

A Practitioner's Guide to
Growth Models



Katherine E. Castellano
University of California, Berkeley

Andrew D. Ho
Harvard Graduate School of Education

February 2013

A Practitioner's Guide to Growth Models

Authored By:

Katherine E. Castellano, University of California, Berkeley

Andrew D. Ho, Harvard Graduate School of Education

*A paper commissioned by the
Technical Issues in Large-Scale Assessment (TILSA)
and
Accountability Systems & Reporting (ASR)
State Collaboratives on Assessment and Student Standards
Council of Chief State School Officers*



Copyright © 2013 by the Council of Chief State School Officers.

All rights reserved.

ACKNOWLEDGEMENTS

This report has benefitted from insightful comments and reviews from State Collaboratives on Assessment and Student Standards (SCASS) members, making it truly a product of collaboration. We extend special thanks to several assessment experts who volunteered their time and energy to improving various drafts. Their insights as practitioners enhanced the utility of this report for its target audience. We thank Bill Bonk (Colorado Department of Education), Beth Cipoletti (West Virginia Department of Education), Juan D’Brot (West Virginia Department of Education), Gary Phillips (American Institutes for Research), and Michelle Rosado (Connecticut State Department of Education) for their constructive reviews. Bill Auty (Education Measurement Consulting) provided assistance through the drafting process, and Frank Brockmann (Center Point Assessment Solutions, Inc.) provided design and production assistance.

We would also like to acknowledge the support of SCASS advisors Charlene Tucker, Duncan MacQuarrie, and Doug Rindone in providing feedback throughout the development of this report. Their vision for clear and accurate descriptions of growth models improved the content and the style of the document. Any remaining errors are ours.

CHAPTER 1

The Gain Score Model

The gain score model is a simple, accessible, and intuitive approach that primarily supports **growth description**. As its name suggests, it is a **gain-based model**, and it serves as a basis for more complex models like the trajectory and categorical models as well as some “value-added” models. The gain score model, also referred to as “growth relative to self” or “raw/simple gain,” addresses the question

How much has a student learned on an absolute scale?

The answer to this is the gain score: the simple difference between a student’s test scores from two time points. For this difference to be meaningful, student test scores from the two time points must be on a common scale. If the two time points represent two grade levels, then the common scale should be linked to a developmental continuum representing increased mastery of a single domain.

Question 1.1:

What *Primary Interpretation* Does the Gain Score Model Best Support?

Of the three primary growth model interpretations — growth description, growth prediction, and value-added — the gain score model supports growth description.

The gain score model describes the absolute change in student performance between two time points. This is sometimes called “growth relative to self” (DePascale, 2006) as the student is only compared to himself or herself over time.

GAIN SCORE MODEL

Aliases and Variants:

- Growth Relative to Self
- Raw Gain
- Simple Gain
- Gains/Slopes-as-Outcomes, Trajectory Model

Primary Interpretation:

Growth Description

Statistical Foundation:

Gain-based model

Metric/Scale:

Gain score – on the common test score scale

Data: Vertically-scaled tests and test scores from two time points

Group-Level Statistic:

Average Gain – describes average change in performance from Time 1 to Time 2

Set Growth Standards:

Determining a minimum gain score needed for “adequate growth”

Operational Examples:

- Pretest/Posttest experimental designs
- Quick growth summaries
- A basis for trajectory models

The sign and magnitude of a gain score are important in indicating a student's change in performance. The magnitude of the gain indicates how much the student has changed, whereas the sign indicates if the gain was positive, signifying improvement, or negative, signifying decline. Gain scores require an understanding of the underlying test score scale in order to be interpreted meaningfully. A 350, a 375, and a difference of 25 carry little meaning unless the scores and the gain refer to well-understood locations on an academic or developmental scale. When the scale is not well known or understood, the gain score can be referenced to a norm or standard, as described in Section 1.5.

Gain scores can be generalized to more than two time points through the calculation of an average gain or a slope. An average gain is equivalent to the difference between the initial and current scores divided by the grade span. A slope is found through a regression model that estimates the best-fit line through the trajectory. This use of regression to describe scores relative to *time* contrasts with the use of regression in conditional status models, which use regression to describe current scores relative to *past scores*.

Question 1.2:

What is the *Statistical Foundation Underlying the Gain Score Model*?

The statistical foundation of the gain score model is, as the name suggests, a gain-based model.

The gain score model produces gain scores, which are sometimes referred to as "raw gains," "simple gains," or just "gains." A gain score is found using test scores from two time points as follows:

$$\begin{aligned}\text{Gain Score} &= \text{Test Score at Current Time Point} - \text{Test Score at Previous Time Point} \\ &= \text{Current Status} - \text{Initial Status}\end{aligned}$$

Figure 1.1
Illustration of the Gain Score Model

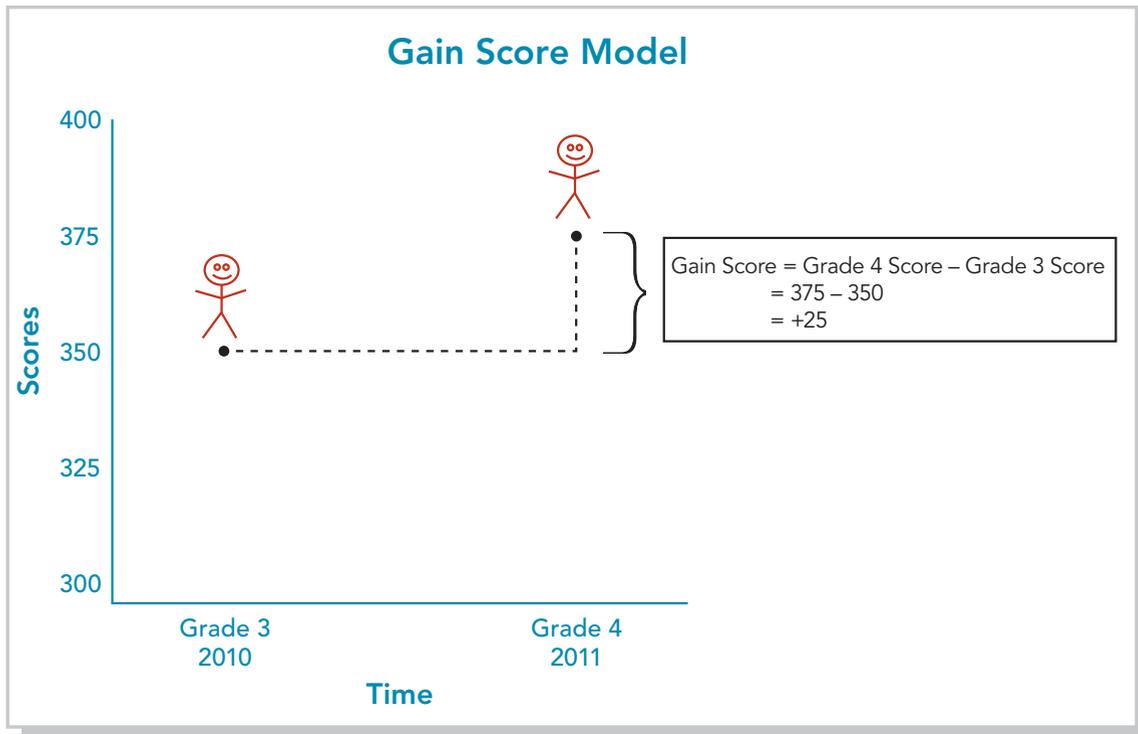


Figure 1.1 illustrates the gain score model calculation using data for a student in Grade 3 in 2010 and in Grade 4 in 2011 on a hypothetical mathematics test. The horizontal axis represents time, and the vertical axis represents the test score scale. For test scores from both the Grade 3 and Grade 4 assessments to be shown on this continuous scale, these two assessments must share an underlying vertical scale.

The solid, black dots in Figure 1.1 mark a particular student's test scores. This student, represented with stick figures, earned a score of 350 in Grade 3 and 375 in Grade 4. The gain score is illustrated by the vertical difference in these two scores, which, as shown in the figure, is $375 - 350 = +25$. The reporting scale for the gain score is the common scale of the two test scores. Combining the positive sign and the magnitude of the gain score, this student gained 25 points from 3rd grade to 4th grade on this hypothetical state mathematics assessment.

Question 1.3:

What are the *Required Data Features* for the Gain Score Model?

The gain score model requires student test score data from at least two time points from tests aligned to a common scale. The student test score data must be linked over time, requiring unique student identifiers.

Gain scores require scores for students from at least two time points. The database requires unique student identifiers that are constant over time, and group-level identifiers are necessary to support group-level analyses. Even given these data, interpretations of gain scores are only appropriate if the test scale is designed to support meaningful differences in test scores. If the scores from the two time points are on different scales, then such a difference is not interpretable. Accordingly, the scores from each time point must be on a common scale. This context is sometimes described as a pretest/posttest design, where the pretest and posttest are either the same test, making their scales equivalent, or are carefully developed tests that share content and technical specifications that allow them to be equated and placed on a common scale. In contrast, when the scores are from different grade-levels as in Figure 1.1, their shared scale is typically called a vertical scale.

Vertical scaling is a difficult enterprise, and casually or poorly constructed scales are a serious threat to the use and interpretation of gain scores and models based on them. To construct a defensible vertical scale, test designers must invest considerable work during the test development process to set content specifications that span a developmental continuum. Other requirements include items that meet these specifications, administration of tests to an appropriate sample of students during the scaling process, attention to statistical models for creating the vertical scale, and evaluation of the results of the scaling (Kolen & Brennan, 2004). Poorly designed vertical scales can result in serious distortions, including ceiling effects that artificially restrict the gains of initially high scoring students or spurious relationships between gains and initial status. This may lead to the illusion that high scoring students have greater gains than low scoring students, or vice versa, when this may not actually be the case. A well-designed vertical scale will minimize ceiling effects, support defensible interpretations about the relationship between gains and status, and be anchored to a substantive domain through which growth can be well understood.

Gain scores are sometimes accused of having low precision and reliability. However, reliability, like validity, is best expressed in terms of a desired purpose. If the primary interest is in ranking individuals by gain scores, then gain scores are often problematic and are best derived from tests that themselves have high reliabilities or data from multiple time points. If the magnitude of the gain is the target of inference, rather than relative rankings, gain scores are both appropriate and can have sufficient precision (Rogosa, 1995). Finally, if group-level, or average gain scores are the target of inference, then gain scores can support precise inferences provided that the underlying vertical scale is defensible.

Question 1.4:

What Kinds of *Group-Level Interpretations* can the Gain Score Model Support?

Gain scores can be aggregated to the group-level by taking the average of a set of students' gain scores. Average gain scores describe the average change in performance for the group. Similar to student-level gain scores, average gain scores are best suited for growth description at the group level.

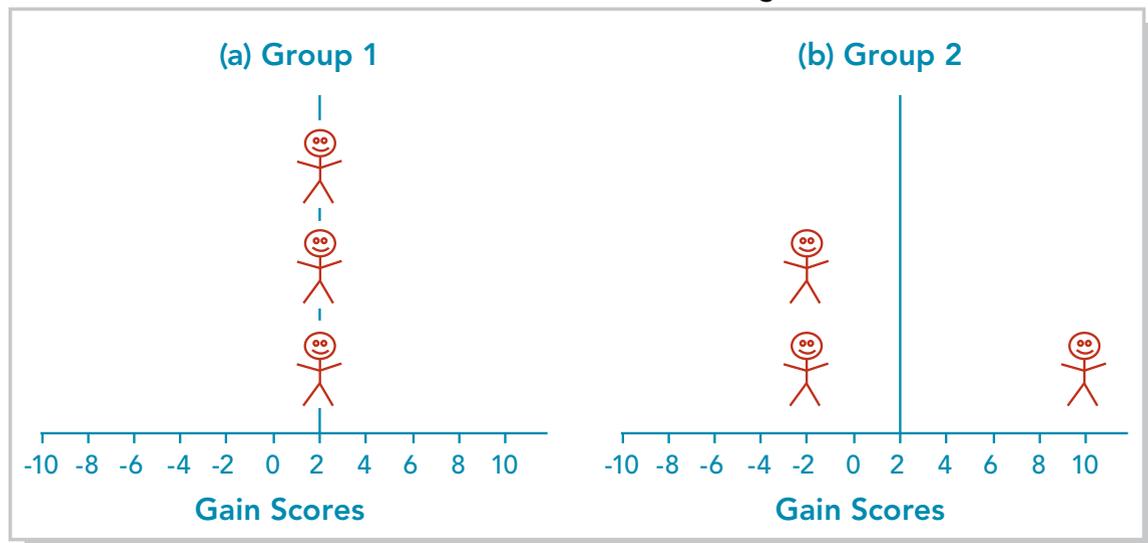
The gain score model supports simple calculations of group-level statistics. Most commonly, the group-level summary statistic for a set of students of interest, such as in a particular classroom, grade level, school, or district, is the average of their individual gain scores. This summary statistic is typically referred to simply as an “average gain score.”

Average gain scores provide descriptions of group-level growth. They describe how much the students in that group have improved on average. A near zero average gain score indicates that either all students had near zero gains or that there was rough balance between positive gains and negative gains that average to near zero. A positive average gain score indicates that students, on average, made positive gains, whereas a negative average gain score indicates that students generally declined in performance.

Simple summary statistics are often insufficient to support full inferences about the distribution of student growth. Graphical displays of student gain scores often provide a clearer picture of the overall growth of a group.

Figure 1.2 illustrates a simplistic case in which two groups of students have the same average gain score but the distributions of gain scores are quite distinct. Both groups of three students have an average gain score of +2, as shown by the thick, vertical line at +2. In Group 1, shown in panel (a), all three students have the same gain score of +2. In contrast, in Group 2, two students have slightly negative gains of -2 and one student has a large positive gain of 10. Although both groups have an average gain score of +2, this single summary statistic provides a limited depiction of the distribution of growth of these groups. These coarse averages are best disaggregated when the primary purpose of reporting is the support of teaching and learning.

Figure 1.2
Different Distributions of Gain Scores with the Same Average Gain Score



An extension of the gain-score model involves using gains as outcome variables in regression models. These models predict growth through individual, classroom, and school variables, and they identify relationships between these variables and magnitudes of growth over time. These types of models can be used to support value-added interpretations. For example, schools or classrooms associated with higher levels of average growth may be investigated to understand the mechanisms through which this growth may have occurred. However, although no model can support value-added inferences on its own, gain-based models are particularly poorly suited to value-added inferences given their dependence on vertical scaling properties.

Vertical scales are typically developed to support growth description and not causal inference about growth. For example, in certain curricular domains, vertical scales often reflect increased variability in student achievement as grade levels increase. This is consistent with a positive correlation between initial status and growth, where higher scoring students in any particular grade are predicted to make greater gains into the future. This is a useful observation for the design of instruction, but an undesirable feature for value-added models where giving credit to higher growth for higher-scoring students seems unfair. This is a reminder of the fundamental importance of specifying the intended interpretations and use of growth models.

Question 1.5:

How Does the Gain Score Model Set Standards for Expected or Adequate Growth?

Value judgments can determine cut points for “low,” “typical,” and “high” gain scores at the individual and group level. Growth expectations can also be norm-referenced by comparing students’ gain scores to the growth distribution of a reference group. A standard can also be set by anticipating whether a student or group is on track to some criterion in the future.

The simple gain score is an index of absolute growth, expressing how much a student grew on an absolute scale. Students, teachers, parents, and school administrators may want to know not only “how much” a student has grown, but also if that growth is “adequate” or “good enough.” As with most growth models, a standard setting committee composed of qualified, informed, and invested stakeholders can be charged with defining adequate growth. The magnitude of the gain score may not be sufficient to communicate the adequacy of growth. Intuitively, it may seem clear that negative gains are inadequate, but to ensure that all data users interpret the gain scores in a uniform manner, clearer reporting categories may be required. These categories can be determined in three different ways: 1) scale-based standard setting, 2) norm-referenced standard setting, and 3) target-based standard setting.

Scale-based standard setting involves setting cut points on the gain-score scale to differentiate among gains, for example, “negative,” “low,” “adequate,” and “high” growth. For determining appropriate cuts on the gain score scale, a standard setting committee may consider the empirical distribution of gain scores to avoid setting unrealistic standards. Although the committee could decide to use the same set of cut scores across grades, the pattern of changes across grades would be unlikely to support common standards, as different gains are likely to vary across grade level. Similar procedures could be completed at the group level for classifying average gain scores as low, typical, or high group growth.

Norm-referenced standard setting uses a distribution of gain scores from a “reference group” to set expectations about adequate growth. This reference group can be a static “norm group” sampled from some representative population. Alternatively, the reference group can be updated, defined each year based on current, operational student performance. A natural reporting metric is the percentile rank of each gain score in the reference group, where a student whose gain is above 75 percent of the reference group’s gains receives a growth percentile of 75.⁴ In this case, the effective reporting scale is the norm-referenced percentile rank scale, and a standard setting committee can identify where cut scores are located on this scale. As with scale-based standard setting, these norm-referenced standard setting procedures can be applied at the group level to set expectations for adequate group gains relative to the distribution of all groups’ average gain scores.

Target-based standard setting classifies students/groups as making adequate growth by determining if they are “on track” to some target standard at a future point in time. For instance, a target may be defined as reaching the proficiency cut point in a particular grade level or exceeding the “College and Career Ready” standard by a particular grade. This intersects with the primary interpretation of growth prediction, and the trajectory model (described in the next chapter) uses the gain score in precisely this way. This extension to the gain score model assumes that students continue on their growth trajectories over time, making the same gains each year.

Question 1.6:

What are the Common Misinterpretations of the Gain Score Model and Possible Unintended Consequences of its Use in Accountability Systems?

The gain-score model aligns well with common intuition about growth over time. Biases and distortions can be introduced through poor vertical scaling. Gains can be inflated by artificially deflating prior scores.

⁴ This contrasts with the Student Growth Percentile (Chapter 6), where the reference group is defined empirically by a subset of students with similar past scores. In this case, the reference group is a full distribution of current or past gains.

The gain score model aligns closely with intuitive notions of growth. However, there are a number of shortcomings of gain-based descriptions that do not follow from common intuition about gains. First, simple gain-based approaches use only two time points and can be unreliable with respect to individual comparisons of gains. For more robust information about an individual's growth trajectory, more than two time points may be required. This is generally addressed by using multiple time points and fitting a simple regression-based estimate of an individual slope over time, resulting in an average gain score for an individual. More advanced estimates of individual growth curves can be supported with multiple time points, nonlinear trajectories, and latent growth curve analyses. These are natural extensions of the simple gain-score model.

Second, properties of the vertical scale may lead to correlations between initial status and growth that are poorly suited for accountability metrics. For example, some vertical scales reflect the observation that variability in individual achievement increases over time. In these cases, high scoring students are more likely to make greater gains than lower scoring students. Although this may be a valid interpretation on a particular developmental score scale, it may be poorly suited for accountability metrics, where expectations for higher and lower scoring students may be required to be equal. On the other hand, these differential, scale-based expectations for lower scoring students may be precisely what the accountability model should reflect. If the vertical scale is well developed, it may reflect the reality that it is more difficult for lower scoring students to catch up without adequate intervention. The interactions between scaling decisions and growth expectations must be evaluated with respect to the inferences and actions that the growth interpretations support.

Third, a vertical scale that is poorly designed will have biases built into the scale. In these cases, associations between initial status and growth may be spurious, and expectations based on growth will be similarly unrealistic for higher and lower scoring students. Hidden ceiling and floor effects will lead to an inability of high or low scoring students to demonstrate their true growth. In general, the considerable reliance of the gain-score model on responsible vertical scaling leads to greater dependence of results on scaling properties. When there are weaknesses, they are likely to arise accidentally, but they are difficult to detect without thoughtful exploratory data analysis.

Finally, another feature of gain scores can be manipulated more cynically when gain scores form the basis of high-stakes accountability decisions. It is apparent from the calculation of the gain score that a student can have a higher gain by increasing his or her current score. This is a desired response to accountability pressures. However, it is also possible to reverse this — a student can have a higher gain by decreasing his or her previous score. This could be achieved by distorting reporting, but also more systematically by pushing less experienced

teachers to early tested grades. Although this may appear cynical, this guidebook would be incomplete without a comprehensive presentation of both intended and unintended consequences of each model as it may function in practice.

References

- DePascale, C.A. (2006). *Measuring growth with the MCAS tests: A consideration of vertical scales and standards*. Dover, NH: National Center for Improvement in Educational Assessment, from http://www.nciea.org/publications/MeasuringGrowthMCASTests_CD06.pdf.
- Kolen, M.J., and Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer Science+Business Media, Inc.
- Rogosa, D.R. (1995). Myth and methods: 'Myths about longitudinal research' plus supplemental questions. In J.M. Gottmann (Ed.), *The analysis of change* (pp. 3-66). Mahwah, NJ: Lawrence Erlbaum.

CHAPTER 3

The Categorical Model

Categorical models characterize growth in terms of changes in performance level categories from one grade to the next. They are also referred to as transition models, transition matrix models, or value tables. These names are often used interchangeably, although the term “value table” typically refers specifically to categorical models that assign differential values or weights to transitions.

The categorical model is a **gain-based** model that is fundamentally similar to the gain score model. Instead of expressing gains as the change in scale score points from one year to the next, the categorical model expresses gains as the *change in performance level categories* from one year to the next. This results in a large reduction in information about student scores, as the entire range of score points is substantially reduced to a small number of reporting categories. Positive gains are associated with moving up one or more performance levels, whereas negative gains are associated with moving down one or more performance levels. In this sense, categorical models support **growth descriptions** like the gain score model. Although, compared to using the scale score, performance level categories are coarser and information is lost, the categorical model is easy to describe and explain, particularly if the category definitions are relevant and well understood.

Categorical models also implicitly support **growth predictions**. Transitions through past categories can support predictions about student location in categories in the future. Categorical models can address both of the following questions:

CATEGORICAL MODEL

Aliases and Variants:

- Transition Model
- Transition Matrix Model
- Value Table

Primary Interpretation:

Growth description and growth prediction

Statistical Foundation:

Gain-based model

Metric/Scale: Change in performance level categories (categorical scale)

Data: Performance levels articulated across years (implicit vertical scale), student status expressed by performance level, and values for transitions if value tables are used

Group-Level Statistic:

Percentage of students “on track” to proficiency or average value across value tables

Set Growth Standards:

Define cut scores for performance levels and values for value tables; specify rules for students being counted as “on track”; establish what average value is good enough

Operational Examples:

NCLB Growth Model (e.g., Delaware and Iowa)

**How has this student grown in terms of transitions through performance level categories over time?
In which category will she likely be in the future?**

An advantage of categorical models is their conceptual simplicity. However, they can rely on a large number of explicit and implicit judgments. Some accountability systems prefer to value certain transitions between performance levels more than others, resulting in a categorical model that is often called a “value table.” There is also a series of less obvious judgments involved in setting the cut scores that delineate each category. These decisions require consideration of several issues, including the transitions that receive weight, the differential weighting of transitions, and cut score articulation across grades.

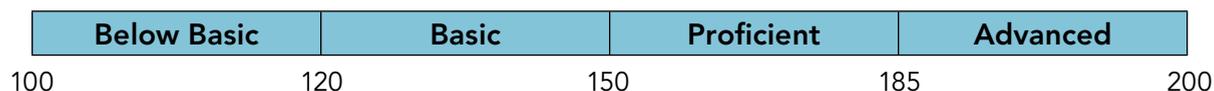
Question 3.1:

What *Primary Interpretation* Does the Categorical Model Best Support?

Categorical models can support both growth description and growth prediction. They describe how much students grow from one year to the next in terms of changes in performance level categories. Categorical models can also implicitly or explicitly predict the category a student will achieve in the future, under an assumption of linear progress across categories.

Categorical models support growth descriptions and growth predictions. Like both the gain score and trajectory model, the categorical model is based on a conceptualization of growth as an increase in score points from one year to the next. The fundamental distinction between the categorical model and the other gain-based models is that the categorical model uses score points that are expressed as a small number of performance level categories as opposed to using the tests’ entire score point scale. Performance level categories are often ascribed names like “Below Basic,” “Basic,” “Proficient,” and “Advanced” that denote varying degrees of mastery. The numerical test score scale is divided into these ordered categories by cut scores on the test scale. Figure 3.1 illustrates this for a hypothetical test scale that ranges from 100 to 200 points.

Figure 3.1
Illustration of a Test Scale Divided into Ordered Performance Level Categories by Cut Scores



As shown in Figure 3.1, ordered performance level categories are just a “chunking” of the numerical test scale. A student who earns a score of 125 is in the “Basic” performance level, as her score falls between 120 and 150. The scores of 120, 150, and 185 are cut scores that divide the four performance level categories. In the usual standards-based testing scenario, a standard setting committee would determine the cut scores with careful consideration of the test scale, item content and difficulty levels, student performance on the items in the tests, and the qualitative descriptions of each category. In this example, they are chosen for illustration. Before cut scores can be determined, the categories must be carefully defined so that they relate to distinct skill sets and mastery levels. Simply dividing the scale into a set of categories is not useful unless each category provides useful information about a student’s achievement level.

To implement a categorical growth model, performance levels are ideally articulated across grade levels, meaning that they are defined with qualitative descriptions and cut scores that reflect not only within grade mastery but a continuum of mastery across several grade levels. The same set of category names are usually used in each grade, but the qualitative descriptions of the categories differ across grades as they reflect different skill sets and ability levels. Accordingly, the cut scores that distinguish among the categories may vary in relative stringency across grades. This is discussed further in Section 3.5.

After articulating cut scores across all the grade levels of interest, the decisions supported by the categorical model can be illustrated by a “transition matrix.” Table 3.1 gives an example of a transition matrix for the change in performance level category from Grade 3 to Grade 4 for a state mathematics test. In this illustrative example, each grade-level test scale is divided into four categories — Below Basic, Basic, Proficient, and Advanced — like in Figure 3.1. The cells along the diagonal are shaded grey. These shaded cells correspond to cases in which a student maintains the same performance level category in Grade 3 and Grade 4. The cells below the diagonal correspond to cases in which a student goes down one or more performance levels from Grade 3 to Grade 4. The remaining cases, the cells above the diagonal, represent growth or moving up one or more performance levels from Grade 3 to Grade 4. A student, represented by a stick figure, falls in one of these cells — in the first row and second column. This student scored at the Below Basic level in Grade 3 but in the Basic level in Grade 4. This change in performance level from Grade 3 to Grade 4 signifies that the student improved, grew, or increased in terms of achievement level categories.

Table 3.1
Example of a Transition Matrix

Performance Level in Grade 4				
Performance Level in Grade 3	Below Basic	Basic	Proficient	Advanced
Below Basic				
Basic				
Proficient				
Advanced				

Table 3.1 illustrates the use of categorical models for growth description. This simple table shows the student of interest increased one performance level category. Within the Grade 3 domain of mathematics, the student only had a Below Basic understanding and mastery of the material. However, in Grade 4, she has improved to a Basic understanding of Grade 4 mathematics. Ostensibly, in terms of achievement level categories, this student has grown.

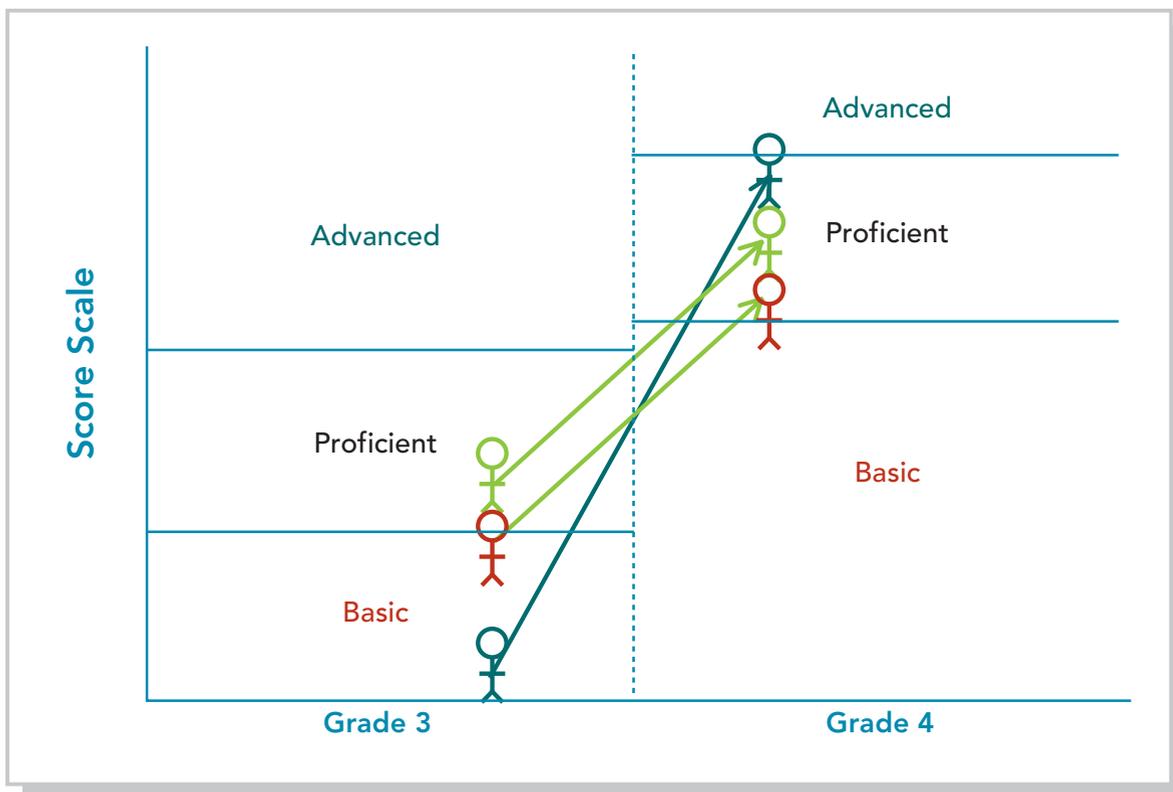
Interpreting a change in achievement level categories as growth can lead to some counterintuitive findings. To clarify these findings, it can be useful to imagine a vertical scale that underlies the achievement level categories across grades. This is shown in Figure 3.2. One counterintuitive finding is that the maintenance of an achievement level over time represents a kind of stasis. This may conflict with commonsense notions of growth, as maintenance of a standard across grades generally requires growth, as shown by the green student in Figure 3.2. This conflict is generally resolved by observing that interpretations of achievement level categories across grades are more relative than they are absolute.

A second counterintuitive finding is that similar levels of growth over time may or may not lead to a change in categories. As Figure 3.2 shows, two students (represented by the green and red stick figures) who make the same absolute scale score gains can either maintain the proficiency category or rise from Basic to Proficient depending on their starting point and their position with

respect to the cut scores. This is explained by the loss of information that arises from dividing the score scale into a small number of categories. As a corollary, a change in categories can be associated with a very wide range in actual gains, simply due to where the student happens to be within the coarse category regions. For example, the blue student scores at the very bottom of the scale in Grade 3 and then at the upper boundary of the Proficient category in Grade 4. The red student scores at the top of the Basic category in Grade 3 and the bottom of the Proficient category in Grade 4. The categorical model treats these two students' gains as equivalent.

Figure 3.2

Illustration of Possible Contradictions when Mapping a Vertical-Scale-Based Definition of Growth onto a Categorical Definition of Growth



As the previous discussion demonstrates, the categorical model affords growth interpretations through the articulation of achievement level categories across grades. Although this does not require an explicit vertical scale, the resulting interpretations of results assume that a vertical scale exists. Through the articulation of cut scores across grades, the categorical model creates an implicit vertical scale. Even if a performance level happens to describe different domains across grades, the implicit assumption is that an increase in achievement levels is desirable and interpretable as growth.

If the categorical model supports growth interpretations, it is essential that the performance level categories are carefully defined and are vertically aligned over an underlying achievement

continuum. If scores at the top of the Basic category reflect markedly different achievement than scores at the bottom of the category, then the category should be further subdivided into finer categories, or alternatives like trajectory models should be considered.

To support growth predictions, categorical models can include the assumption that transitions across categories will continue in a linear fashion over time. This is a coarser, categorical version of the trajectory model that assumes that students continue to make the same gains each year as they have in recent years. If a student improves one performance level category from last year to this year, it might seem reasonable to then assume she will improve one more performance level category next year. In our illustrative example, our student of interest went from Below Basic to Basic from Grade 3 to Grade 4. Thus, if the student continues to make such growth, we would predict that she would move up yet another performance level next year and be Proficient. Rules can be set to label students as “on track” to reaching a desired performance level, such as Proficient or College and Career Ready. Section 3.5 discusses these rules further.

Question 3.2:

What is the *Statistical Foundation Underlying the Categorical Model?*

The categorical model is a re-expression of the gain score model using performance level categories instead of scale scores. It is implicitly a gain-based model of growth.

The categorical model and the gain score model (Chapter 1) are similar in concept, although they express growth on different scales. The gain score model requires that each grade level test be linked to a common vertical scale, allowing for scores across grades to be comparable. It then defines gain scores as the difference in scale score points from one year to the next. In contrast, the categorical model requires that each grade level test scale be divided into distinct achievement level categories that have accompanying qualitative descriptions of the skills and mastery level students at that level should have. It then defines gain scores as the difference in performance level categories from one grade to the next.

Gains in the categorical model can be expressed qualitatively, for example, “She was Below Basic in Grade 3 and Basic in Grade 4.” The gains can also be expressed numerically, as in “a gain of one achievement level.” The range of possible gains is substantially reduced from the gain score model to the categorical model. The gain score model uses the entire range of possible score scale points, whereas as the categorical model collapses the score scale into a far smaller number of categories.

Categorical models allow for flexibility in the assignment of numbers or values to each category or to each transition. In the previous example, the transition could be weighted by

the number of categories that each student changed. This numerical assignment would result in any increase of one performance level to correspond to a gain of +1, any decrease in two performance levels corresponds to a gain of -2, and so on. In contrast, all positive transitions might be valued as +1 regardless of how many categories a student jumped. In other cases, certain transitions might be valued higher than others.

A categorical model that uses careful assignment of different values to each transition is often referred to specifically as a “value table.” Table 3.2 provides an example of a value table. In response to the allowance of growth models under the Growth Model Pilot Program, Delaware, like several other states, adopted a categorical model for determining accountability calculations under NCLB. In this example, there are four performance level categories below Proficient. Any non-proficient student that gains in terms of achievement level categories receives a particular number of points. Students that reach the desired performance level category of Proficient receive the highest weight of 300 points. For the remaining positive transitions, larger jumps and jumps starting from performance level categories closer to Proficient are weighted highly. For instance, a student transitioning one category from Level 1A to Level 1B counts for 150 points, whereas a student transitioning one category from Level 1B to Level 2A counts for 175 points.

Table 3.2
Example of a Value Table

	Year 2 Level				
Year 1 Level	Level 1A	Level 1B	Level 2A	Level 2B	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
Proficient	0	0	0	0	300

Source: Delaware Department of Education. (2010). *For the 2009-2010 school year: State accountability in Delaware*. Retrieved from, http://www.doe.k12.de.us/aab/accountability/Accountability_Files/School_Acct_2009-2010.pdf

The choice of values for a transition matrix can depend on several factors, such as policy and accountability decisions, the number of performance levels, the perceived difficulty in making

certain jumps in performance levels, and the time horizon for reaching a desired performance level. The relative advantage of the value table is that it can set clear incentives for schools for particular achievement level transitions. Although the accuracy of individual growth reporting and prediction may degrade due to the loss of information into broad categories, the categorical model can clearly communicate the relative priorities of educational policies. Section 3.5 further delves into important considerations when setting values.

Question 3.3:

What are the *Required Data Features* for the Categorical Model?

The categorical model requires student achievement levels at each time point of interest. These achievement levels are defined by cut scores and qualitative descriptions relating to student proficiency. Interpreting the transition between achievement level categories as growth requires an implicit vertical scale.

The categorical model only requires student test scores reported in achievement levels like Basic, Proficient, and Advanced. The mapping of scores to achievement levels requires decisions about the number of achievement levels, the descriptions of these levels in terms of student performance, and the cut scores that divide the achievement categories on the score scale.

State testing programs commonly set achievement level cut scores in the process of test development. However, these categories may be insufficient for supporting growth interpretations in a categorical model. If a state decides to use a categorical model for reporting growth to proficiency but only has three performance levels currently in place — Basic, Proficient, and Advanced — then a student cannot be deemed as “on track” to Proficient without actually reaching the proficiency performance level. If a Basic student moves up one level, that student is not on track to proficiency, that student is simply Proficient. In these situations, it is useful to subdivide the Basic category to facilitate finer-grain tracking of student progress toward proficiency.

An essential requirement of the categorical model is that achievement levels must be articulated across the grade levels for which the growth model is applicable. Cross grade-level performance levels are linked in several fundamental ways. First, tests in each grade-level of interest must have the same set of performance levels. In other words, if the Grade 3 levels are Low-1, Low-2, Intermediate, Proficient, and Advanced, then the Grade 4 levels must also be Low-1, Low-2, Intermediate, Proficient, and Advanced and likewise for all other grades of interest. Second, although the cut scores that classify students into each of these categories may change for each grade-level, compared to the other performance

levels, a particular performance level should correspond to the same *relative* achievement level each year. Moreover, the performance levels in and across grades must be aligned to some underlying continuum of mastery. Under these conditions, it is meaningful to attach interpretations of progress or growth to a change from Low-1 in Grade 3 to Low-2 in Grade 4. Once such interpretations are made, however, even if the tests do not have an explicit vertical scale, model users are implicitly assuming a vertical scale exists across all the grade levels of interest.

Question 3.4:

What Kinds of *Group-Level Interpretations* can the Categorical Model Support?

At the group-level, the two most typical statistics reported for the categorical model are the percentage of students “on track” to a desired performance level, like proficiency or college and career readiness, and the average transition value over all the students in a group.

Like the trajectory model, the categorical model is often implemented as a way to monitor and incentivize progress toward a desired performance level, such as proficiency or college and career readiness. Accordingly, a natural statistic to summarize group-level growth under this model is the percentage of students on track to the desired performance level. An alternative group-level statistic, particularly when weights are differentially attached to transitions (see Table 3.2), is the average transition value for all the students in the group.

The percentage of on-track students describes group growth in terms of progress toward a desired goal. If a large percentage of students is making progress, this suggests that the group is generally improving with respect to a future standard. As with trajectory models, the percentage of on track students is either added to the percentage of proficient students or re-expressed as a percentage of students eligible to be on track. These percentages can themselves be compared to benchmarks such as Annual Measurable Objectives or other minimum required percentages.

Another useful feature of value tables is that average values for groups are interpretable as a kind of average growth. For a simple case where a value table's cells correspond to the number of categories a student has gained or declined, the average over all students is the average gain in categories for that particular group. More generally, value tables like those in Figure 3.2 can be compared against the value scheme, in this case, a 0 to 300 scale, to gauge whether students are generally making transitions toward the desired target. An additional standard setting procedure may be used to determine whether averages of value tables are sufficient for particular groups.

Question 3.5:

How Does the Categorical Model Set Standards for Expected or Adequate Growth?

The categorical model is more dependent on judgmental standard setting procedures than most growth models. The scores that support growth calculations are achievement level categories determined by standards. Additional judgments must be incorporated to determine which category transitions are sufficient or what value they should be assigned. A third level of standard setting may be useful for evaluating whether group-level average growth is sufficient.

In categorical models, growth is operationalized as a transition between categories. Any increase in a category may be deemed as adequate. Or, a relative value can be assigned to each transition as in Table 3.2. The value table framework adequately captures the scope of the standard setting task. It also illustrates the amount of control that policy designers can have in communicating the desired incentive structure to stakeholders.

In simple models where any category gain is sufficient, an additional implication is that the student is on track to successively higher categories in the future. In this way, the categorical model functions as a coarse trajectory model, where a gain of one category is extrapolated and assumed to extend to future time points until proficiency is eventually met.

For group growth, whether the growth statistic is the percentage of on-track students or the average of value table scores across students, separate standard-setting procedures will be required to establish whether these group growth magnitudes are sufficient.

A feature of the categorical model is that no intuitive standard for growth arises naturally from the model. There is instead a degree of control in the form of the value table. The value table is at once transparent in its dependence on user input and deceptive in its coarseness and in its functioning as an implicit vertical scale.

Question 3.6:

What are the Common Misinterpretations of the Categorical Model and Possible Unintended Consequences of its Use in Accountability Systems?

Although categorical models do not require a vertical scale in a strict sense, the articulation of multiple cut scores across grades represents an implicit vertical scale that requires the same critical attention as vertical scaling. The grouping of scores into coarse categories leads to a loss of information in reporting both status and growth.

Although the categorical model does not require a vertical scale in the strict sense, the previous sections have demonstrated that growth interpretations from categorical models require interpretation of the articulated cut scores as an implicit vertical scale. If a transition from Below Basic in one grade to Basic in the next grade is interpretable as growth, then the cut score must share some common meaning across grades, not just in relative stringency, but in the content domain as well. If the model also assumes that a transition across one category boundary predicts a transition across subsequent category boundaries, then the categorical model acts as a coarse trajectory model and requires the same attention to its underlying vertical scale.

The grouping of the scores into categories leads to a loss of information both in the reporting of scores and the description and prediction of growth. As Figure 3.2 demonstrates, the categories represent a kind of relative stringency that may or may not conflict with user intuition about growth. More problematically, a broad range of implicit gain scores will be mapped into the same transitions, and gain scores that are equal lead to a category gain in some cases and not in others. These facts suggest that the reporting of categorical model results should be limited or withheld at the student level.

At the school level, the categorical model is clearer than other models in its communication of differentiated incentives for different transitions, particularly when values in value tables are carefully considered. Although the values may seem arbitrary, they are no less arbitrary than assuming that gain scores should count equally, as the gain score model generally does, or that students should be on track to a particular standard by a particular time horizon, as a trajectory model can do. However, because the categorical model shares the same underlying statistical foundation as gain score and trajectory models, it also shares the undesirable feature where the artificial deflation of initial scores (in this case, categories) will inflate the observed transitions of students. This can be seen in Table 3.2, where, in any given column, points are maximized when students are in lower initial categories. This is the same underlying, “gaming” mechanism that can inflate gain scores and trajectories in the models in the two previous chapters.

CHAPTER 6

The Student Growth Percentile Model

The Student Growth Percentile (SGP) model offers a normative foundation for the calculation and interpretation of growth. Although this model uses a relatively complex statistical framework, the procedure is open-source, well described, and explainable with accessible, visually appealing graphics (Betebenner, 2009). Because the SGP model is a relatively recent and popular development, this chapter will offer a particularly detailed exposition.

Damien Betebenner's SPG model (Betebenner, 2010b) involves two related procedures resulting in 1) student growth percentiles, which will be referred to as "SGPs," and 2) percentile growth trajectories (see further discussion of Betebenner's model in the following pages). These primarily support interpretations of **growth description** and **growth prediction**, respectively. SGPs locate current student status relative to past performance history and thus use a **conditional status** statistical foundation. SGPs answer the question

What is the percentile rank of a student compared to students with similar score histories?

Simplistically, SGPs describe the relative location of a student's current score compared to the current scores of students with similar score histories. The location in this reference group of "academic peers" is expressed as a percentile rank. For example, a student earning an SGP of 80 performed as well as or better than 80 percent of her academic peers.

A strict implementation of this procedure would seem to involve the selection of "academic peers" that have identical previous scores. This is impractical and imprecise with large numbers of prior grade scores. Regression-based methods can address this problem, but, as described in previous chapters, linear

STUDENT GROWTH PERCENTILE MODEL

Aliases and Variants:

- The Colorado Model
- Percentile Growth Trajectories
- Conditional Status Percentile Ranks

Primary Interpretation:

Growth description
Growth prediction

Statistical Foundation:

Conditional status model

Metric/Scale: Percentile rank (whole numbers 1 - 99)

Data: Set of psychometrically sound tests over two or more grade levels in a single domain and large sample sizes

Group-Level Statistic: Median/mean SGP – describes the average/typical status of students relative to their past performance, or percentage of students on-track (to a future standard)

Set Growth Standards:

Requires judgment about an adequate SGP or median/average SGP. Predictions require a future standard and a time horizon to meet the standard.

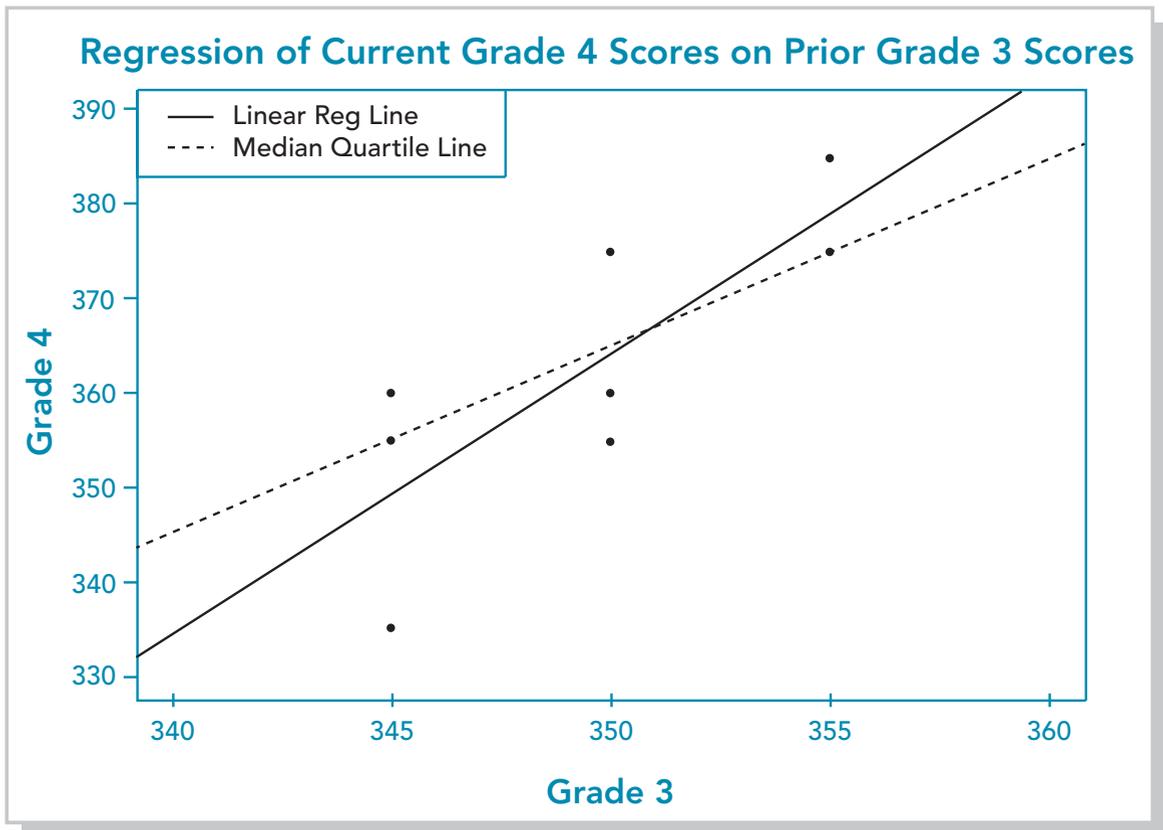
Operational Examples:

NCLB Growth Model (e.g., Colorado and Massachusetts)

regression methods require 1) assumptions of linear relationships between predictors and outcomes and 2) equal variability in current scores across prior-year scores. The computation of SGPs involves a more flexible statistical tool called *quantile regression* that loosens these requirements to fit a broader range of test score distributions in practice. The software that estimates SGPs is open-source and freely available in the statistical software package, R.

Figure 6.1

Illustration of a Simple Linear Regression Line (that models the conditional average) and the Median Quartile Regression Line (that models the conditional median)



A simple linear regression model, like the one shown by the solid black line in Figure 6.1, results in a single line that represents the best prediction of an outcome variable (current status) by a predictor variable (past performance). Equivalently, this line represents a “conditional average,” the average value of the outcome at each level of the predictor. In Figure 6.1 and in real data, the line represents an approximation of the conditional averages — a best guess about the value of an outcome given a predictor.

Instead of fitting one line for the conditional average, the SGP model fits 99 lines, one for each conditional percentile, 1 through 99. As a point of reference, the 50th line is the line for the

conditional median, and it is shown by the dashed black line in Figure 6.1. Typically, for real statewide datasets, the median quantile regression line and the simple linear regression line will likely be closer together than they are in this illustrative example, which is based on a very small dataset. This conditional median line represents the best guess about the median of an outcome given a predictor, just as the usual regression line represents the best guess about the average of an outcome given a predictor. Points closest to this conditional median line will be assigned an SGP of 50. For instance, two students actually lie on this line — the middle Grade 4 scoring student of the three students who scored 345 in Grade 3 and the lower Grade 4 scoring of the two students who scored 355 in Grade 3. These two students will receive SGPs of 50. Students at points above the conditional median line will be assigned SGPs higher than 50 according to the conditional percentile lines to which they are closest and vice versa for students at points below this line.

For illustrative purposes, this chapter explains the empirical calculation of SGPs in a simplistic case with limited data. This empirical method is analogous to operational SGP calculations and provides intuition about the statistical machinery underlying SGPs. We refer the interested reader to the SGP R package and references by its primary author, Betebenner, for a full description of operational SGP computations.⁵

An extension of the SGP model known as “percentile growth trajectories” supports **growth predictions**. The approach has similarities to both the trajectory model and the projection model, where SGPs are extrapolated and assumed to be maintained over time. This prediction helps to answer the question

Assuming the student maintains her SGP over time, what will her future score be?

This future score can be compared to a target future standard to support an “on track” designation. In this standards-based context, an alternative framing is captured by the question

What is the minimum SGP a student must maintain to reach a target future standard?

When determining whether students are “on track,” these two questions are functionally equivalent. Determining whether a student’s predicted future status exceeds the future standard is equivalent to determining whether the student’s trajectory exceeds the minimum required trajectory. This equivalence was established in the context of the trajectory model in Section 2.5. Both the trajectory model and the percentile growth trajectories procedures involve an assumption of students continuing on their same “growth” path. The trajectory model operates under the assumption of linear growth, where students maintain constant gains each year. The percentile growth trajectories, in contrast, assume students maintain constant ranks with respect to their academic peers each year.

The percentile growth trajectory procedure is also similar to the projection model, in that growth

⁵ See Betebenner (2009; 2010a; 2010b).

predictions require data from a cohort of students that has already reached the target grade of interest. These reference cohorts provide the hypothetical trajectories for each student's extrapolated SGP over time. However, percentile growth trajectories are less data driven than the projection model. Previous data are used to estimate where consecutively maintained SGPs will lead into the future, but the data are not used to predict whether or not students will actually consecutively maintain these SGPs. Thus, percentile growth trajectories, like the trajectory model, make an aspirational, descriptive assumption that a measure of growth is maintained over time.

Question 6.1:

What *Primary Interpretation* Does the Student Growth Percentile Model Best Support?

The SGP model supports growth description with SGPs and growth prediction with percentile growth trajectories.

This guide considers growth models less as coherent packages than as collections of definitions, calculations, and rules. The SGP model is an example of this, where SGPs describe growth through one procedure, and percentile growth trajectories predict growth through an additional layer of assumptions. These latter assumptions include students' maintenance of SGPs over consecutive years. The distinction between SGPs and percentile growth trajectories is analogous to the distinction between the gain-score model and the trajectory model, but this chapter discusses both given the unfamiliar statistical machinery that they both share.

SGPs describe the relative performance of students by comparing their current scores to those of a set of students with similar scores on prior grade-level tests. The SGP metric expresses this relative status in terms of percentile ranks. Typically, SGPs are expressed as whole number values from 1 to 99. By creating norm groups of students with similar past scores, both low- and high-performing students can theoretically receive any SGP from 1 to 99. In other words, SGP models will typically have zero or near-zero associations between status and SGPs, a unifying feature of conditional status models. In contrast, gain-based models can have these associations built into the vertical scale, ideally to reflect true changes in the variability of student achievement over time. From the perspective of growth description, these associations may be desirable to the extent that they reflect true growth over time. From the perspective of evaluation for accountability, these associations may seem unfair.

If the desired use of the growth model is to predict future student performance, the SGP model can be extended to provide percentile growth trajectories. These trajectories assume that students will maintain their SGPs through to the future, continuing to obtain scores at the same relative rank with respect to their academic peers. In practice, 99 different percentile growth trajectories can be computed starting at each score point and continuing into the future. For a group of 30 students who happen to have

30 different current scores, there will be $30 \times 99 = 2970$ possible trajectories, 99 for each student. The predicted trajectory for each student is the one that corresponds to his or her current SGP.

Each percentile growth trajectory assumes that a student at a particular starting score will have a particular SGP and maintain that SGP each year. In this way, the percentile growth trajectory that corresponds to a student's actual SGP will lead to a predicted score in the future. This score can be compared to a target score at a time horizon, or, equivalently, the student's actual SGP can be compared to the SGP required to reach the target future score. The derivation of these trajectories is described later in this chapter.

Question 6.2:

What is the *Statistical Foundation Underlying the Student Growth Percentile Model*?

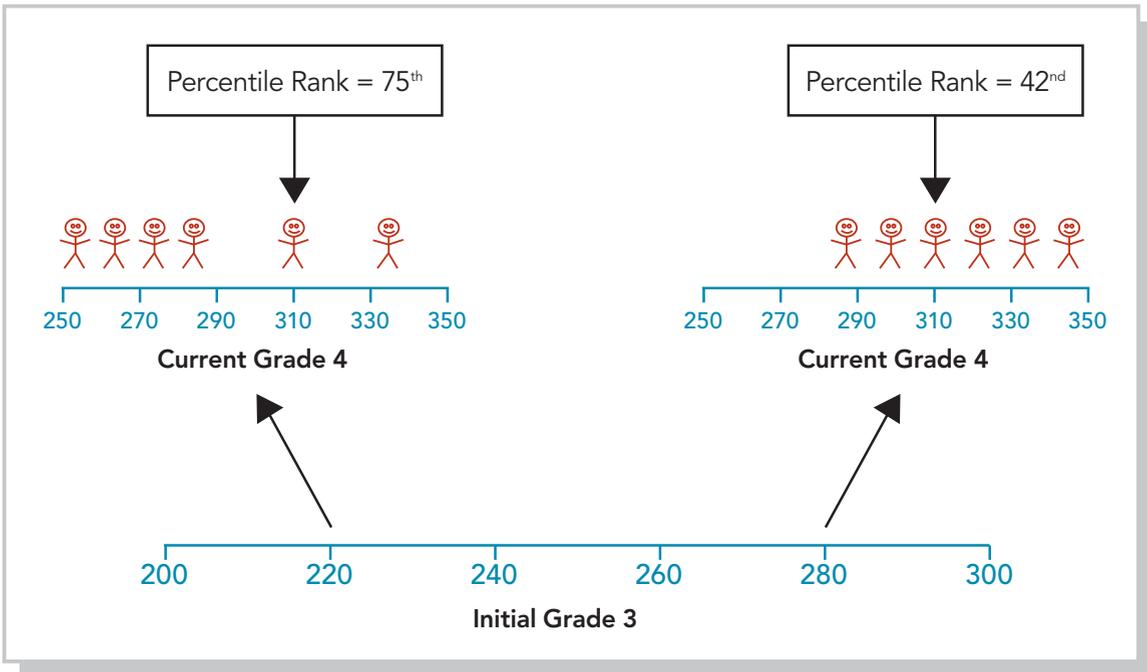
The SGP model is a conditional status model.

SGPs represent conditional status. They re-express a student's current score as a percentile rank in a theoretical distribution of students with identical past scores. This statistical foundation is best understood through an illustration of the computation of SGPs. The SGPs currently used by states like Colorado and Massachusetts rely on a statistical tool called quantile regression. The term "quantile" is general and includes "percentile" as a special case, and, in fact, the statistical method underlying the SGP model is more literally "percentile regression." We begin with a heuristic example that introduces the central idea supporting interpretations of SGPs — the academic peer group. Although this is not precisely the way SGPs are estimated in practice, it is a useful intuitive aid that supports understanding of the actual procedure.

Figure 6.2 introduces a longitudinal dataset for a cohort of Grade 4 students with one prior year of Grade 3 scores. Like the conditional status models from the two previous chapters, SGPs can accommodate scores from any number of prior grade levels and other non-test-score variables as well, but this one-prior-year case will suffice as an illustration. The initial Grade 3 score scale has scores ranging from 200 to 300 and represents the "initial status" of students in this cohort. Arrows are located at Grade 3 scores of 220 and 280 to focus exclusively on the students who earned these particular Grade 3 scores. Six students earned a score of 220 on the Grade 3 test, and six other students earned a score of 280. These students are represented by stick figures located above their "current" Grade 4 score on a score scale that ranges from 250 to 350. In each set of students, one student earned a score of 310 on the Grade 4 test, which, in this hypothetical scenario, reflects an above-average score. Although these two students earned the same current Grade 4 score, they are in different relative positions among their "academic peers," their peers with the same Grade 3 scores. The percentile ranks of these two students are displayed in boxes above their heads.

Figure 6.2

Illustration of a Heuristic Approach to Computing Student Growth Percentiles



The percentile ranks of these two students are heuristic estimates of their SGPs, their percentile ranks within their group of “academic peers.” The percentile rank calculation follows simply from their ranks. Given the small number of students in each group of academic peers, we use the following percentile rank formula that has a slight adjustment for small, discrete variables.

$$\text{Percentile Rank} = \frac{\text{Number of students below Score} + (.5 * \text{Number of students at Score})}{\text{Number of students in the academic peer group}}$$

This formula allows for calculation of any student’s percentile rank relative to their academic peers by simply counting the number of students below and at the student’s score. Among the six students who scored 220 in Grade 3, the student who scored a 310 in Grade 4 has four students scoring strictly below her and only one student, herself, scoring at her score. Her percentile rank is then

$$\begin{aligned} \text{Percentile Rank} &= \frac{\text{Number of students at or below 310} + (.5 * \text{Number of students at 310})}{\text{Number of students in the academic peer group}} \times 100 \\ &= \frac{4 + (.5 * 1)}{6} \times 100 \\ &= \frac{4.5}{6} \times 100 = 75 \end{aligned}$$

This supports a statement like, “This student performed as well as or better than 75 percent of her academic peers.” Among the six students who scored a 280 in Grade 3, the student who scored a 310 in Grade 4 has two students scoring strictly below his score and only himself scoring at his score. His percentile rank is then

$$\begin{aligned}\text{Percentile Rank} &= \frac{\text{Number of students at or below 310} + (.5 * \text{Number of students at 310})}{\text{Number of students in the academic peer group}} \times 100 \\ &= \frac{2 + (.5 * 1)}{6} \times 100 \\ &= \frac{2.5}{6} \times 100 \approx 75\end{aligned}$$

This supports a similar statement, “This student performed as well as or better than 42 percent of his academic peers.”

The SGP model does not actually divide students into groups with identical past scores. This heuristic approach would result in intractably small groups when there are multiple prior year scores. With one prior year as in Figure 6.2, the numbers of students with the same prior year scores may be large. However, with two or more years, the numbers of students with the exact same prior year scores will dwindle and become unsupportable as a reference group. Instead, the SGP model performs a kind of smoothing that borrows information from nearby academic peer groups to support the estimation of percentile ranks. Even though increasing the number of prior year scores will diminish the sizes of groups of students with identical past scores, this borrowing of information allows for continued support of SGP estimation.

The actual calculation of SGPs involves the estimation of 99 regression lines,⁶ one for each percentile from 1 to 99. In Figure 6.1, this can be visualized by 99 lines that curve from the lower left to the upper right and try to slice through their respective percentiles at each level of the Grade 3 score. For example, the 50th regression line is given by the dashed black line and estimates the median Grade 4 score at each Grade 3 score. This line passes through the central score of the trio of students who scored 345 in Grade 3. It does not pass exactly through the central score of the trio of students who scored 350 in Grade 3 because the line is pulled upwards by the students who scored a 355 in Grade 3. This median regression line can support interpretations like, “Students with a Grade 3 score of 350 have a predicted median Grade 4 score of 365.” Accordingly, students with Grade 3 scores of 350 and observed Grade 4 scores of 365 have a SGP of 50. The 90th regression line will lie above the 50th regression line

⁶ Technically, the SGP model estimates regression lines only when there is a single prior year score. With two prior year scores, these are regression surfaces in a three dimensional space. With three or more prior year scores, these are regression hypersurfaces in multidimensional space.

and may, for example, predict a Grade 4 90th percentile of 375. Students that are closest to the 90th regression line will be above the median regression line shown in Figure 6.1 and will be assigned an SGP of 90.

This SGP of 90 indicates that this student performed as well as or better than 90 percent of her academic peers. In practice, this will be an estimate that not only estimates percentile ranks for students with the exact same previous scores, but also borrows information from “nearby” students with similar, but not identical, past scores. This frames the academic peer group as more of an academic neighborhood. This is illustrated by the fact that that median regression line in Figure 6.1 does not go directly through the central score for students who scored a 350 in Grade 3; rather, the line is pulled up by the students who scored a 355 in Grade 3.

This metaphor extends to all conditional status metrics. SGPs, like residual gain scores, describe growth in terms of relative status in an academic neighborhood. This conditional status is normative and cannot be interpreted in terms of an absolute amount of growth on any developmental scale. If there is an underlying vertical scale score with sound properties, there would be no way to tell which SGPs, if any, would be associated with negative growth. Conditional status is also dependent on the definition of the academic neighborhood, which changes with the addition of additional prior grade scores or other predictor variables. These are not shortcomings but reminders that conditional status metrics support a contrasting perspective on growth.

Question 6.3:

What are the Required Data Features for the Student Growth Percentile Model?

The SGP model requires test scores for large numbers of students to support stable estimation of SGPs.

Part of the appeal of SGPs and other conditional status metrics is that they do not require test scores from multiple time points to share a common vertical scale. The SGP model is also more flexible than the residual gain model in that neither linear relationships nor common outcome variance across predictor levels is required. However, this flexibility can come at a cost, as SGPs require estimation of large numbers of parameters for the 99 regression lines. This requires sufficient data. A loose rule of thumb is to include at least 5,000 students, but, like all guidelines, this can depend on a number of factors; in this case, it depends on the interrelationships between the variables and the number of prior years of data included (Castellano & Ho, in press). Estimation tends to be most problematic for outlying students on one or more test score distributions. These students can receive highly unstable SGPs as there are too few students in the same academic neighborhood to obtain stable relative ranks.

Question 6.4:

What Kinds of *Group-Level Interpretations* can the Student Growth Percentile Model Support?

SGPs are often summarized at the group-level with a median SGP that represents the SGP of a typical student. It is also possible to use a simple average of SGPs for a group. In either case, aggregated SGPs provide descriptive measures of group growth. In the context of growth prediction, percentile growth trajectories can support calculation of percentages of students predicted to be on track to reaching a desired standard.

The SGP model provides useful norm groups for describing student status. However, school administrators and policymakers are often more interested in summary measures of student growth than individual growth results. SGPs can easily be aggregated for any group of students by taking the median or mean of the SGPs. In practice, median SGPs are the most common aggregate SGP metric. The median function is motivated by the fact that SGPs are percentile ranks and are thus on a scale that is generally not recommended for averaging (Betebenner, 2009). Others have shown that averages or averages of transformed percentile ranks can in some cases support more stable aggregate statistics (Castellano & Ho, in press). Castellano, K. E. (2012). Contrasting OLS and quantile regression approaches to student “growth” percentiles. *Journal of Educational and Behavioral Statistics*. Advance online publication. doi: 10.3102/1076998611435413

These simple aggregates of SGPs support descriptions of group growth, whether the groups are classrooms, schools, or districts. They summarize the distribution of SGP with an average or typical value from the group. These measures can thus be described with statements like, “The average fourth grade student in School A performed as well as or better than 55 percent of her academic peers.” SGPs are generally not recommended for the support of causal, or value-added, interpretations on their own (Betebenner, 2009). That is, they are not recommended in support of interpretations like, “The fourth grade teachers at School A are the cause of this higher-than-expected performance.”

SGPs for a group can also be summarized by other statistics and graphical displays. These can augment simple averages to provide a fuller picture of the distribution of SGPs for particular groups. Additionally, the relationship between group SGPs and group status can be displayed to communicate the distinction between high and low average status and high and low average growth.⁷

In the context of growth prediction, percentile growth trajectories can be summarized at the group level by calculating the percentage of students who are designated as on track to the target future score. This is described in further detail in this next section.

⁷ For further information, this Colorado Department of Education website includes examples of attractive SGP-related graphics summarizing school and district performance: <http://www.schoolview.org/ColoradoGrowthModel.asp>.

Question 6.5:

How Does the Student Growth Percentile Model Set Standards for Expected or Adequate Growth?

Like the residual gain model, the SGP model sets empirical expectations for growth through the estimation of percentile regression lines. However, this statistical machinery is not sufficient to determine which SGPs are “good enough,” and additional standards may be desired to support interpretations on the SGP scale. For growth prediction, percentile growth trajectories can be compared to a future target score, such as the Proficient cut score in a target grade level. They can also be used to determine the minimum SGP a student must maintain to reach the future target score.

An essential step in implementing most growth models is the definition and communication of adequate growth. These determinations are useful at both the student and the group level. The Colorado Department of Education (CDE) uses SGPs of 35 and 65 to distinguish among low, typical, and high growth (CDE, 2009). In contrast, the Massachusetts Department of Elementary and Secondary Education (MDESE) defines 5 growth categories at the student level: Very Low, Low, Moderate, High, and Very High. These are delineated by SGP cuts of 20, 40, 60, and 80 (MDESE, 2009). These classifications support growth reporting and accurate user interpretation of SGPs. At the aggregate level, median SGPs can also be evaluated with respect to standards, where the most common standard in practice is a simple cut score set at 50 that delineates groups with higher and lower growth than expected.

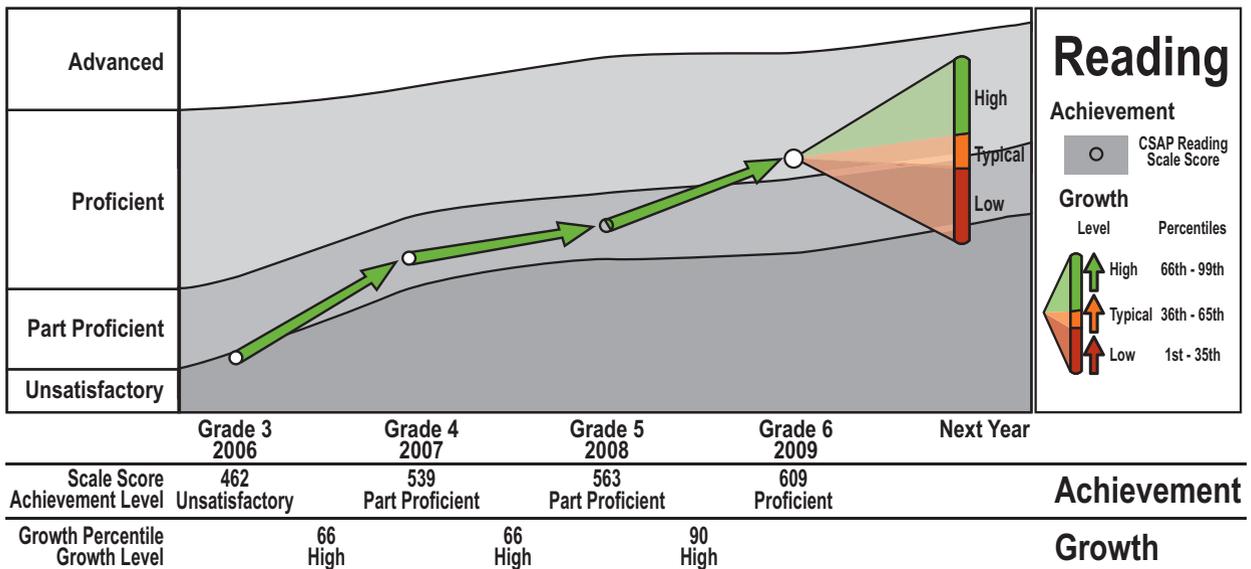
A higher-level standard setting approach arises from an extension of SGPs to support growth predictions. These “percentile growth trajectories” can support inferences about student trajectories toward a particular standard, such as Proficient or College and Career Ready. Percentile growth trajectories combine aspects of the projection and trajectory models. Like the projection model, percentile growth trajectories are found by estimating regression equations using cohorts of students who already have scores from the future target grade level. These prediction equations are then applied to students whose future trajectories are of interest. Like the trajectory model, percentile growth trajectories assume that students will maintain constant gains each year. For percentile growth trajectories, a constant gain is the maintenance of the same SGP each year into the future. This is akin to an assumption of continued relative gains.

The trajectory model can both predict a future score and report the minimum gain necessary to achieve a future standard. Similarly, percentile growth trajectories can predict where a student will be in the future and also report the minimum SGP that must be maintained to reach the future target. Percentile growth trajectories can also report a range of future outcomes associated with the maintenance of different SGP levels. Figure 6.3 reproduces a plot from a presentation by

Betebenner (2011) that shows a range of percentile growth trajectories for a student. These plots are rich with information about student status, growth, and predicted growth.

Figure 6.3 shows one student’s observed Reading scores from Grades 3 to 6 with predictions to Grade 7. This student is currently in Grade 6, scored a 609 on the reading achievement test, is Proficient, and given her scores in Grades 3, 4, and 5, scored an SGP of 90. In the next year, there is a distribution of colors — green, yellow, and red — showing where the student is predicted to fall if the student scores a high, typical, or low SGP next year. These predictions are constructed from percentile growth trajectories one year into the future. Although all 99 percentile growth trajectories are not specified in the figure, the color bands summarize the span of trajectories across the SGP range. The color classifications are based on Colorado’s SGP cut scores of 35 and 65.

Figure 6.3
An Illustration of Percentile Growth Trajectories



Source: Betebenner (2011). Retrieved March 29, 2012, from <http://ccsso.confex.com/ccsso/2011/webprogram/Session2199.html>. This figure was generated using the “studentGrowthPlot” function using the SGP package and R software. Several states are currently using this package to produce student reports for their state assessment programs.

Figure 6.3 also shows that the student will continue to be proficient if she has a high SGP, but a typical SGP will result in a decline from proficient to partially proficient. A particularly low SGP could result in a decline to the “unsatisfactory” category. The figure emphasizes the importance of standard setting, not only in the definition of high, typical, and low growth, but in the articulation of standards across grades. The figure also masks an essential assumption

underlying the plot: a vertical scale underlies all of the grade level tests. Without an assumed or actual vertical scale, these kinds of plots cannot be constructed. With a vertical scale, alternative gain-based models become possible and represent useful contrasts.

Question 6.6:

What are the *Common Misinterpretations of the Student Growth Percentile Model and Possible Unintended Consequences of its Use in Accountability Systems?*

Student Growth Percentiles are often incorrectly assumed to describe an absolute amount of growth in a normative frame of reference. They are instead a relative metric in two ways, both with respect to the variables included as predictors and with respect to other students in the model. Group-level SGPs may be overinterpreted as value-added measures when they are not intended to support these inferences on their own.

A literal interpretation of a growth percentile is one where growth is expressed as a percentile rank. This might entail describing an absolute growth measure like a gain score in terms of its rank relative to other gain scores. This percentile rank of gain scores is a gain-based expression that is a natural extension of a gain-score model. In contrast, SGPs represent a relative metric in at least two ways. First and most intuitively, like any percentile rank, SGPs describe growth normatively with respect to a particular reference group. Second and less intuitively, the SGP — and any conditional status approach to growth — defines status relative to other variables in the model.

In the case of SGPs, these predictor variables are the prior grade scores that set expectations for current status. As such, adding or removing prior grade variables will alter SGPs, because expectations about status will change when expectations are based on different pieces of information. Of course, gain-based models will also change as prior-grade variables are added, but the quantity estimated in gain-based models (the average gain or slope) generally improves as more information is added. In conditional status models like SGPs, the addition of information fundamentally changes the expectations and therefore the substantive definition of the quantity being estimated.

As an example of this, assume that a fifth grade student with a prior year of fourth grade data has an SGP of 90. Say that a research analyst uncovers an additional previous year of data from third grade, recalculates all SGPs, and finds that the student now has an SGP of 50. Is the student's true SGP 50, 90, or somewhere in between? There is no single answer to this question. The SGP of 90 compares the student's current status to academic peers defined by fourth grade scores. The SGP of 50 compares the student's current status to academic peers defined by third and fourth grade scores. If it seems that more grades allow for an improved definition of academic peers, then why not improve the definition further by including demographic variables?

Expectations change based on the predictors used to set expectations, thus there is no immediately obvious answer to the question of which SGP is “true.” In contrast, if a student gains 10 points from Grades 3 to 4 and 90 points from Grades 4 to 5, there is a clearer argument for averaging these gains to obtain an average gain. This is not an inherent advantage of gain-based models or a disadvantage to conditional status models. Conditional status should depend upon the variables used to set expectations, and this is preferred if there is substantive interest in these expectations. The distinction emphasizes that these two statistical foundations support fundamentally different conceptions of growth.

Like gain-based models and, more directly, residual gain models, SGPs can be artificially increased by deflating initial year scores. In the intuition of SGPs, this deflation changes the academic peer group of students to one that will tend to be lower scoring, resulting in an inflated SGP. As a corollary, this will also inflate percentile growth trajectories. As with other models, these incentives can be diminished through a thoughtful combination of status and growth model.

References

- Betebenner, D.W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4), 42–51, from <http://www.ksde.org/LinkClick.aspx?leticket=UssiNoSZks8%3D&tabid=4421&mid=10564>.
- Betebenner, D.W. (2010a). *New Directions for Student Growth Models*. Dover, NH: National Center for the Improvement of Educational Assessment. Presentation dated December 13, 2010 from <http://www.ksde.org/LinkClick.aspx?fileticket=UssiNoSZks8%3D&tabid=4421&mid=10564>.
- Betebenner, D.W. (2010b). *SGP: Student Growth Percentile and Percentile Growth Projection/Trajectory Functions*. (R package version 0.0-6).
- Betebenner, D.W. (2011). *New directions in student growth: The Colorado growth model*. Paper presented at the National Conference on Student Assessment, Orlando, FL, June 19, 2011, from <http://ccsso.confex.com/ccsso/2011/webprogram/Session2199.html>.
- Castellano, K.E., and Ho, A.D. (in press). Contrasting OLS and quantile regression approaches to student “growth” percentiles. *Journal of Educational and Behavioral Statistics*.
- Colorado Department of Education (CDE). (2009). *The Colorado growth model: Frequently asked questions*. Retrieved April 27, 2012, from <http://www.schoolview.org/GMFAQ.asp>.
- Massachusetts Department of Elementary and Secondary Education (MDESE). (2009). *MCAS student growth percentiles: State report*. Retrieved March 29, 2012, from <http://www.doe.mass.edu/mcas/growth/StateReport.pdf>.

