Research Report

A Summary of Models and Standards-Based Applications for Grade-to-Grade Growth on Statewide Assessments and Implications for Students With Disabilities

Heather M. Buzick

Cara Cahalan Laitusis

June 2010

ETS RR-10-14

A Summary of Models and Standards-Based Applications for Grade-to-Grade Growth on Statewide Assessments and Implications for Students With Disabilities

Heather M. Buzick and Cara Cahalan Laitusis
ETS, Princeton, New Jersey

June 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Dan Eignor

Technical Reviewers: Wendy Yen and Dianne Henderson-Montero

Abstract

Recently growth-based approaches to accountability have received considerable attention because they have the potential to reward schools and teachers for improving student performance over time by measuring the progress of students at all levels of the performance spectrum (including those who have not yet reached proficiency on state accountability assessments). While the use of growth in accountability holds promise for students with disabilities, measuring changes over time in their academic performance is complex. This paper summarizes models and approaches that use individual student test scores from multiple years for 3 different purposes: determination of adequate yearly progress under the federal accountability system, research on individual growth trajectories, and evaluation of the contribution of teachers and schools to student learning. Practical challenges in measuring and modeling growth for students with disabilities are then discussed. Finally, 3 areas in need of research on the measurement of growth from large-scale annual accountability assessments are identified and described: testing accommodations, test difficulty, and understanding the longitudinal characteristics of the population of students with disabilities.

Key words: growth models, value-added, longitudinal, accountability, students with disabilities, NCLB, ESEA, state assessments

Acknowledgments

The authors would like to thank the following people for their help and comments: Henry Braun, Linda Cook, Matthias von Davier, Dianne-Henderson Montero, Wendy Yen, Lu Ying, and John Young.

Table of Contents

Impending reauthorization of the Elementary and Secondary Education Act of 1965 (ESEA) has cultivated interest in exploring ways to allow states more flexibility in the federal accountability system. One area of particular interest is the measurement and use of student academic growth, specifically changes in test scores from grade to grade for individual students on statewide standards-based assessments. Aggregates of individual student growth across schools, districts, and a state can be used to support decisions and actions that hold schools and teachers accountable for providing students with a high quality education.

The last reauthorization of ESEA (the No Child Left Behind Act of 2001 [NCLB]) introduced requirements for all students to be proficient in reading and mathematics by 2014 and aimed to close gaps in achievement between demographic subgroups including students with disabilities. According to the original intent of the NCLB legislation, schools would receive credit under federal accountability guidelines for their students making adequate yearly progress (AYP) according to a status model. Under the status model, the percentage of students who are proficient based on the current year's test scores is compared to the state's annual targets. Schools failing to make AYP using the status model would be able to use an improvement model. The improvement model (referred to as a safe harbor provision) gives credit to schools that decrease by at least 10% from one year to the next the percentage of students in a particular subgroup who are not proficient. Schools could also use an index model to receive partial credit for student subgroups scoring below proficient if the current year students were to make gains (e.g., below basic to basic) over the previous year students in the same grade. Under the original NCLB legislation, federal accountability approaches that use changes over time do not necessarily compare the same students.

Status and improvement models under NCLB have been criticized for not valuing the academic improvement of students at each possible proficiency level and for perceived unfairness to schools with students who are low performing at a given point in time because of factors beyond the control of the teacher, school, or district. In response to these concerns, two programs issued by the U.S. Department of Education were initiated to offer more flexibility in the federal accountability system and to help students and schools with students who are low performing upon enrollment. The Growth Model Pilot program, initiated in 2005, allows states to take into account individual student test scores from the previous one or more years when comparing current or projected scores to a growth target for making AYP decisions. As intended,

schools would receive credit for their contribution to learning and improved achievement for their students, including those who are initially underperforming. The Differentiated Accountability pilot program (2008), which differentiates among schools that are in need of improvement at various levels, allows a school that did not meet AYP to target interventions directly toward areas most in need. This program reflects an interest in finding more effective ways of improving education in particular for students who are low performing and a desire to make the implementation of the federal accountability system more practical (Mintrop & Sunderman, 2009). The existence of both programs reflects a shift toward valuing academic growth and a focus beyond the average performance of a school, district, or state. This shift is important for students with disabilities, who are receiving standards-based instruction and participating in federal accountability assessments in greater numbers but are still performing (on average) well below their nondisabled peers (Center on Education Policy [CEP], 2009). In addition, the use of academic growth measures may help schools with large numbers of students with disabilities that fail to meet AYP because of the lower performance of students with disabilities on state accountability assessments (see Eckes & Swando, 2009, for AYP results for three states).

Measures of change over time in scores from statewide standards-based assessments can be used not only for determining AYP but also to address policy questions that contribute to improved educational outcomes and to make decisions about teacher and school accountability. It is evident that the pending reauthorization of ESEA has created a culture that supports rewarding schools and teachers for the academic growth of their students across all levels of the proficiency spectrum and has placed value on longitudinal approaches that use data from large-scale statewide accountability assessments. This movement toward using growth measures can also be seen in the proposed guidelines for the Race to the Top Fund program (U.S. Department of Education [USED], 2009), whereby the USED provided support for assessments designed to measure growth and longitudinal data systems with the ability to link teacher and student data. In addition, the Institute of Education Sciences (IES) issued the Statewide Longitudinal Data System (SLDS) Grant Program in 2002 to assist states in collecting more individual-level longitudinal data from large-scale, statewide assessments to support longitudinal models (National Center for Educational Achievement, 2008).

Organizations that advocate for students with disabilities have also become more interested in the measurement of growth and its use in federal accountability. For example, the Council on Exceptional Children (CEC) issued a statement in support of growth-based approaches for accountability and value-added[1] models (VAMs) while calling for more research and funding for states to continue exploring models that use changes in performance over time (CEC, 2009). In their 2009 public policy agenda, the American Speech-Language-Hearing Association (ASHA) listed as one of their priorities to "promote use of a growth model to assess students with disabilities and meet the requirements of adequate yearly progress (AYP) under the reauthorization of NCLB, and consider legislative or regulatory remedies for inconsistencies created between NCLB and [the Individuals with Disabilities Education Act]" (ASHA, 2009, Issue Objective, para. 3). Recently, the National Center on Learning Disabilities convened a growth model task force (GMTF) of university measurement experts and disability policy and advocacy centers to develop a policy paper on the use of growth measures. The GMTF recommended that federal accountability policy include approaches that use measures of individual academic growth in combination with status models for AYP, provided that they fully include students and that they count all students in the same way (GMTF, 2009).

While AYP determinations based on growth have the potential to benefit schools with students who are initially low performing and growing over time, including some students with disabilities, several important challenges pertain to students with disabilities that must be addressed. These challenges may impact growth-based interpretations and the decisions supported by models that use longitudinal test scores from students with disabilities. This paper will (a) summarize models for longitudinal data and growth-based approaches to accountability, (b) discuss characteristics of students with disabilities and the assessments that may impact the measurement of growth and growth-based interpretations, and (c) identify areas for future research on growth-based applications using test scores from students with disabilities.

A Summary of Models for Academic Growth

Measures of student academic growth can be used for AYP determinations (i.e., growth models), for research on individual growth trajectories (see, for example, Morgan, Farkas, & Wu, 2009) which can theoretically be used for making formative decisions about learning and also to determine the appropriateness of policy-driven growth targets (i.e., growth curve models),

and for evaluations of teachers and schools for educational effectiveness (i.e., VAMs). Below is a summary of longitudinal models and growth-based applications for these purposes and a discussion of theoretical advantages and limitations for students with disabilities. Practical limitations of modeling growth for students with disabilities are described in the following section on measurement challenges. The model summary herein is organized in three sections based on model purpose: (a) growth-based approaches for determining AYP; (b) models for individual growth trajectories; and (c) growth-based models for teacher and school accountability. Preceding the model summaries is a brief description of vertical scaling, a statistical technique that is required for some models that use test scores from multiple grades, particularly when describing or inferring about the academic growth of individual students.

Vertical Scaling

Vertical scaling refers to the process of placing scores from two or more tests with different difficulty levels but similar test content on the same scale in order to measure grade-to-grade academic growth (see Kolen & Brennan, 2004, for a technical discussion; see Lissitz & Huynh, 2003, and Yen, 2007, for applications-oriented discussions). Assessments that are vertically scaled are intended to support inferences about the academic progress of individual students over time. In addition, change scores from vertically scaled assessments can be aggregated for use in school, district, or state level growth-based AYP calculations. However, in practice, vertically scaled scores have limitations that can adversely impact the validity of inferences drawn from them. Specifically, vertical scaling is a more complex linking process than equating parallel test forms because both the content and the statistical specification of assessments typically change across grades (see Kolen & Brennan, 2004, for a discussion of linking error and violations of dimensionality in vertical scaling). In addition, numerous statistical strategies can be used to create a vertical scale, and the manner in which it is created can change the interpretation of results that use scores from more than one grade (e.g., Briggs & Weeks, 2009). Due to such complexities and the need for additional resources required to link assessments across grades, not all standards-based state assessments are vertically scaled. Some of the longitudinal methods described below require vertically scaled scores while others do not; this information is indicated for each of the models described in the following three sections that summarize models for academic growth.

Growth-Based Approaches for Determining Adequate Yearly Progress (AYP)

The federal Growth Model Pilot program (2005) has permitted states to count students who are *on track to proficiency* as proficient for the purpose of making AYP decisions. The term *growth models* is being used in the federal accountability system as a label for approaches that take into account individual student test scores from the previous one or more years when comparing current or projected scores to a growth target. Individual student growth is aggregated across schools, districts, and the state to make standards-referenced AYP decisions (see Slater, Wentzel, & Chard, 2009, and Council of Chief State School Officers [CCSSO], 2009, for summaries of specific approaches used by states for making growth-based AYP determinations). All models used for accountability under the current NCLB legislation hold all students to the same standard of being proficient by a fixed point in time (i.e., growth to proficiency), consequently keeping high expectations for instruction and achievement. Growth models can be described in three categories: value tables, an observed growth trajectory compared to intermediate growth targets, and projection models. Models in the first two categories are descriptive approaches to using longitudinally linked individual student test scores, whereas projection models are statistical models that are used to obtain estimates of future scores based on current and past test scores linked within individual students. The three categories are described below.

Value tables. Value tables, used by Delaware and Minnesota, are a descriptive approach for determining AYP whereby numerical values are assigned to transitions in academic performance across two consecutive years (Hill, Gong, Marion, DePascale, Dunn, & Simpson, 2006). Human judgment is used to assign values to the proficiency level transition matrix. For example, a value of 100 may be assigned for moving from below proficient in the previous year to proficient in the current year. The values from the table can then be aggregated across students in a school to form an index for accountability. As an example, Delaware's value table is illustrated in Appendix A. Most value tables are used for AYP, a *school-level* accountability measure, and are intended to hold all students to the same standard of reaching proficiency by a chosen point in the future. Value tables do not produce growth trajectories and do not require statistical modeling or a vertically scaled assessment (Yen, 2008).

The use of a value table for AYP calculations is transparent to stakeholders since there is no statistical modeling. Growth can be valued differently in different grades, affording the

potential to align values with results from theoretical and empirical research on natural growth trajectories. However, this would have high costs in terms of time and money to implement, and as such, states that use this approach would likely have one value table for all grades (as is the case with Delaware and Minnesota). Students with disabilities taking the modified assessment[2] (and possibly the alternate assessment) could be combined with students taking the general assessment via standard setting and alignment of standards. However, value tables only use one prior year of scores for one content area so they are not likely to be accurate for making individual student-level decisions, particularly for low performing students. In addition, value tables cannot be used to predict future growth for use in program evaluation or instructional intervention. Finally, values tables rely on two levels of human judgment (to set proficiency cut scores and to assign values to transitions); as such, they are highly subjective and cannot be evaluated empirically for accuracy. Other limitations depend on the exact method that a state uses (the reader is referred to Slater et al., 2009, and CCSSO, 2009, for some discussion of limitations of specific state models for all students).

Observed growth trajectory compared to intermediate growth targets. Six states (Alaska, Arkansas, Florida, Iowa, Missouri, and North Carolina) have approached growth-based accountability by choosing a criterion at the end of a time horizon and creating intermediate growth targets with which to compare observed scores (vertically scaled or standardized in a baseline year) across time. The methods that are being used by states differ in how growth targets are calculated, the chosen time horizon, and the final grade included in the model (Dunn & Allen, 2009). An exemplary model from Alaska is illustrated in Appendix B; the reader is again referred to CCSSO (2009) for specific state-by-state details. In such models, individual students will have different growth targets for each year prior to the final year based on the student's initial score. This approach can be used for both aggregate (e.g., school-level) accountability and student-level decisions.

Comparing an observed growth trajectory to interim growth targets is less subjective than the use of a value table. However, the intermediate growth targets for individual students are derived from the student's score in the baseline year; as such, intermediate growth targets may be influenced by sources of construct-irrelevant variance or accommodation use that impact the baseline test score. In addition, scores from multiple years must be on the same or a comparable scale in order to compare them to the growth targets, so some students may not be included in

the model (e.g., a student who takes the general assessment in the baseline year and the modified assessment in the following two years with no link between assessments). Finally, given that the shape of the trajectory formed by the intermediate growth targets is policy driven, the choice (e.g., linear) may be inconsistent with the growth of students, including some students with disabilities, who may grow very little in some grades and more in others.

Projection models. This approach projects whether students are on track to become proficient by a certain point in the future. Statistical models can be employed to predict future performance based on past performance and include regression, multilevel modeling,[3] and latent variable modeling. For AYP decisions, the predicted values are compared to proficiency cut-scores. Student characteristics are not permitted to be included in statistical models under NCLB; as such, only unconditional forms or models conditioned on academic variables such as previous test scores can be used for determining AYP. The models, as permitted under NCLB, are described below.

Linear regression. The classical linear regression model estimated with ordinary least squares (OLS) relies on the assumption of normality to estimate an average score conditional on previous test scores. A single linear regression line can be used to predict a future score for each individual student conditional on scores or proficiency levels from one previous assessment cycle (i.e., univariate linear regression) or multiple prior assessments (i.e., multivariate linear regression). A vertical scale is not required. The linear regression model assumes that the rate of growth is the same for all subgroups of students, which may be untenable, particularly for some students with disabilities. As such, examination of systematic bias of model predictions should be conducted prior to implementation and periodically during the course of model use.

The benefit of a statistical model over descriptive approaches is that it yields a standard error of measurement that can be used to determine the precision of estimates and the appropriateness of inferences supported by test scores. This benefit is particularly important for comparing models across subgroups and evaluating the performance of models for some students with disabilities whose single year test scores are less precise because they are low performing on a statewide assessment. Multivariate linear regression is more accurate than univariate regression because it uses more information, which may reduce the influence of measurement error from sources of construct-irrelevant variance such as barriers to access. However, more data points are required, which can be costly and increases the risk of missing data. While the

linear regression model is a relatively simple statistical model to understand, a limitation over some other statistical models is that parameter estimation depends on the assumption that the data arise from one population with the same rate of linear growth. As such, it does not address changes in growth trajectory that may be influenced by testing accommodations, proficiency level, or disability subtype.

Quantile regression. Quantile regression describes a family of models, with each model defined by a specific percentile and its own estimated rate of change. Quantile regression does not require scores from a vertically scaled assessment. Quantile regression is not based on normal theory like OLS; as such, it is less dependent on model assumptions (e.g., the model can accommodate a larger number of students at the extremes of the proficiency distribution, an asymmetric distribution of scores is permissible, and multiple modes can exist). Once the family of conditional quantile functions is estimated, conditional growth percentiles can be calculated to help understand academic progress compared to peers with a similar growth trajectory and can also be used to predict future performance (Betebenner, 2009). *Student growth percentiles*, which employ quantile regression, are used for AYP determinations by Colorado and are being considered by other states. While an approach using student growth percentiles has the same limitations as OLS regression in terms of the inability to model the impact of accommodations and disability subtypes on academic growth, it offers a benefit over a single population linear regression model by estimating multiple growth trajectories that are not necessarily linear and may better reflect true academic progress (see Betebenner, 2009, p. 11 for an illustration).

Multilevel modeling. Multilevel modeling (see, for example, Raudenbush & Bryk, 2002) describes a category of models that take into account the nested structure of data (e.g., students nested within classrooms within schools). Arizona, Ohio, Tennessee, and Texas use such models, which include several variables in the prediction of a future expected score or proficiency level. Among these may be several scores from previous years' assessments, as well as additional covariates. Such forms of multilevel models do not require a vertical scale. Taking into account the nested structure of data is appropriate if students in the same classroom, for example, have similar characteristics. The forms of the multilevel models approved for use in AYP determinations, like linear regression, are not targeted at modeling or detecting differences in growth trajectories across subgroups of students. They do not address changes in growth trajectory based on testing accommodations or modifications, or changes based on proficiency

level or disability subtypes. However, multilevel models have the capability to answer substantively interesting questions about how growth is impacted by these student characteristics (described in the next section). Time points can also be treated as nested within individuals in multilevel models provided that the scores are from a vertically scaled assessment. The impact of student-level and aggregate (e.g., school-level) characteristics on estimates of average initial status and rate of growth can be modeled with such a specification. However, without several points in time and scores from many individual students with disabilities, the benefits of multilevel models over a simpler model would not likely be realized.

Latent growth curve modeling. Latent variable models are not currently used by any states in making AYP decisions; however, the unconditional latent growth curve model (see Hancock & Choi, 2006), which is appropriate for continuous longitudinal test scores from vertically scaled assessments, can be used to estimate growth trajectories and predict individual student growth. A latent growth curve model combines repeated measures from discrete, observed time points into a set of multiple indicators that is subdivided by sections that represent time points at which the assessment took place. These repeated measures are then regarded as indicators of one (or more) underlying common factor(s) or student variable(s). The model provides estimates of average initial status and rate of change that can be used to estimate individual growth trajectories. However, this model requires at least three time points and a large sample size, which would make it an infeasible model for most states.

It is highly likely that growth models will continue to be an option for determining AYP when ESEA is reauthorized. While the intention of using growth models for AYP is to recognize the achievement of initially low performing students, some school-level AYP decisions from growth models have been shown to be similar to results from status models (Tong & O'Malley, 2006; Weiss, 2008). When proficient students were not included, growth model results were shown to measure growth to proficiency (Dunn & Allen, 2009), a result that, while derived from a model not permitted by current federal guidelines, would be desirable for some students with disabilities who are initially low performing. In addition, as described in more detail below, there are several challenges when measuring academic growth for students with disabilities that may have an impact on the inclusion of some students with disabilities in growth models for AYP and the interpretation of results from growth models. It is important that these challenges be considered when making AYP decisions based on growth.

The methods described above that are currently being used for AYP determinations are not permitted to include student demographic characteristics by law. This limitation may be especially problematic when using longitudinal test scores from a complex population such as students with disabilities. While the goal of all students achieving proficiency is laudable, the application of growth models as defined by NCLB has been criticized for setting unrealistically high expectations because of the high rate of growth low scoring students must demonstrate to meet growth targets (Dunn & Allen, 2009).

Longitudinal test scores can be used for educational improvement in other ways that extend beyond AYP determinations. The models described below are more sophisticated and take into account more of the information that may be provided in a longitudinal dataset and theoretically can be used for research to inform instructional decisions or employed to study the natural academic growth trajectories of students.

Models for Individual Growth Trajectories

Models that can be used to estimate individual growth trajectories are generally referred to as growth curve models. In growth curve modeling, the functional form of growth is assumed to be the same for all individuals (e.g., linear, quadratic) and individual growth trajectories can be described by a single average initial status and average growth rate with random effects or measurement error representing variation in growth due to individuals, classrooms, or schools, depending on the specification of the model. Growth curve models can be specified as multilevel models with individuals nested within time points (e.g., longitudinal mixed-effects, Raudenbush & Bryk, 2002) or they can be specified within a structural equation modeling (SEM; Jöreskog, 1977) framework (i.e., latent growth curve models, Bollen & Curran, 2006). Traditional latent growth curve models can be extended to multigroup or mixture modeling, which allow for different subgroups to have different average initial status scores and average growth rates. Multilevel longitudinal models can also be specified to estimate different average growth rates and initial status scores for different subgroups of students.

Under ideal conditions, growth curve models can answer questions such as when did growth begin, what is the growth rate for a given individual, and what is the predicted future score for an individual. This family of models can accommodate variables such as characteristics of students with disabilities including disability type and severity, accommodation use, and

instructional setting. The SEM approach to growth curve modeling allows for flexibility in specifying measurement error, which can be an advantage for measuring the growth of complex populations, such as students with disabilities. The multilevel approach to growth curve modeling is able to accommodate unbalanced measurement occasions across individual students, a benefit for modeling longitudinal test scores from students with disabilities who may have missing values in some years due to, for example, moving between the general assessment and the modified assessment.

Theoretically, growth curve models are useful for determining if students with different disability classifications or accommodation use differ in their initial status, in addition to the rate and shape of their growth. However, to obtain the most information, the number of measurement occasions and the accuracy of those measurements must be high, a characteristic that may be difficult to obtain from students with disabilities, particularly for students who use an accommodation in which it is unclear whether it introduced construct-relevant or irrelevant variance. Sample size requirements are the highest for mixture modeling, which may require combining students based on statistical necessity instead of substantive considerations.

Growth curve models assume that scores arise from the same or a closely related skill. As such, to apply a growth curve model to scores from different grades, the assessments are required to be vertically scaled. Nonetheless, the extent to which the assumption of skill invariance over time holds for students with disabilities relative to the general population is unclear. It is likely that since the measurements from this population are more complex, it will be harder to obtain data that has the desired measurement properties sufficient for making meaningful inferences.

The benefit of applying growth curve models to scores from state standards-based assessments is that we can learn about the natural growth trajectories of students with disabilities in order to inform expectations about growth and develop growth targets for accountability that remain high yet are more realistic than those currently established in accountability systems. For example, research could be conducted by applying mixture modeling to longitudinal models in order to differentiate groups in terms of their growth trajectory and ultimately to aggregate groups that are meaningful and practical while ensuring inferences supported from the results of the model are appropriate. While these more sophisticated models may hold promise, the data requirements may be difficult to fulfill when using annual statewide assessments, especially for

students with disabilities whose assessments usually have measurement properties that are more complex than those for students without disabilities.

Growth-Based Models for Teacher and School Accountability

The methods described in the previous sections were for making AYP decisions based on a fixed criterion for federal accountability and for modeling empirically and theoretically derived individual growth trajectories. This section describes an approach that uses evidence from students' academic growth to evaluate teacher or school performance, called value-added modeling (Braun, 2005; see also Braun, Chudowsky, & Koenig, 2010, for measurement and policy issues). Value-added describes a family of models that use student-level test score data to estimate the contributions of schools or teachers to student academic growth. VAMs require data across two or more years, class-level data, and student-level data be matched to teacher- or school-level data. Student characteristics such as accommodation use or disability status can be included in these models; however, since little is known about the impact of disability status or accommodation use on student growth, research is needed on the appropriate specification of VAMs for teacher and school accountability.

VAMs do not yield individual growth trajectories but can be layered on top of statistical models that aim to parsimoniously model estimates of growth parameters and their standard errors. As such, they are not well suited to answer questions such as how much an individual student has grown or is expected to grow academically (Betebenner, 2008). Instead, VAMs infer contributions of teachers or schools relative to the average and can be used for accountability and educational improvement. Some growth models can be specified to shift the focus from questions about individual student growth to norm-referenced decisions about teacher or school effectiveness. As a simple example, average fitted residuals from linear regression could be used to estimate the contribution of a school to student growth. Multilevel models can also be specified as value-added; for example, the Dallas model estimates teacher effects using a two-stage model with linear regression in stage one and a two-level random-intercept model in stage two (Braun & Wainer, 2007). As such, some growth models used for AYP could be expanded to provide value-added analyses of teachers and schools.

VAMs have been gaining popularity among state policy makers and the Race to the Top Fund program suggests that VAMs may be a part of the revised federal accountability system. It

is imperative that researchers examine the appropriateness of decisions supported by VAMs about schools and teachers that educate students with disabilities, particularly in terms of the impact of missing data, inconsistent accommodation use, and low test scores, as well as the fairness of VAMs for special education teachers whose students have the most significant cognitive disabilities and take the alternate assessment.[4] In addition, research would be needed on the feasibility of VAMs in the presence of coteaching (e.g., Cook & Friend, 1995), which is used primarily for students with disabilities who are in a general education classroom.

Measurement Challenges

As described above, using measures of academic progress over time is of particular interest for students with disabilities. As a subgroup, the participation of students with disabilities in large-scale, standards-based assessments has been increasing, and performance on these assessments has historically been low but improving (CEP, 2009; Thurlow, Quenemoen, Altman, & Cuthbert, 2008). However, students with disabilities are not a homogeneous subpopulation, a characteristic that may have implications for the utility of longitudinal statistical models and the appropriateness of standards-based applications that use grade-to-grade changes in academic performance on annual state assessments. Referring to K–12 accountability assessments, Koretz and Hamilton (2006) made this statement:

> The heterogeneity of students with special needs, the small size of many subgroups, the strikingly inconsistent classification of students with disabilities, the construct-relevance of some of the impediments caused by special needs, and the inability in many cases to clearly delineate construct-relevant from construct-irrelevant impediments – pose formidable barriers to the validation of inferences about the performance of students with special needs. (pp. 563–564)

More specifically, the following characteristics of the subpopulation of students with disabilities may have an impact on applications that use test scores from individual students across two or more grades:

1. use of and changes in testing accommodations and modifications
2. a large portion of students performing significantly below grade level
3. tracking of student growth across testing programs

4. low incidence disability subgroups
5. psychometric properties of alternate and modified assessments

We discuss each of these challenges below.

Accommodation/Modification Use

Testing accommodations and modifications are changes to the standardized administration to improve accessibility for individuals with disabilities. In most cases, accommodations refer to changes that do not alter the construct being measured and modifications refer to changes that do alter the construct being measured. Examples of testing accommodations include large print test forms on tests that do not measure vision or extended testing time on tests that do not measure speed. Testing modifications would include reading aloud a test that is intended to measure decoding of text. As Koretz and Hamilton (2006) stated, "The psychometric function of accommodations is to increase the validity of inferences about students with [disabilities] by offsetting specific disability-related, construct-irrelevant impediments to performance" (p. 562).

Although testing accommodations are not intended to change the construct being measured or the comparability of test scores between accommodated and nonaccommodated testing conditions, research has shown that some testing accommodations result in differential score changes on some assessments for students with disabilities (see Pitoniak & Royer, 2001, or Sireci, Scarpeti, & Li, 2005, for a review). For example, a differential performance boost has been found with both testing accommodations (e.g., read aloud on mathematics tests; extra time on reading and mathematics tests) and testing modifications (read aloud on reading tests that measure word recognition). In addition, research on accommodation use indicates that the number and type of testing accommodations vary from year to year (Fuchs & Fuchs, 2001; Shriner & DeStefano, 2007). Such variation may be due to the changing needs of students over time or the changing preferences of students over time (i.e., refusing to receive testing accommodations). Variation across years may also be due to factors external to students, including changes in state policy (Christensen, Lazarus, Crone, & Thurlow, 2008), limits on resources to administer testing accommodations, or errors in decision-making by the Individualized Education Program (IEP) team. The implication is that the change in test scores from year to year may be related to consistency in the use of accommodations and modifications

rather than true changes in knowledge, skills, and abilities over time. Failing to incorporate information about accommodation or modification use that benefits students with disabilities when using multiple test scores over time may result in less valid inferences or incorrect decisions about the academic progress of such students.

Performing Significantly Below Grade Level

As a subgroup, students with disabilities tend to have lower performance levels than students without disabilities on state standards-based accountability assessments (e.g., CEP, 2009; Klein, Wiley, & Thurlow, 2006; Ysseldyke, Thurlow, Langenfeld, Nelson, & Teelucksingh, 1998). Due to time constraints and limited funding resources, state assessments have typically been built with a difficulty level appropriate for the majority of test takers, which may not be sufficiently aligned with the ability of some students with disabilities. A nonadaptive assessment that is administered to a population with a wide range of proficiency can lead to less precise estimates, lower reliability, and consequently, inappropriate inferences from test scores for individuals at the extremes of the proficiency distribution. While existing large-scale state assessments have been designed to be used to reliably classify students as proficient or not for AYP purposes, they may not have been designed to measure student academic growth. Low performance and low precision of estimates, in combination with the fact that measures of growth are less reliable than measures from a single point in time, may elicit concern about using test scores across grades from assessments that have not explicitly been designed for growth to make high stakes decisions involving students with disabilities.

Tracking Students Across Testing Programs

Under NCLB, all students are required to participate in statewide annual testing and at least 95% of students, measured by school and by subgroup, must participate in a state assessment in order for a school to receive earned credit for AYP. Students with disabilities participate in either the state's general assessment or, for students with the most significant cognitive disabilities, the alternate assessment. Only 1% of the total student population can be counted as proficient for AYP reporting using scores from the alternate assessment. A more recent testing option for AYP reporting is the modified assessment. This option, which is used in at least nine states (Albus, Lazarus, Thurlow, & Cormier, 2009), allows states to create an assessment for students who demonstrate persistent academic difficulties to demonstrate their

proficiency on grade level achievement standards. While any number of students can take the modified assessment, only 2% can be counted as proficient for AYP reporting, which in effect limits the number of students who take the test. Students may move between the general, modified, and alternate assessments in different years for a variety of reasons such as access to the general curriculum and decisions by the students' IEP teams.

States that employ the modified assessment would decrease the number of students with very low scores on the general state assessment but introduce a new problem: tracking of students from year to year and comparability of test forms across the three assessments. Scores from the general, modified, and alternate assessments may not be able to be combined in some longitudinal models, in effect excluding scores from some students with disabilities in some growth-based applications. For example, a recent survey of 15 states with growth models found that 13 states did not include students who took the alternate assessment in their growth model results (Ahearn, 2009). Developing methods for including scores from the alternate assessment in growth models and linking scores on the modified assessment to the general state assessment could help increase the inclusion of test scores for students with disabilities in some growth-based applications.

Low Incidence Disability Subgroups/Changing Profile

While the number of students with disabilities taking annual state standards-based assessments has increased dramatically, some low incidence disability subgroups still exist due to the small number of students with particular types of disabilities in the population and how the population of students with disabilities is distributed within states, particularly in small states. Table 1 shows the breakdown of disability categories in the United States for the most current year available. As can be seen in Table 1, for example, students with visual impairments comprise very small portions of the total population of students with disabilities. In addition students from many of the disability categories are broken down into smaller groups based on grade and the type of assessment the students takes (i.e., general, modified, or alternate).

The number of students with specific disabilities taking statewide accountability assessments generally reflects the proportions in the population. Table 2 is an example of the breakdown by disability subtype from a state with a large general population.

Table 1

Students With Disabilities in the United States Receiving Services Under the Individuals With Disabilities Education Act (IDEA), Fall 2007

	Age			
	3-5	6-11	12-17	18-21
All disabilities	710,371	2,733,616	2,938,905	335,311
		Percent of total		
Specific learning disabilities	1.9%	31.1%	55.1%	45.0%
Speech or language impairments	46.2%	36.2%	5.4%	1.8%
Mental retardation	1.8%	5.7%	9.1%	21.7%
Emotional disturbance	0.5%	4.5%	9.8%	8.9%
Multiple disabilities	1.0%	1.8%	2.2%	5.8%
Hearing impairments	1.1%	1.2%	1.2%	1.4%
Deaf-blindness	< 0.1%	< 0.1%	< 0.1%	0.1%
Orthopedic impairments	1.1%	1.0%	0.9%	1.5%
Other health impairments	2.3%	9.2%	12.0%	8.0%
Visual impairments	0.5%	0.4%	0.4%	0.5%
Autism	5.6%	5.2%	3.4%	4.6%
Traumatic brain injury	0.1%	0.3%	0.5%	0.7%
Developmental delay	38.0%	3.2%	N/A	N/A

Note. Data are from the U.S. Department of Education, Office of Special Education Programs, Data Accountability Center.

Table 2 shows that while the number of students with disabilities taking the general assessment comprises 13–15% of the total student population, the sample size for some disability categories is low in a given year. In addition, the profile of students with disabilities has been shown to change over time (Abt Associates, 2006). This has implications not only for sample sizes across time, but also for the identification of disability classification across time. Furthermore, as Koretz and Hamilton (2006) stated, “The identification and classification of students with disabilities are strikingly inconsistent across states, schools, and classrooms”

(p. 563). Consequently, the match rate for scores across years from students with disabilities may be lower because of these characteristics of the population. While these concerns exist for individual state-specific accountability assessments, such issues may be mitigated by future development of common assessments, whereby scores from the same assessment taken by students from multiple states in a consortium could be combined to increase sample sizes, affording the potential to use a wider variety of longitudinal models.

Table 2

Number of Students by Disability Subgroup Participating in the 2008–2009 Reading Florida Comprehensive Assessment Test

	Grade					
Disability subtype	3	4	5	6	7	8
Specific learning disabled	13,311	14,593	15,386	16,148	16,609	16,535
Speech impaired	7,835	5,619	3,496	1,703	1,069	726
Language impaired	3,933	3,196	2,659	2,190	1,813	1,506
Intellectual disability	813	618	678	686	698	767
Emotional/behavioral disability	1,730	1,900	2,064	2,320	2,394	2,541
Deaf/hard of hearing	291	298	262	265	264	256
Orthopedically impaired	215	178	178	172	180	157
Other health impaired	1,771	1,953	1,968	2,058	1,949	1,884
Visually impaired	86	79	88	97	71	75
Autism spectrum disorder	698	581	548	448	349	346
Dual-sensory impaired	3	-	2	1	1	-
Traumatic brain injured	22	21	32	35	28	33
Hospital/homebound	92	94	89	156	166	229
Total students with disabilities	30,803	29,135	27,450	26,279	25,595	25,055
Percent of total student population	15%	15%	14%	13%	13%	13%

Note. Data are from https://app1.fldoe.org/FCATDemographics/Selections.aspx?level=State&subj=Reading

Movement in or out of special education has implications for AYP reporting since results are reported by subgroup, including the special education subgroup, provided there is sufficient sample size. Systematic changes, including late diagnosis of a disability (such as a learning disability) and early exit disabilities (such as speech and language impairment), may influence growth-based interpretations for the students with disabilities subgroup if the performance of such students differs from students with a consistent classification across years. It may also decrease the sample size such that some schools may not be able to report results by subgroup in some grades. In addition to moving in or out of special education, there may also be obvious clerical errors in reporting that cause changes in special education status to occur in a dataset (e.g., a student classified as blind in year 2 but as having no disability in year 1 or year 3). Finally, changing disability classification across years within the subgroup of students with disabilities can impact research studies that disaggregate by subgroup and may not allow for some disability subtypes to be studied independently. This inconsistency may also be due to errors in reporting, legitimate changes in diagnosed disabilities (e.g., from low vision to blind), or switching primary and secondary disabilities over time (e.g., other health impaired and autism).

Psychometric Quality of Alternate/Modified Assessments

When using some models for change over time, the psychometric properties of the assessment must allow for valid inferences from the use of multiple test scores across time. Challenges to the validity of inferences and the reliability of scores from alternate and modified assessments include limitations such as small sample sizes, fewer items, limited content coverage, and a decreased number of distracters. In addition, no definitive research exists on the validity of interpretations supported by scores from tests taken with accommodations (on the alternate and modified assessments as well as the general assessment). States that develop an alternate or modified assessment are required by the federal government to demonstrate validity, reliability, and accessibility evidence; yet many have not fully done so (CEP, 2009). Without sound psychometric properties, these assessments would not be able to be used for measuring growth in some students with disabilities who take them consistently over time or move between assessments.

Recommendations for Further Research

Given the challenges described above as well as the need to ensure that high stakes decisions based on growth measures for students with disabilities are appropriate, research is needed to address both AYP growth models and VAMs. In this section we discuss three areas for future growth-based psychometric research: testing accommodations, test difficulty, and understanding the population of students with disabilities.

Impact of Testing Accommodations

A critical area in need of research is the impact that testing accommodations and modifications have on inferences and decisions supported by individual test scores from annual statewide standards-based assessments across two or more years. As previously described, changes in accommodation use from year to year may impact a student's growth trajectory; when making growth-based accountability decisions, this could result in testing accommodations being withheld in early grades or forced upon students who no longer need the accommodation in later grades simply to alter the student's growth trajectory. One potential solution is to include specific accommodation use in the model. Another potential solution would be to provide evidence that some accommodations do not obscure growth-based interpretations from certain models. Each of these potential solutions requires that research be conducted to determine the impact that different testing accommodations and changes in accommodation use have on growth trajectories and interpretations from AYP growth models and VAMs. This research could then inform decisions about how to treat and control for the impact of different accommodations when using measures of academic growth.

Impact of Test Difficulty

The increased difficulty of standards-based accountability assessments for students with disabilities raises concern for the accurate determination of AYP using growth models and fair use of VAMs for teacher and school accountability. Research on the impact of low reliability and less precise proficiency estimates on results from such models is important since growth-based applications use multiple test scores that are likely to be measured with higher error relative to the general population. In addition, research on the impact of low initial performance on growth trajectories can potentially help inform state accountability growth targets. Finally, how best to

use academic growth in accountability so that results provide valuable information in conjunction with the current year's scores, yet keep expectations aligned with grade-level standards, is also important to help ensure that students with disabilities are receiving a high quality education and that teachers and schools are being appropriately rewarded for providing it.

Understanding the Population

Preliminary results from some states participating in the Growth Model Pilot have shown lower year-to-year match rates for students with disabilities (e.g., Alaska Department of Education, 2007; Texas Department of Education, 2008). When match rates are low, some schools or districts may not benefit from the use of a growth model, some teachers or schools may not have usable VAM results because of insufficient sample size, and selection bias may be introduced which can impact descriptive and inferential statistics. In order to ensure that appropriate and fair decisions are made from the use of growth models and VAMs in accountability systems, more empirical research is needed to understand characteristics of the population of students with disabilities that impact longitudinal measurements.

The following are some specific questions about the population of students with disabilities that can be answered empirically:

1. What are the characteristics of students excluded from AYP growth models and VAMs (e.g., do early exit disability subtypes change the average proficiency level of the subgroup of students with disabilities in later grades)?
2. How does disability classification change across years in longitudinal databases?
3. To what extent do students with disabilities move between assessments (general, modified, alternate) across years?

Answers to such questions would inform follow-up research on, for example, statistical approaches for dealing with exclusion, substantive and empirical research on the shape of growth trajectories by disability subtype, the accuracy of growth-based decisions by disability subgroup, and best practices for linking scores between assessments (e.g., modified to general). Furthermore, delineating the content being measured on general and alternate assessments would help inform the feasibility of measuring the growth of students with disabilities who move to a different assessment over time.

Using measures of academic growth from annual state standards-based assessments in accountability systems is supported in the educational policy community and among disabilities advocates. Growth is seen as an important tool to help improve the quality of education that students receive; yet measuring growth is complex, particularly for the subgroup of students with disabilities. Assessment developers and policy makers should be mindful of such complexities as ESEA is reauthorized and future state longitudinal data systems are created. While students with disabilities should continue to be included in growth models for accountability purposes, research is needed to address the challenges described herein. Growth-based interpretations can add to the body of evidence about student learning and performance and the contribution of teachers and schools to this effort. Understanding growth measures for students with disabilities and the consequences of their use through empirical research will contribute to making meaningful and appropriate the decisions that rely on such measures.

References

Abt Associates. (2006). *Improving results for students with disabilities: Key findings from the 1997 national assessment studies*. Bethesda, MD: Author.

Ahearn, E. (2009). *Growth models and students with disabilities: Report of state interviews.* Retrieved from http://www.projectforum.org

Alaska Department of Education. (2007). *Match rate – Students Fall 2005 to Fall 2006.* Retrieved from http://www.eed.state.ak.us/tls/assessment/AKGrowthModel/May2007/Studentmatch%20rate%2005%20to%2006.pdf

Albus, D., Lazarus, S. S., Thurlow, M. L., & Cormier, D. (2009). *Characteristics of states' alternate assessments based on modified academic achievement standards in 2008* (Synthesis Report 72). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

American Speech-Language-Hearing Association. (2009). *2009 public policy agenda.* Retrieved from http://www.asha.org/advocacy/briefs-agenda/09PPA.htm

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28,* 42–51.

Betebenner, D. (2008). *A primer on student growth percentiles.* Dover, NH: National Center for the Improvement of Educational Assessment.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective.* New York, NY: Wiley.

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models.* Princeton, NJ: ETS.

Braun, H., & Wainer, H. (2007). Value-added modeling. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 475–501). Amsterdam, Netherlands: Elsevier.

Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting value out of value-added: Report of a workshop.* Washington, DC: National Academies Press.

Briggs, D., & Weeks, J. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28,* 3–14.

Council of Chief State School Officers Accountability Systems and Reporting State Collaborative. (2009). *Guide to United States Department of Education growth model pilot program 2005-2008.* Retrieved from http://www.ccsso.org/content/pdfs/ CCSSO%20ASR%2009%20Guide %20 to%20Growth%20Pilot%20projects.pdf

Center on Education Policy. (November, 2009). *State Test Score Trends Through 2007-08: Has progress been made in raising achievement for students with disabilities?* Washington, DC: Author.

Christensen, L. L., Lazarus, S. S., Crone, M., & Thurlow, M. L. (2008). *2007 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 69). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Cook, L., & Friend, M. (1995). Co-teaching: Guidelines for creating effective practices. *Focus on Exceptional Children, 28*(3), 1–16.

Council on Exceptional Children. (2009). *Race to the top, Comments of Council of Exceptional Children.* Retrieved from http://www.cec.sped.org/AM/Template.cfm?Section=Home& TEMPLATE=/CM/ ContentDisplay.cfm&CONTENTID=12938

Delaware Department of Education. (2009, November). *Accountability technical manual.* Retreived from http://www.doe.k12.de.us/aab/files/DDOE-TM%202008-09-Final.pdf#Accountability Technical Manual 2008-2009

Dunn, J., & Allen, J. (2009). Holding schools accountable for non-proficient students: Coordinating measurement and accountability. *Educational Measurement: Issues and Practice, 28,* 27–41.

Eckes, S., & Swando, J. (2009). Special education subgroups under NCLB: Issues to consider. *Teachers College Record, 111*(11), 2479-2504.

Fuchs, L., & Fuchs, D. (2001). Helping teachers formulate sound test accommodations decisions for students with learning disabilities. *Learning Disabilities Research and Practice, 16,* 174–181.

Growth Model Task Force. (2009). *Growth models for accountability: Considerations and recommendations for including students with disabilities.* New York, NY: National Center for Learning Disabilities.

Hancock, G. R., & Choi, J. (2006). A vernacular for linear latent growth models. *Structural Equation Modeling, 13,* 352–377.

Hill, R., Gong, B., Marion, S., DePascale, C., Dunn, J., & Simpson, M. A. (2006). Using value tables to explicitly value growth. In R. Lissitz (Ed.), *Longitudinal and value added models of student performance* (pp. 255–290). Maple Grove, MN: JAM Press.

Jöreskog, K. G. (1977). Structural equation models in the social sciences: Specification, estimation, and testing. In P. R. Krishnaiah (Ed.), *Applications of statistics.* Amsterdam, Netherlands: North-Holland.

Klein, J. A., Wiley, H. I., & Thurlow, M. L. (2006). *Uneven transparency: NCLB tests take precedence in public assessment reporting for students with disabilities* (Technical Report 43). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York, NY: Springer.

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

Lissitz, R. W., & Huynh H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation.* Retrieved from http://PAREonline.net/getvn.asp?v=8&n=10

Mintrop, H., & Sunderman, G. L. (2009). Predictable failure of federal sanctions-driven accountability for school improvement and why we may retain it anyway. *Educational Research, 38,* 353–364.

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of Kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities, 42,* 306–321.

National Center for Educational Achievement. (2008). *2008 survey of state longitudinal data systems.* Retrieved from http://www.dataqualitycampaign.org/

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 et seq. (2002).

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71*(1), 53–104.

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

Shriner, J., & DeStefano, L. (2007). Assessment accommodation considerations for middle school students with disabilities. In C. C. Laitusis & C. Cook (Eds.), *Large-scale assessment and accommodations: What works?* Washington, DC: Council for Exceptional Children.

Sireci, S. G., Scarpeti, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75,* 457–490.

Slater, S. C., Wentzel, C., & Chard, L. (2009, April). *Applications of growth models within the framework of No Child Left Behind.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Texas Department of Education. (2008). *Response to United States Department of Education clarification questions on Texas' growth model pilot proposal.* Retrieved from http://ritter.tea.state.tx.us/student.assessment/resources/growth_proposal/111208_TXResponseto_USDE_ClarificationQuestions.pdf

Thurlow, M. L., Quenemoen, R., Altman, J. R., & Cuthbert, M. (2008). *Trends in the participation and performance of students with disabilities* (Technical Report 50). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Tong, Y., & O'Malley, K. (2006). *An empirical investigation of growth models.* Austin, TX: Pearson Educational Measurement.

U.S. Department of Education. (2009). *Race to the Top Fund: Notice of proposed priorities, requirements, definitions, and selection criteria.* Retrieved from http://www.ed.gov/legislation/FedRegister/proprule/2009-3/072909d.html

Weiss, M. J. (2008). *Using a yardstick to measure a meter: Growth, projection, and value-added models in the context of school accountability* (Doctoral dissertation). Retrieved from ProQuest.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273–282). New York, NY: Springer.

Yen, W. (2008). *Measuring academic growth in California.* Princeton, NJ: ETS.

Ysseldyke, J. E., Thurlow, M. L., Langenfeld, K. L., Nelson, J. R., Teelucksingh, E., & Seyfarth, A. (1998). *Educational results for students with disabilities: What do the data tell us?* Minneapolis, MN: National Center on Educational Outcomes.

Notes

[1] Value-added describes a family of models that estimate the contributions of schools or teachers to student academic growth using student test scores that are adjusted for prior academic achievement and/or other student characteristics. See Braun (2005) for an introduction and Braun, Chudowsky, and Koenig (2010) for measurement and policy issues.

[2] *Modified assessment* herein refers to an alternate assessment based on modified achievement standards, and *alternate assessment* herein refers to an alternate assessment based on alternate achievement standards.

[3] Hierarchical linear modeling and multilevel modeling can be used interchangeably. The term *multilevel modeling* is used herein.

[4] It is difficult to show validity evidence using traditional psychometric approaches because of the typical characteristics of an alternate assessment such as few items, small test taker volumes, and limited content coverage.

Appendix A

Delaware's Growth Model

Table A1

Value Table

Year 1 level (Grades 3–9)	Year 2 level (Grades 4–10)				
	Level 1A	Level 1B	Level 2A	Level 2B	Levels 3–5 proficient
Level 1A	0	150	225	250	300
Level 1B	0	0	175	225	300
Level 2A	0	0	0	200	300
Level 2B	0	0	0	0	300
Levels 3-5 Proficient	0	0	0	0	300

Year 1 Grade 2 level	Grade 2 to Grade 3				
	Level 1A	Level 1B	Level 2A	Level 2B	Levels 3-5 proficient
Level 1A	0	0	0	200	300
Level 1B	0	0	0	0	300

Note. Data are from the *Accountability technical manual* by the Delaware Department of Education, November 2009. Retreived from http://www.doe.k12.de.us/aab/files/DDOE-TM%202008-09-Final.pdf#Accountability Technical Manual 2008-2009.

The values assigned to individual student transitions from the value table above are averaged across students in a school for mathematics and reading separately. This average is computed across all students and across student subgroups and compared to a growth target (see Table A2). If the average meets or exceeds the growth target, the school gets credit for AYP.

Table A2

Growth Targets

	Reading	Math
2006	186	123
2007	204	150
2008	204	150
2009	219	174
2010	237	201
2011	252	225
2012	267	249
2013	285	276
2014	300	300

Note. Data are from the *Accountability technical manual* by the Delaware Department of Education, November 2009. Retreived from http://www.doe.k12.de.us/aab/files/DDOE-TM%202008-09-Final.pdf#Accountability Technical Manual 2008-2009.

Appendix B

Alaska's Growth Model

Figure B1 is a graphical representation of Alaska's growth model for two hypothetical students who took the annual state assessment in the same local education agency (LEA) in grades 3 and 4. Growth targets for individual students are calculated based on the student's initial score in grade 3 or the first year the student entered the LEA using the formula, $y_g = y_{g-1} + \frac{1}{\mathrm{T}}(\mathrm{PROF} - y_0)$, where g is the current grade, y_0 is the student's initial score, PROF is the proficiency cutscore, and T is the total number of years to reach proficiency. The number of students who are *on track to proficient* is counted along with the number of students who are proficient for both reading and mathematics separately.

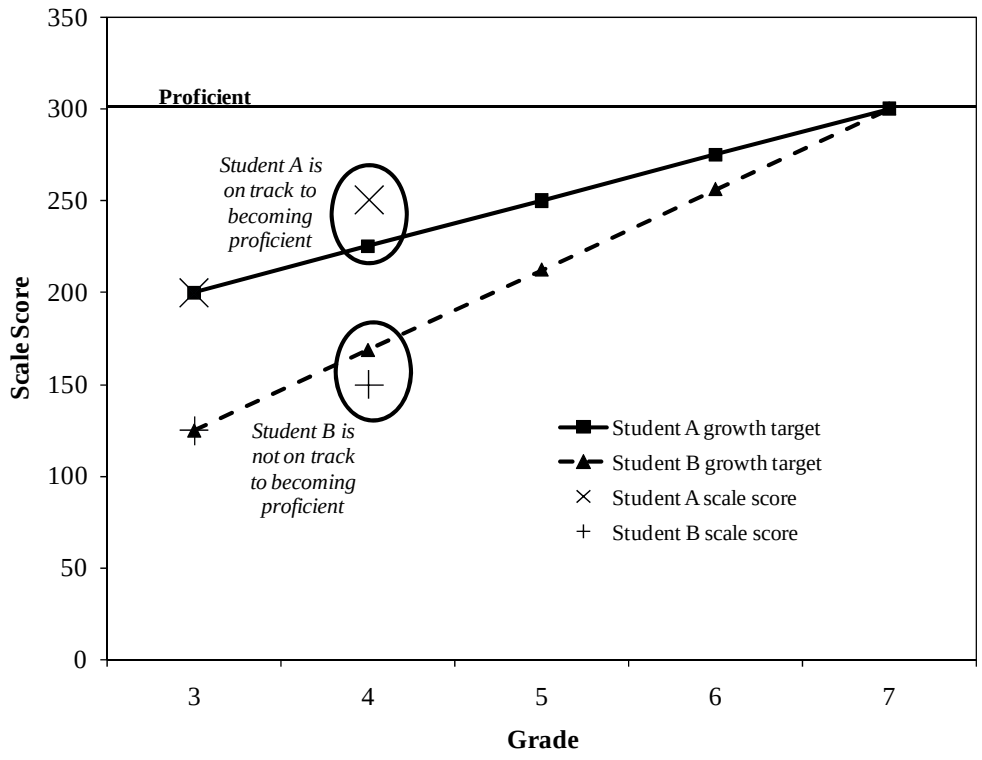

Figure B1. **Alaska's growth model for two hypothetical students who took the annual state assessment.**