British Journal of Educational Psychology (2015), 85, 138–153

The British Psychological Society

www.wileyonlinelibrary.com

Special Issue Article

Approaches to academic growth assessment

Eric M. Anderman*, Belinda Gimbert, Ann A. O'Connell and Lisa Riegel

The Ohio State University, Columbus, Ohio, USA

Background. There is much interest in assessing growth in student learning. Assessments of growth have important implications and affect many policy decisions at many levels.

Aims. In the present article, we review some of the different approaches to measuring growth and examine the implications of their usage.

Sample. Samples used in research on growth models typically include students enrolled in public schools that primarily serve kindergarten through the 12th grade.

Method. Definitions of growth and gain are reviewed, and five types of growth models are examined: (1) Student Gain Score Model, (2) The Covariate Adjustment Model, (3) The Student Percentile Gain Model – referred to as single-wave value-added models, (4) Univariate Value-Added Response Models, and (5) Multivariate Value-Added Response Models.

Results. Modelling approaches are vastly different, whereas Student Gain Models are mathematically and conceptually simple, Multivariate Models are highly complex.

Conclusion. Educators assessing growth must make critical decisions about measurement. The type of instrument that is selected and the type of analytic techniques selected are of great importance. Growth must be considered from technical, pedagogical, and policy perspectives.

Researchers and policymakers have acknowledged for many years that assessment of student learning is complex and involves the intersections of a number of disciplines, including psychometrics, education, psychology, and sociology (Brookhart, 2004). In recent years, researchers and policymakers have become particularly focused on measuring student growth (i.e., how indices of student learning change over time within the same student). Throughout much of the history of educational assessment, the emphasis has been on measurement that occurs at one time point; more recently, the focus on growth has become a priority for reasons that we will discuss in this article.

Models that attempt to capture growth and change are at the foundation of our quest to better understand human learning. Research examining growth and change are encountered in all fields of human study, including psychology, medicine, mental health, sociology, and education. Only in education, however, is the collection and monitoring of data nearly universal: Nearly every country, state, district, and school uses some form of student assessment to track and monitor student academic progress and academic

*Correspondence should be addressed to Eric M. Anderman, Department of Educational Studies, The Ohio State University, 121 Ramseyer Hall, Columbus, OH 43210, USA (email: anderman.1@osu.edu).

DOI:10.1111/bjep.12053

attainment. Across the globe, efforts to improve academic assessment of students is an ongoing endeavour, primarily due to accountability requirements aimed at examining not only student preparation and learning, but also quality of teachers, quality of teacher education, and quality of schools. Cross-nationally, the triennial Programme for International Student Assessment assesses 15-year olds in reading, mathematics, and science (one subject every 3 years) as well as samples of teachers and schools from 65 countries and provides a mechanism for comparing and examining changes in student performance over time or evaluating systemwide education policy. However, as individual students are not tracked, individual growth cannot be examined.

In the United States, for example, efforts are underway to facilitate the assessment of individual growth with standards based on national goals. The national-level Common Core State Standards (*Common Core*) have been accepted by 90% of states, replacing previous state-level standards (ETS, 2013). Corresponding newly developed national assessments, such as the *Partnership for the Assessment of Readiness for College and Career* (PARCC) and *Smarter Balanced*, will enable teachers, administrators, parents, and other stakeholders to better understand how schools are preparing students.

It is essential that researchers, practitioners, and policymakers understand the various approaches to the assessment of academic growth, because the high-stake nature of accountability and assessment in many nations affords these indicators of student learning more importance than ever. Whereas in the past, only students were evaluated by such assessment systems, we now live in an era where – in addition to students – teachers, schools, districts, and even entire countries are evaluated by the results of various assessments (Ercikan, 2006). Indeed, researchers are now able to merge data from multiple nations to examine growth across countries. For example, Lee and Fish (2010) merged data from the National Assessment of Educational Progress (NAEP) in the United States with data from the Trends in International Mathematics and Science Study to compare growth across a variety of nations.

Advantages to assessing growth

There are several reasons why assessments of academic growth may be more useful than approaches that only gauge learning at one static point in time. To date, student learning has been measured by two related, but distinct paradigms in high-stake testing environments: *Achievement* and *progress*, known also as *growth* (Finn, Ryan, & Partin, 2008).

Measuring achievement at only one time point raises concerns regarding interpretability, comparability, and usefulness for policy decisions. For example, US school districts, historically, have applied status-based methods, which used current student scores and ignored the incoming knowledge level of students and other school variables (Coleman, Campbell, & Kilgore, 1982). In some districts, achievement as a result of a learning experience has been measured by a student's demonstration of competence or attainment of proficiency at a particular point in time, with proficiency pre-defined by a specific academic standard. In other instances, achievement has been measured by comparing students' performance on a criterion-referenced assessment with a benchmark, similarly as a time-sensitive snapshot. Aggregated achievement scores are then used to compare schools and districts, to determine which schools may be labelled as effective, in need of greater support through additional resources and monitored assistance, or

underperforming and in need of significant intervention. Consequently, the disparate definitions and arbitrary scaling of student achievement among individuals who prepare, recommend and approve, and implement policy – as well as within the educational community at large – have resulted in confounded efforts to successfully assess the influence of innovative initiatives that aim to both nurture and accelerate education reforms. These issues are particularly problematic when achievement based on assessment data from a single point in time is used to make decisions about student, teacher, or school performance.

As policymakers have turned their attention to teacher quality and accountability, they have recognized the need to consider student growth, rather than absolute student achievement scores, when designing systems of accountability (Aitkin & Longford, 1986; Olson *et al.*, 1998). McCoach, Rambo, and Welsh (2013) identified three advantages of models that assess growth of student learning: (1) growth models are more equitable than are other measurement models, because schools vary greatly in terms of students' initial levels of achievement; (2) growth is less strongly related to socioeconomic status than is overall achievement; and (3) growth models allow schools to be recognized for improvements in student learning, acknowledging that the overall (i.e., mean or median) measured academic performance of the students in a given school is distinct from growth in performance. Thus, growth models may be considered more equitable, because they can acknowledge students' progress, even when initial levels of achievement (and mean school achievement) may vary considerably.

Growth and its relation to development and learning

Whether or not educators choose to assess achievement at one time point or across several time points is related in important ways to basic tenets of cognition and learning. Indeed, developmental issues until recently have been given limited attention in assessments of student achievement. This is in some ways not surprising, as pre-service teachers often do not receive much training in areas of child development. In fact, in the United States, a recent report issued by the National Council on the Accreditation of Teacher Education (Pianta, Hitz, & West, 2010) summarized research indicating that developmental issues are sorely lacking in teacher preparation programmes. Thus, it should come as little surprise that developmental issues and their correspondence to academic growth are not considered seriously as important factors in the assessment of student learning; indeed, issues related to cognitive, social, moral, emotional, and biological development are seldom considered when examining student achievement. Despite the acknowledged importance of developmental issues in learning, static assessments given at one time point are often assumed to equitably assess academic learning.

Students at any given grade level or at any given age, however, are not equal developmentally. Consider a typical classroom of 8-year old children. Some may be reading as well as 11-year old peers, whereas others may be reading at the level of a 5-year old child. Based on an aggregate such as the arithmetic mean, results of an assessment, that is, administered to these children would indicate that some students are clearly below average and others are above average in their reading abilities. Yet, the results tell us nothing about the child's *potential* to learn or the child's *progress* from year to year. Using performance data in this way fails to capture the impact of a child's prior knowledge and skills and does not help to clarify how these change over time.

What would an assessment system that is based on development, or 'growth', tell us? Such a system would provide information about how much a student has progressed between 1 year and the next. That information could be considered simply as a measure of the child's individual progress or in comparison to similarly achieving peers. In addition, repeated administrations of an assessment across several years – as long as the scores from the assessment can be vertically scaled across academic years – can provide educators with information about rates of growth in learning (in addition to absolute learning).

How is growth defined and measured?

It is important to understand the different ways in which growth can be defined and measured, to select an appropriate growth model. All growth involves estimation of change in a construct measured over a span of time. The quality of this estimation is based on assumptions that the construct of interest can be measured on a continuous scale and that the observations obtained across time points or within different groups possess the same measurement properties, so that the interpretation of scores is stable over time and across groups, allowing for meaningful and valid comparisons (Bontempo, Grouzet & Hofer, 2012; Hofer, Thorvaldsson, & Piccinin, 2012).

First, it is important to acknowledge that academic 'achievement' is a broad topic; the term 'achievement' encompasses many concepts (Guskey, 2013). With regard to measuring growth in achievement, it is important to distinguish between growth in *psychometric constructs* and growth in *measured academic performance*. Many researchers in particular employ measures of students' self-beliefs about learning and motivation (e.g., self-efficacy beliefs, goal orientations, expectancies, and values). These are psychological constructs, and from a developmental perspective, researchers are interested in how these self-beliefs change over time. For example, Eccles, Wigfield, and their colleagues have conducted a number of studies examining changes in children's, adolescents', and emerging adults' expectancies and values (e.g., Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002; Wigfield *et al.*, 1997).

In contrast, assessments of academic achievement are focused on measuring students'-specific knowledge and skills in a particular content area. Growth models focus on how such knowledge changes over time; as we will discuss below, many of the newer models allow researchers and policymakers to statistically isolate variables that are related to knowledge growth (e.g., attending a specific school or being taught by a specific teacher). In this study, we focus primarily on growth in academic achievement and not on growth in psychological constructs.

Measuring growth: How many time points are necessary?

The simplest form of academic growth can be estimated through analysis of two-wave change and is defined as the difference (D_i) between an individual's pre-test score ($X_{i,\text{pre}}$) and a subsequent post-test score ($X_{i,\text{post}}$) on a measure for a selected academic outcome: $D_i = X_{i,\text{post}} - X_{i,\text{pre}}$. This definition of change is commonly referred to as 'gain', as it only refers to a simple *increment* or amount of change (positive or negative) rather than to the process involved in obtaining that change (Singer & Willet, 2003). Whereas simple difference scores are an unbiased estimate of true gain (Ragosa, Brandt, & Zimowski, 1982; Zumbo, 1999), their use is limited in academic settings because these gain scores cannot inform us about the shape (e.g., linear, curvilinear, and discontinuous) of an individual's

pattern of change – or trajectory – over time. As a consequence, limited information is garnered about how to influence or capitalize on students' different learning trajectories in order to contribute to improved learning outcomes.

Despite their ease of implementation, true change and measurement error are confounded in two-wave designs (Singer & Willet, 2003; Willett, 1997). Historically, the challenge in separating true score from measurement error led to a series of corrections and estimation strategies that attempted to improve estimation of systematic true change (Willett, 1988). One of these is the modified difference score, which incorporates information about sample or group change (μ_D) and weights the observed individual differences by reliability of the difference scores for the group ($\rho(D)$). Replacing unknown parameters by their sample estimates, the modified difference score is as follows: $\text{mod}\,\hat{D}_i = (\hat{\rho}(D)) \cdot D_i + (1 - \hat{\rho}(D)) \cdot \bar{D}$. Advantages of this approach are that the reliability of the modified difference scores remains unchanged, yet the set of modified gains tend to have smaller variance (mean-squared error or MSE) relative to the original gain scores. In addition, as the transformation is linear, patterns of association between modified gains and external covariates are unchanged from their original counterparts (Willett, 1988). Thus, while this adjustment improves precision of estimates of change for individuals, it has little benefit over original gain scores (Ragosa *et al.*, 1982; Willett, 1988), and still only provides a 'static' estimate of change.

A second alternative utilizes residualized change scores (Zumbo, 1999), although there is debate among methodologists as to their interpretation and meaning (Willett, 1997). Residualized change is the difference between actual scores at time 2 and predicted scores at time 2, when the time 2 scores are regressed on time 1 scores. The strength of the correlation between the scores over time, however, impacts the size of the residualized change scores, and this method of measuring change is subject to the assumptions of linear regression (e.g., normality of residuals and homoscedasticity). Despite these limitations, residual gain scores are still being analysed in research examining student outcomes.

A third-related strategy used to estimate academic growth from two-wave studies involves covariate-adjusted gains. That is, academic gains are estimated after adjusting for pre-test scores, and perhaps other covariates such as socio-economic status, gender, or other individual-level variables. This approach is identical to estimating post-test scores, while adjusting for pre-test scores (and other covariates; Hedeker & Gibbons, 2006). Of concern in this approach is that educational settings are inherently lacking in random assignment of students to classrooms or schools, and thus covariates are not guaranteed to adequately adjust for selection bias due to individual factors or by resource allocation/access differences. Heckman and Rubinstein (2001) and other scholars (e.g., Anderman, Anderman, Yough, & Gimbert, 2010) have pointed out that individual factors such as general ability or student motivation, and effort can contribute to differences in academic achievement and variability in growth. Consequently, adjusted gains controlling for pre-test and individual-level differences can provide information over and above that obtained through simple gains, but should be cautiously interpreted.

According to Raudenbush (2001), observations from two time points provide biased information about differences in growth across groups. Even after adjusting the gains, it is likely that hidden differences across schools or classrooms due to unmeasured variables will contribute additional bias. Thus, models examining the influence of contextual factors on academic growth call for more sophisticated definitions of growth and change than can be articulated through simple two-wave assessments.

A validity assumption underlying the two-wave methods discussed above is that we are accurately capturing the construct of 'academic growth' through the use of gain scores (Willett, 1997). Education scientists and methodologists have long struggled with this issue, and while perspectives may differ on the best methods for estimating change from two-wave studies, there is clear and growing consensus that if growth is to be understood at the level of the individual, information on change over time using more than two points in time is necessary (Hedeker & Gibbons, 2006; McCoach, Madura, Rambo-Hernandez, O'Connell, & Welsh, 2013; Raudenbush, 2001; Singer & Willet, 2003). Gain scores are estimated irrespective of the fact that the underlying change in academic constructs such as motivation or achievement occurs as a continual process, not a static 'jump' that happens at a specific point in time. Thus, estimates of academic growth from two-wave studies are not optimal when the *processes* of change, often the focus for educational psychologists, may be of primary interest. In addition, little advantage is afforded through gain scores and their modified, residualized, or covariance-adjusted counterparts in terms of furthering *theories* of change, a necessary aspect of identifying interventions or strategies that might influence and accelerate academic growth.

Current statistical software is keeping pace with rapid methodological advances regarding growth and change, and there are many options for investigating academic growth when three or more waves of data are available. Among these are individual growth curves, latent growth curve or latent transition models, growth mixture models, and autoregressive or Markov chain models. While more complex than conventional repeated-measures analysis of variance or multivariate analysis of variance models, they maintain distinct advantages, most notably that of clarifying patterns of change rather than isolating degree of change or testing for mean change across distinct groups (O'Connell & McCoach, 2004). In addition, these newer advances to examining growth allow for a strong match between the selected model for change and associated developmental theories on how change may be structured (Ram & Grimm, 2007). Interestingly, latent growth curves can be viewed as the methodological umbrella for these conventional techniques as well as many other complex models used for growth and change. For example, Voelkle (2007) uses respondent scores on a complex learning task across four waves of data to demonstrate the integration of conventional growth models as special cases of latent growth systems. Tomarken and Waller (2005) provide an excellent review on the advantages of multiwave data over two-wave approaches to change. Finally, readers interested in additional demonstrations of the newer, but complex growth models beyond the traditional repeated-measures ANOVA or MANOVA methods are referred to Singer and Willet (2003) or Laursen, Little, and Card (2012). Both of these volumes provide excellent examples of complex approaches to assessing growth.

The individual growth model

Despite the proliferation of models for assessing growth and change, educational environments are among the most complex settings for examining and understanding change. Of the methods available, an appealing and flexible model that emphasizes estimation of individual change, and easily incorporates correlates of change, is the individual growth model (IGM). The IGM is situated in the general framework of multilevel analysis and, as noted above, also falls under the general methodological umbrella of latent growth curve models. Multilevel models are also referred to as HLM or mixed-effects regression models (Hedeker & Gibbons, 2006; Raudenbush & Bryk, 2002). The primary advantage of IGM models lies in their ability to simultaneously capture and

describe each student's individual trajectory over time. These individual trajectories are then compared across a sample of persons to estimate variation in trajectories for the group(s) of interest. Thus, both intra-individual (within-person) and interindividual (between-person) change is assessed. These trajectories are represented in a multilevel model with the simplest form of growth as a linear trajectory, requiring a minimum of three observations per person. At the individual level of the model (level 1), observations at time *t* are predicted from the value of time, which could be measured (typically from the initial marker of 0) in days, months, years, or other desired metrics. Two parameters are estimated for each individual: Linear rate of change, π_{1i}, which represents the linear increase (or decrease) in the outcome for a given value of TIME, and π_{0i}, which represents the expected value when TIME = 0. At level 2, the between-persons level, we obtain a summary or point estimate of each of these parameters across individuals in the sample; variability in the residuals, r_{0i} and r_{1i}, represents variation in the level-1 estimates across the sample.

$$\begin{aligned}
\text{Level 1}: Y_{ti} &= \pi_{0i} + \pi_{1i}(\text{TIME})_{ti} + e_{ti};\\
\text{Level 2}: \pi_{0i} &= \beta_{00} + r_{0i}\\
\pi_{1i} &= \beta_{10} + r_{1i}.
\end{aligned}$$

This form of the IGM is referred to as an 'unconditional growth model', as variation between people is captured through differences in intercepts and slopes, without conditioning on another variable (i.e., such as gender). In this model, e_{ti} represents the occasion-level residuals, that is, the difference between student *i*'s actual score at time *t* from his or her model predicted score ($\pi_{0i} + \pi_{1i}(\text{TIME})_{ti}$) at that time point. The person-level residuals, r_{0i} and r_{1i}, are random effects that represent the deviation of person *i*'s intercept and slope from an overall intercept (β_{00}) and overall slope (β_{10}). Generally, it is the point estimates and the variance components (variances for the random effects and the level-1 residuals) that are presented and interpreted.

Value-added approaches to measuring academic growth

In selecting measures and analyses to assess academic growth, it is important to acknowledge the overall purpose of the assessment. When selecting measures for student growth, there are important issues that must be considered. In a report commissioned by the Gates Foundation, Battelle for Kids (2011) noted the seven most important questions that need to be answered before selecting a growth model for assessing student learning. These include:

1. What are your intended uses and outcomes for growth measures?
2. What information do you have and/or want to include as 'inputs' in the analysis?
3. How does a potential growth model handle measurement error and uncertainty?
4. What types of results, outputs, or information do you want, and in what format?
5. How much and what types of communications, training, and support will you need to be successful?
6. What experience, expertise, and capacity do you need to implement growth measures?
7. What costs are involved in implementing growth measures, and how will you sustain those costs? (Battelle for Kids, 2011, p. 2).

Thus, if student learning is of primary concern, and more than two data points are available, then IGM might be appropriate. However, whereas IGM is a statistical technique for measuring growth in achievement and learning, it has been operationalized in the assessment community through what have come to be known as 'value-added models', which often are used to assess more than just student achievement. Value-added models vary widely and may be either basic or complex; they often employ sophisticated statistical procedures to estimate or make inferences about the quality and influence of effective teaching (Battelle for Kids, 2011). Value-added models are typically used to estimate effects of teachers or schools on student achievement gains (or growth). Relative to the estimated growth of students with similar prior achievement scores, a teacher providing positive value-added has students who have achieved higher than expected achievement growth. The principal challenge to researchers and policymakers lies in understanding the differences, strengths, and limitations of the various value-added models available.

Much of the development and research on value-added models has occurred in the United States. However, value-added models are being used in other nations. For example, value-added models and slight variations on those models have been used in China (Peng, Thomas, Yang, & Li, 2006), Great Britain (Homer, Ryder, & Donnelly, 2011), and Australia (Darmawan & Keeves, 2006). More generally, the use of standardized assessments of student learning is commonplace across many nations, particularly across the European continent (Education, Audiovisual, & Culture Executive Agency, 2009); nevertheless, there is much variation in terms of the purposes of these tests and the timing of the administration of these assessments. In addition, there is little consideration of multiple test scores in making decisions related to students' educational careers in Europe. For example, as of 2009 (the most recent date when large-scale cross-European data on standardized testing was released), data indicated that Malta was the only country that utilized more than one test score in determining whether or not students could move on to the next class level (Education, Audiovisual, and Culture Executive Agency).

Student performance academic growth models that measure whether, or not, positive value is added to an individual student's or group of students' academic performance use both basic and/or advanced statistical approaches to create a prediction of student scores that measure or estimate such an effect. In such cases, a reliable estimate of an educator's effectiveness often appears to be the desired primary outcome, but this is based on performance or growth of their students. Currently, in the United States, value-added methods are used in some states for state reading and math assessments and end-of-course exams, such as ACT Quality Core. However, among those states with high-quality longitudinal data systems necessary for implementation of value-added approaches, states use a variety of models, so caution should be used when comparing results. Many states are tying student growth estimates to teacher effectiveness, and, as we discuss below, this should be performed with extreme caution. Inherently, each model's statistical mechanics may generate error that can both skew and invalidate the teacher effect estimates; this potential for error warrants noteworthy concern when considering how to measure the influence of teaching quality on student academic success.

Types of growth models

Although a number of growth models can be labelled as *value added*, it is important to acknowledge that all growth models provide estimated effects, and some models yield

more reliable estimates than others. The validity of each model's estimated effects may remain questionable. Some estimates are more appropriate to use with certain decisions than others (Hershberg & Robertson-Kraft, 2009), be it for measuring changes in student achievement, predicting future student performance, or estimating teacher and school effects or formulating teacher and school improvement plans. Understanding each value-added model's inputs, limitations, and shortcomings can help policymakers and educators determine which models to employ for different levels of accountability related to student growth and/or measuring teacher effectiveness. We review five general, overarching types of models. These five represent common options that state-level policymakers tend to choose among, when deciding upon models. Here, the United States is selected as a point of focus because various states use different models, thus providing relevant examples. The first three, (1) The Student Gain Score Model, (2) Covariate Adjustment Model, and (3) Student Percentile Gain Models, are considered '*single-wave*' – each model explores, and may predict, changes in academic achievement based on two data entries in a single subject across a single period of time (e.g., fall and spring mathematics scores). Given that each model's statistical mechanics are 'limited to a single wave of two assessment scores, separate analyses are necessary for multiple cohorts and subjects' (Wiley, 2006, p. 24). The latter two models, (4) Univariate Value-Added Response Models, and (5) Multivariate Value-Added Response Models (MRM), consider more than two sets of measurements and are sometimes referred to as 'multiple wave' or layered models. Such approaches use 'multiple assessments to estimate how multiple teachers across multiple years affect student achievement' and 'provide teacher effects estimates based on both reading and math assessments for several groups of students over several years' (Wiley, 2006, pp. 27–28.).

Student gain score model

Accepted as the simplest value-added approach, the gain score model calculates a student's change in achievement while in a specific teacher's classroom. By calculating the difference between two of a student's achievement test scores (typically from consecutive years) and then comparing the average gains of one teacher's students to other teachers' students, within a school or across the same district, the 'teacher effect' can identify the *best* teachers, those teachers whose students show the largest change in achievement test data (year-to-year). As an example, the Texas Growth Index offers an estimate of student performance as well as projected academic growth using two consecutive grade level using scores from the State of Texas Assessments of Academic Readiness test.

While the Student Gain Score Model is transparent, linear, easily computed and can include statistical adjustments for student and school characteristics, several noteworthy shortcomings may invalidate the 'teacher effect' estimates: (1) A typical Student Gain Score Model does not include data from students with missing test scores. Consequently, highly mobile and frequently absent students may have data missing at one of the two time points. (2) Data used from standardized achievement tests may not be vertically aligned; vertically aligned measures are sensitive to both measuring change and to the fact that students' knowledge and abilities change as they develop (Patz, 2007); (3) Student achievement in previous years is not considered, and therefore, this model assumes that 'teacher effects persist undiminished' (Wiley, 2006, p. 21; i.e., both the influence of any prior teachers and students' previous educational experiences [e.g., participation in a preschool programme, or single-gendered classroom, or class size] do

not change over time). Likewise, as test scores used in the model must be appropriately vertically scaled, schools that have changed tests or students taking alternative forms of a test may be excluded (Hershberg & Robertson-Kraft, 2009; Sanders, 2006). As a result, appropriations of teacher effects may be biased (Wiley, 2006). To some extent, applying an analysis of variance model for student gains (i.e., a 'gain as response' model) with classroom as a fixed- or random effect will minimize the instability in standard errors of the classroom gain estimates, but issues from missing data from high absenteeism students still prevail.

Covariate adjustment model

Like the 'single-wave' Gain Score Model, a Covariate Adjustment Model considers change in student achievement across a period of time in a single subject determined by only two points of data measurement. Established in 1992, the Dallas Value-Added Assessment System (DVAAS) ranks among the historic value-added assessment-based accountability programmes (Wiley, 2006). Initially, this model was used to discern differences in school-level performance across the Dallas Independent School District, that is to determine effective schools using a value-added accountability approach. Recently, this model has been used to estimate teacher and school effects on student academic performance for the purpose of designing teacher and school improvement plans. Using different calculations each year (with only two time points), student achievement data are regressed and adjusted for a variety of both student- and school-level characteristics, such as 'ethnicity, gender, language proficiency, student mobility, crowding, socio-economic level, and percentage minority' (Wiley, 2006, p. 26). Unlike the gain score model, the DVAAS, for example, treats the effects of previous teachers differently: Persistence of teacher effects is estimated. Similar to weaknesses of the Student Gain Score Model, students with missing data are excluded, and gain is 'treated the same no matter where this occurs along a development scale' (Wiley, 2006, p. 24).

Student percentile gain models

These models adjust for test scores that are not equally scaled over different years by comparing the percentiles students earned from 1 year to the next. In particular, normal curve equivalent scores can be used to address issues with vertical scaling prior to calculating gains. However, as noted earlier for the Student Gain Score Model, extensive missing data, primarily from frequently absent students, are problematic for estimating a teacher effect. The student growth percentile method (Betebenner, 2009), used by Colorado, includes both a measure of student growth (through individual growth percentile calculations) and a projection value indicating whether students are on track to meet the proficiency performance level in the future.

Univariate value-added response model

This model applies statistical regressions of current test scores on previous test scores to generate a prediction line from which estimated student scores can be compared. Classrooms (or schools) can be included as either fixed- or random effects. This model allows for a comparison between a student's actual score and a predicted score based on the student's testing history and any number of other variables, including demographic characteristics. Importantly, measurement error can be mitigated if at least three prior

student test scores are available (Sanders, 2006), and the model embraces flexibility as to the specific test score (i.e., subject and grade) that can be included; this attenuation of measurement error is a distinct advantage over the gain models.

Multivariate value-added response models

More complex value-added models are multivariate, longitudinal, and sometimes *layered*. For example, models for teacher effects are layered onto those from earlier years (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Schools or teachers can be treated as random effects, and the models have been designed to analyse incomplete data sets and minimize measurement error in value-added estimates (Hershberg & Robertson-Kraft, 2009).

As more data can be included within and across multiple waves, this model tends to produce more reliable estimates and employs sophisticated statistical procedures that control and account for gain at a much more specific level; with these models, researchers can isolate and measure gains that can be attributed to a particular teacher (Tekwe *et al.*, 2004).

In summary, the choice of model has important implications for drawing conclusions about student growth. As we noted, each of the five general value-added growth models described above make assumptions about information that is included in each model's statistical design, and those assumptions notably affect outcomes and applicability. For example, some models treat school effects as fixed – this means they control for the unmeasured aspects of the environments that differ across schools (e.g., differences in class sizes, curriculum materials, availability of instructional supports, or the competence of principals and peers), and adjust for any school-level differences in average teacher quality. School fixed effects essentially compare teachers within schools only. Thus, with school fixed effects in the model, comparisons among teachers in different schools rest on the assumption that on average, all schools routinely hire equally capable teachers, an assumption that is unsupported in many cases (Newton, Darling-Hammond, Haertel, & Thomas, 2010). In contrast, value-added models that treat school effects as random control for unmeasured aspects of the environment and contain the assumption that not all teachers across a district or a region are equally capable.

Discussion

Educators, researchers, and policymakers would benefit greatly from understanding the nuances of the assessment of growth in student learning, particularly as many curricular-, policy-, and personnel-related decisions may result from interpretations of growth scores. In the present study, we have reviewed the reasons why it is important to consider measurements of growth in student learning, some of the basic assumptions of simple versus more complex growth models, and some of the ways that these techniques are being used in the current high-stake testing climate (e.g., value-added models).

One of the most important conclusions that we can draw from our review is that the nature of the assessment of growth that is selected truly matters (Herman, Heritage, & Goldschmidt, 2011). If an inappropriate or psychometrically unreliable or invalid assessment is used to measure growth, serious consequences may occur, particularly in terms of making decisions about an individual student's progress, and about school personnel decisions (i.e., whether or not a teacher is being effective). The potential

injustices of inappropriate educational decisions being made for individual students, and of teachers and other educators losing their positions because of the psychometric shortcomings of assessments (e.g., the inappropriate use of assessments designed for use at one time point) are genuine.

For example, Ready (2013) notes that although measures of student growth are now quite sophisticated, we still know relatively little about the relations between students' initial achievement and subsequent achievement growth (i.e., do high achieving students exhibit growth at faster rates than do lower achieving students?); such relations may affect the validity of the results generated from these models. Using data from one state in the US, Ready recently demonstrated that in both literacy and mathematics, growth was greater in students with initially low scores than in students with higher scores (Ready, 2013). Thus value-added data that are used to evaluate schools may be biased, as students are not randomly assigned to schools (i.e., there may be more high- or low-achieving students in some schools compared with others).

In addition, many assessment tools are designed to measure knowledge that has been learned during a particular grade level, within a particular subject domain; for example, an assessment might focus on eighth grade mathematics or third grade reading comprehension. As Ravitch (2010, p. 153) notes, 'Testing experts also warn that test scores should be used only for the purpose for which the test was designed: For example, a fifth-grade reading test measures fifth-grade reading skills and cannot reliably serve as a measure of the teacher's skill'. Nevertheless, an important issue that must be seriously considered in the assessment of growth is the fact that *many instruments were simply not designed to measure growth* and thus, may be inappropriate for use in such ventures (Herman *et al.*, 2011). Indeed, growth models can be very powerful tools when they are used for their intended purposes; however, when models are used inappropriately (e.g., models specifically designed just to measure student achievement growth being used to evaluate teacher performance), then both researchers and policymakers should be seriously concerned.

Recently, the American Statistical Association (ASA) issued a public statement about the use of value-added models in educational assessment. Although the ASA in general endorses the use of statistical models to improve education, they offer several important cautions. For example, they note that high levels of statistical expertise are needed to interpret data from these models; estimates generated from value-added models always should include measures of precision; value-added models measure correlation, not causation; and scores and ranking generated from value-added models can change dramatically when different types of models are employed (American Statistical Association, 2014). Thus, the use of these data to make decisions about students and teachers must be carried out very cautiously and carefully.

In response to public critiques of the use of standardized test data to measure teachers' influence on growth in student learning, emerging evidence has suggested alternative student growth measures may reliably identify highly effective teachers (Gill, Bruch, & Booker, 2013). Rather than applying value-added analysis, simple or otherwise, to commercially available tests and/or end-of-course assessments limited only to certain grades and content areas, a different approach allows teachers to propose locally developed curriculum assessments to assess student growth. This is becoming particularly important, given the more prominent use of classroom-based assessments (e.g., self-assessments) to monitor progress in a formative manner (Andrade & Valtcheva, 2009). Student learning objectives (SLOs) may be described as 'classroom-specific growth targets that are chosen by individual teachers and approved by principals as an alternative

measure of student learning' (Gill *et al.*, 2013, p. ii). A rapidly expanding advocacy by the general education community in support of the alternative assessment of both student learning and teacher effectiveness in school districts is heralded by a clear advantage of SLOs that resides in their widespread applicability for evaluating any teacher in any grade or content area. Although SLOs may potentially discern one teacher's influence on student growth from another better than state-generated or commercially available assessments, studies to date have not explored the reliability of SLOs (Gill *et al.*, 2013). Due diligence is required if SLOs are to be used for measuring growth in learning, instructional planning, and high-stake teacher evaluations. To understand whether or not SLOs may better distinguish teacher effectiveness, researchers will need to test the relations between SLO ratings and standardized assessment-based (value-added) growth measures (Gill *et al.*, 2013).

There also still is much to learn about the use of growth models in education. This work is nascent, and many new developments will occur in the coming years. For example, little attention has been paid thus far to the timing of measurements in value-added models. Whereas most high-stake assessments are administered annually, which may not always be the case. Assessments that are given earlier in the academic year will yield different types of results than assessments administered later in the year, since the former often will assess learning that occurred during the prior school year, whereas the latter often will focus on what has been learned during the current school year. Although MRM models can statistically account for different timings of test administration, this issue has not been seriously considered to date.

Conclusions

In this study, we have demonstrated that there are important advantages to measuring student growth, as opposed to measuring static achievement at one time point. As we stated at the beginning of this study, a focus on growth and development is clearly in line with the basic tenets of educational psychology; human learning is not a static process, and cannot and should not be measured as such. Nevertheless, the assessment of growth is highly complex; there are many important decision points involved in assessing growth (e.g., type of instrument to be used, number of measurements, time between measurements, analytic techniques, etc.), and the importance of each of these should not be underestimated. Measuring growth affords educators enormous advantages and extremely powerful information, but reliable and valid measurement rests on the many assumptions that we have reviewed in this study.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling in school effectiveness studies. *Journal of the Royal Statistical Society, Series A*, *149*, 1–43.

American Statistical Association (2014). *ASA statement on using value-added models for educational assessment*. Alexandria, VA: Author. Retrieved from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Anderman, E. M., Anderman, L. H., Yough, M. S., & Gimbert, B. G. (2010). Value-added models of assessment: Implications for motivation and accountability. *Educational Psychologist*, *45*, 123–137. doi:10.1080/00461521003703045

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory into Practice*, *48* (1), 12–19. doi:10.1080/00405840802577544

Battelle for Kids (2011). *Factors to consider when selecting measures of student growth: A guide for education leaders*. Columbus, OH: The Bill & Melinda Gates Foundation.

Betebenner, D. (2009, April 6). *Growth, standards, and accountability*. The Center for Assessment. White paper.

Bontempo, D. E., Grouzet, F. M. E., & Hofer, S. M. (2012). Measurement issues in the analysis of within-person change. In J. T. Newsom, R. N. Jones & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences* (pp. 97–142). New York, NY: Routledge/Taylor & Francis.

Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *The Teachers College Record*, *106*, 429–458. doi:10.1111/j.1467-9620.2004.00346.x

Coleman, J. S., Campbell, T. E., & Kilgore, S. B. (1982). *High achievement: Public, catholic, and other private schools compared*. New York, NY: Basic.

Darmawan, I., & Keeves, J. P. (2006). Accountability of teachers and schools: A value-added approach. *International Education Journal*, *7*, 174–188.

Education, Audiovisual, and Culture Executive Agency (2009). *National testing of pupils in Europe: Objectives, organisation, and use of results*. Brussels, Belgium: Author. Retrieved from http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/109EN.pdf

Educational Testing Service (2013). *Seeing the future: How the common core will affect Mathematics and English language arts in grades 3–12 across America*. NJ: Author. Retrieved from http://www.k12center.org/rsc/pdf/seeing_the_future.pdf

Ercikan, K. (2006). Development in assessment of student learning. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 929–952). Mawhah, NJ: Lawrence Erlbaum.

Finn, C. E., Ryan, T., & Partin, E. L. (2008). *Ohio value-added primer: A user's guide*. Washington, DC: Thomas Fordham Institute.

Gill, B., Bruch, J., & Booker, K. (2013). *Using alternative student growth measures for evaluating teacher performance: What the literature says (REL 2013–002)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Guskey, T. (2013). Defining student achievement. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 3–6). New York, NY: Routledge.

Heckman, J. J., & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the GED testing program. *The American Economic Review*, *91*, 145–149.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: Wiley.

Hershberg, T., & Robertson-Kraft, C. (2009). *A Grand Bargain for education reform: New rewards and supports for new accountability*. Cambridge, MA: Harvard Education Press.

Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). *Developing and selecting assessments of student growth for use in teacher evaluation systems*. Los Angeles, CA: WestEd.

Hofer, S. M., Thorvaldsson, V., & Piccinin, A. M. (2012). Foundational issues of design and measurement in developmental research. In B. Laursen, T. D. Little & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 788–790). New York, NY: Guilford Press.

Homer, M., Ryder, J., & Donnelly, J. (2011). The use of national data sets to baseline science education reform: Exploring value-added approaches. *International Journal of Research & Method in Education*, *34*, 309–325.

Jacobs, J. E., Lanza, S., Osgood, D., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development*, *73*, 509–527. doi:10.1111/1467-8624.00421

Laursen, B. P., Little, T. D., & Card, N. A. (Eds.) (2012). *Handbook of developmental research methods*. New York, NY: Guilford Press.

Lee, J., & Fish, R. M. (2010). International and interstate gaps in value-added math achievement: Multilevel instrumental variable analysis of age effect and grade effect. *American Journal of Education*, *117* (1), 109–137. doi:10.1086/656348

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.

McCoach, D. B., Madura, J., Rambo-Hernandez, K. E., O'Connell, A. A., & Welsh, M. (2013). Longitudinal data analysis. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 199–230). Rotterdam, The Netherlands: Sense Publications.

McCoach, D. B., Rambo, K. E., & Welsh, M. (2013). Assessing the growth of gifted students. *Gifted Child Quarterly*, *57*(1), 56–67. doi:10.1177/0016986212463873

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education policy analysis archives*, *18*(23), 1–24.

O'Connell, A. A., & McCoach, D. B. (2004). Applications of hierarchical linear modeling for evaluation of health interventions: Demystifying the methods and interpretation of multilevel models. *Evaluation and the Health Professions*, *27*(2), 119–151. doi:10.1177/01632787042 64049

Olson, L. (1998). A question of value. *Education Week on the Web*, *17*. Retrieved from http://www.edweek.org/ew//vol-17/35value.h17. Cited in Tekwe *et al.*

Patz, R. J. (2007). *Vertical scaling in standards-based educational assessment and accountability systems*. Washington, DC: The Council of Chief State of School Officers.

Peng, W., Thomas, S. M., Yang, X., & Li, J. (2006). Developing school evaluation methods to improve the quality of schooling in china: A pilot "value-added" study. *Assessment in Education: Principles, Policy and Practice*, *13*, 135–154.

Pianta, R. C., Hitz, R., & West, B. (2010). *Increasing the application of developmental sciences knowledge in educator preparation: Policy and practice issues*. Washington, DC: National Council for Accreditation of Teacher Education.

Ragosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*, 726–748. doi:10.1037/0033-2909.92.3.726

Ram, N., & Grimm, K. (2007). Using simple and complex growth models to articulate developmental change: Matching theory to method. *International Journal of Behavioral Development*, *31*, 303–316. doi:10.1177/0165025407077751

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, *52*(1), 501–525. doi:10.1146/annurev.psych. 52.1.501

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

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York, NY: Basic Books.

Ready, D. D. (2013). Associations between student achievement and student learning: Implications for value-added school accountability models. *Educational Policy*, *27*(1), 92–120. doi:10.1177/ 0895904811429289

Sanders, W. (2006). *Comparison among various educational assessment value-added models*. White paper. Presented at The power of Two – National Value-Added Conference, Columbus, OH.

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

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., . . . Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment and school performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 11–36. doi:10.3102/ 10769986029001011

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, *1*, 31–65. doi:10.1146/annurev. clinpsy.1.102803.144239

Voelkle, M. C. (2007). Latent growth curve modeling as an integrative approach to the analysis of change. *Psychology Science*, *49*, 375–414.

Wigfield, A., Eccles, J. S., Yoon, K., Harold, R. D., Arbreton, A. A., Freedman-Doan, C., & Blumenfeld, P. C. (1997). Change in children's competence beliefs and subjective task values across the

elementary school years: A 3-year study. *Journal of Educational Psychology*, *89*, 451–469. doi:10.1037/0022-0663.89.3.451

Wiley, E. W. (2006). *A practitioner's guide to value-added assessment*. Retrieved from http://nepc.colorado.edu/files/Wiley_APractitionersGuide.pdf

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, pp. 345–422). Washington, DC: American Educational Research Association.

Willett, J. B. (1997). Measuring change: What individual growth modeling buys you. In E. A. Amsel & K. A. Renninger (Eds.), *Change and development: Issues of theory, method, and application* (pp. 213–243). Mahwah, NJ: Lawrence Erlbaum.

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 269–304). Bingley, UK: Emerald Group.

Received 14 February 2014; revised version received 1 August 2014