes likely to introduce bias because mobility rates and patterns correlated with student socio-demographics
- Researchers often fail to report details on missing data, attrition, and mobility; list-wise deletion most common
- Luo & Kwok (2012) simulation study found that, unless cross-classified models used:
  - School estimates are biased
  - Extent of bias a function of the size and pattern of mobility
  - Spurious results obtained even when the mobility rate was relatively low
- Example:

Table 1
Student Demographic Characteristics by Analytic Sample (from Zvoch & Stevens, 2005)

	Accountability Sample (*N* = 3,334)		Complete Cohort Sample (*N* = 6,098)	
Student Characteristic	*Frequency*	*Percent*	*Frequency*	*Percent*
Female	1,710	51.3	3,016	49.5
Non-Anglo	1,797	53.9	3,536	58.0
English Language Learner	397	11.9	1,121	18.4
Free Lunch Recipient	1,170	35.1	2,628	43.1
Special Education	101	3.0	1,092	17.9

Functional Form of Growth

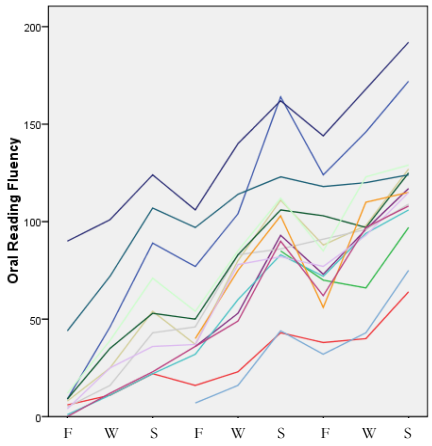
- Common assumption, expectation, model specification, prediction and projection, etc. is that growth is linear
- However, there is a weight of evidence that growth is often at least curvilinear (if not more complex)
- Nonlinearity occurs both within grade and between grade
- Accurate estimation of shape of growth function a critical step in evaluating the course of learning or development (Rogosa, 1979; Willett, et al, 1998)
  - With only two waves of data it is not possible to evaluate the shape or form of the growth function; only a linear function can be fit to the data
  - In this circumstance, it is also not possible to determine the goodness of fit of the function to the data
- See table below

Functional Form of Growth

Function	Number Waves Required for GOF	Points of Inflection	Exponential Form
Linear	3	0	1
Quadratic	4	1	2
Cubic	5	2	3
Quartic	6	3	4
Quintic	7	4	5

Individual Student Growth in Reading Fluency
(fall, winter, spring for grades 1-3)

Note annual summer drop in performance



21

Individual Student Growth in Reading Fluency (fall, winter, spring for grades 1-2) for Students Enrolled in an Intensive Summer School Program

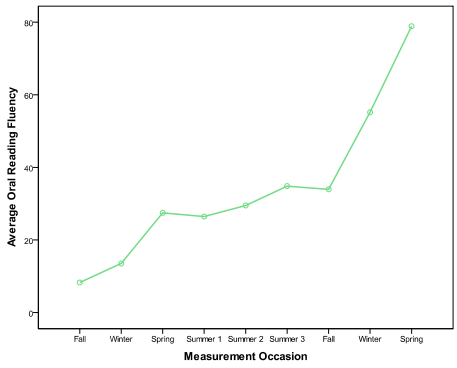


Figure 7. Reading Fluency Over First Grade, Summer, and Second Grade for Summer School Participants.

22

Functional Form of Growth

- Fit all functions beginning with linear up to the highest order polynomial possible
- Graphical analysis should be used as well; aggregate growth can mask individual differences in trajectories
- Test interactions between functional form and key student characteristics or demographics
  - Example: no interaction for special education students, their intercept significantly lower but no statistically significant difference in annual rates of linear or curvilinear growth (Stevens, 2005)
  - Example: interaction present for English language learners, statistically significant ($p < .001$) higher linear growth rate but also statistically significant ($p < .001$) curvilinear deceleration of growth

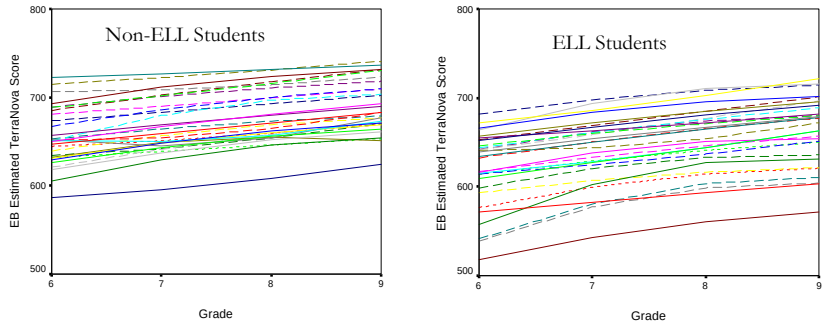


Figure 8. Estimated growth trajectories for Non-ELL and ELL students.

Uncertainty in Estimation

- Sample size
- Standard errors and confidence intervals
- In addition to sample size, parameter reliability (i.e., intercept, slope) impacted by:
  - Magnitude of the intraclass correlation with higher units (students, classes, schools)
  - Number of measurement occasions for slopes
- When coupled with low variability in school average rates of growth, school-level growth estimates can have very low reliability (Willett, 1988)
- Reliability of growth estimates improves dramatically with additional waves, even three occasions are considerably more accurate than two (Raudenbush, 2001)

Sensitivity and Validity

- Use of longitudinal designs introduces a number of potential difficulties and threats to internal validity not present in cross-sectional designs
- Sensitivity of model estimates to variations that can impact inference such as:
  - Sample size for total and disaggregated groups
  - Inclusion or omission of covariates
  - Cohort stability
  - Regression to the Mean in two-wave models
- Validation of longitudinal models for accountability systems must include evaluation of whether inferences drawn about assessment systems and school effectiveness are warranted (Forte-Fast & Hebbler, 2004; Plake, 2002)

Sensitivity and Validity

- Evidence is needed that demonstrates that the method legitimately captures the effects of school policy and practice and that the method is relatively immune to the influences of construct irrelevant sources of variation
- Examination of plausible rival hypotheses (Rindskopf, 2000; Riechardt, 2000) provides another mechanism for studying and validating alternative methods
- Shadish, Cook, & Campbell (2002) describe a process of pattern matching that involves the logical, theoretical consideration of the attributes and characteristics of a construct that should be present followed by a process of observation and matching of actual attributes and characteristics as a method of determining validity

Sensitivity and Validity

- Need for additional study of alternative growth models
- Some evidence on relationship of models to confounding factors such as student socioeconomic status and school outcome measures:
  - Stronger relationships for status scores, percent proficient (Ballou et al., 2004; Hill & DePascale, 2003; Linn & Haug, 2002; Ponisciak & Bryk, 2005; Sammons, Mortimore, & Thomas, 1996; Willms, 1992; Teddlie, Reynolds, & Sammons, 2000)
  - Smaller associations found for growth measures (Ballou, Sanders, & Wright 2004; Bryk et al., 1998; Bryk & Raudenbush, 1988; Ponisciak & Bryk, 2005; Willms, 1992; Stevens, 2000; Stone & Lane, 2003; von Hippel, 2009)

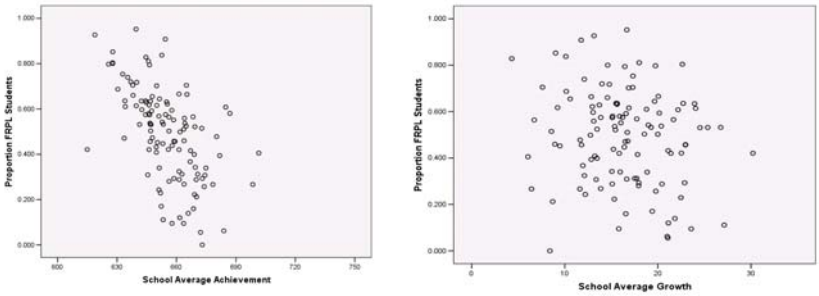


Figure 9. Relationships Between the Proportion of Free-Reduced Price Lunch (FRPL) in the School and School Average Status (r = -.56) or Growth (r = -.17) for Middle School Mathematics Achievement.

29

Contact Information

Joseph Stevens, Ph.D.
College of Education
5267 University of Oregon
Eugene, OR 97403
(541) 346-2445
stevensj@uoregon.edu

Presentation available at:
http://www.uoregon.edu/~stevensj/stevens2012.pdf