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

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.

TABLE OF CONTENTS

LIST OF TABLES

PART I

PART II

APPENDIX A

LIST OF FIGURES

PART I

PART II

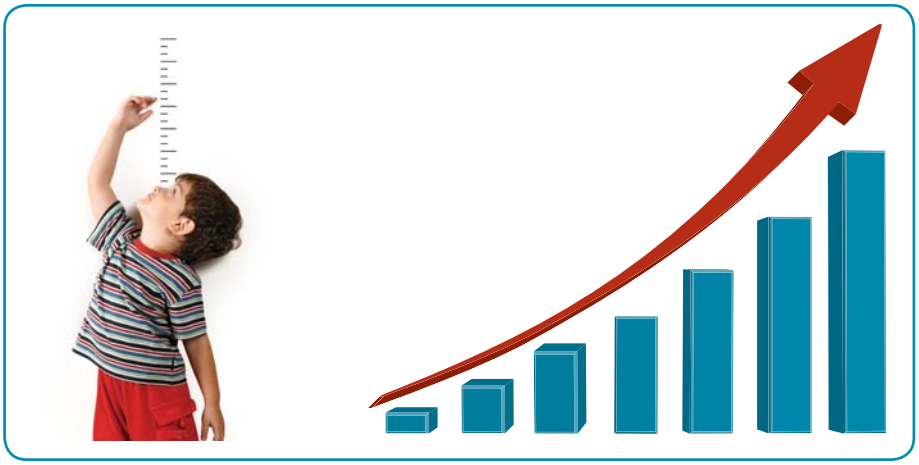
PART I
A FRAMEWORK FOR OPERATIONAL GROWTH MODELS

If names are not correct, then language is not in accord with the truth of things.
If language is not in accord with the truth of things,
then affairs cannot be carried out successfully.
— Confucius

1 - Growth and Growth Models

Growth refers to an increase, expansion, or change over time. A common metaphor is that of a child growing in height or weight, where growth is tracked easily as the change in inches and ounces over time. Asked to pantomime "growth," one might shrink into a crouch, mimicking a small child, and then jump up and out with arms and legs spread, emphasizing a two-stage, transformative process. Asked to draw growth, one might draw a graph with an arrow starting in the lower left and pointing to the upper right. Implicit in this graph is a vertical axis indicating a quantity of interest and a horizontal axis representing time. Figure 1.1 shows two of these intuitive representations of growth.

Figure 1
Intuitive Depictions of Growth



If growth models for educational policy followed this commonsense intuition about growth, there would be little need for this guide. Instead, statistical models and accountability systems have become increasingly varied and complex, resulting in growth models with interpretations that do not always align with intuition. This guide does not promote one type of interpretation over another. Rather, it describes growth models in terms of the interpretations they best support and, in turn, the questions they are best designed to answer. The goal of this guide is thus to increase alignment between user interpretations and model function in order for models to best serve their desired purposes: increasing student achievement, decreasing achievement gaps, and improving the effectiveness of educators and schools.

A Practitioner's Guide to Growth Models begins by overviewing the growth model landscape, establishing naming conventions for models and grouping them by similarities and contrasts. It continues by listing a series of critical questions or analytical lenses that should be applied to any growth model in current or proposed use. The remainder of the guide delves systematically into each growth model, viewing it through these lenses.

This guide is structured like a guidebook to a foreign country. Like a guidebook, it begins with an overview of central features and a presentation of the landscape before proceeding to specific regions and destinations. Although it can be read from beginning to end, a typical user may flip to a model that he or she is using or considering for future use. Although the guide is structured to support this use, readers are encouraged to peruse the beginning sections so that, following the analogy, they can appreciate the full expanse of this landscape.

2 - Growth: Beyond Status

In the practice of modeling growth, the operational definition of growth does not always align with the intuitive definition of growth. If this were a guide only for the growth models that aligned with intuition, it would be a short guide that excluded a number of models in active use across states. Although these models may be less intuitive, they often answer useful questions about longitudinal data that "intuitive" growth models do not answer. To be useful, a broader working definition of growth is necessary.

When defining a term, it is often easier to begin with what it is not. Among all the discussions of student and group growth using educational assessment data, there is one underlying common thread — "growth is not status." Accordingly, to develop a definition of growth we must first define status. Fortunately, defining status is a much easier task than defining growth.

Status describes the academic performance of a student or group (a collection of students) at a single point in time.

This simple definition of status provides a contrast that allows us to define growth. Student status is determined by data from a single time point and provides a single snapshot of student achievement, whereas any conception of academic growth is determined from data over two or more time points, taking into account multiple snapshots of student achievement. With this distinction from status, a simple working definition of growth arises.

Growth describes the academic performance of a student or group (a collection of students) over two or more time points.

Growth models, in turn, use some systematic method, usually mathematical or statistical, to describe the academic performance of a student or group over two or more time points. This growth definition is deliberately broader than the more intuitive definition of growth as the change in academic achievement over time. The essential components of the definition are 1) multiple time points and 2) a temporal distinction between at least two of these time points. For example, the average of two student test scores from a fall and spring test administration is not a growth metric, because the average is blind to which score came earlier and which score came later in time. The following sections review additional conceptions of growth and, in turn, growth models.

3 - Different Ways to Slice the Data: Status, Improvement, and Growth

This guide's general definition of growth is an entry point into the tangled web of descriptions for growth models. Table 1.1 below and the following Tables 1.2 to 1.4 all show the same hypothetical aggregated data for a particular school but highlight different cells to emphasize additional distinctions between status and growth. In each of these tables, the rows designate grades, and the columns designate years. The cells contain hypothetical average Mathematics test scores for all students in a particular school. In Table 1.1 in particular, the shaded cell reports 320 as the average Mathematics score of 3rd graders in 2007: a single grade at a single point in time, or simply, a school's status score for a particular grade-level. Useful contrasts and interpretations arise when this cell is grouped with other cells in the table. Different groupings or "slices" of the table support different interpretations about student performance, as we review below.

Table 1.1
Example of a School Status Score

	Year					
Grade	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
3	320	380	350	400	390	420
4	400	450	420	450	480	500
5	510	550	600	650	620	620
6	610	620	630	620	650	660
7	710	780	750	750	800	800
8	810	810	820	820	810	840

3.1 The Vertical Slice: Across-Grade Status

A vertical slice through the data table as shown in Table 1.2 provides a representation of **school status across grades**. Instead of a single shaded cell that summarizes achievement at a single grade, this full shaded column summarizes 2007 school achievement across grades. Descriptions of status are useful, but they represent a single point in time and do not allow for growth interpretations. Although it may seem that differences across grades — from 320 to 400 to 510 and so on — imply growth, these are not the same students across grades, and all scores occurred at the same point in time. The differences in these average scores are best interpreted as differences in achievement across grades at a particular point in time.[1]

Table 1.2
Example of School Status Scores across Grade Levels

	Year					
Grade	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
3	320	380	350	400	390	420
4	400	450	420	450	480	500
5	510	550	600	650	620	620
6	610	620	630	620	650	660
7	710	780	750	750	800	800
8	810	810	820	820	810	840

3.2 The Horizontal Slice: Improvement over Time

Table 1.3 highlights a horizontal slice through the data table to provide a representation of **within-grade improvement over time**. The shaded row in Table 1.3 describes 3rd grade scores from 2007 to 2012. Such horizontal slices are sometimes described as an *improvement model* or a *cohort-to-cohort perspective*. Each cell in the row represents a different cohort of students. Comparison of the cells, from 320 to 380 to 350 and so on, reveals change in achievement at a particular grade level over time. These comparisons are commonplace in large-scale assessment, from the National Assessment of Educational Progress (NAEP) that reports on the achievement of 4th, 8th, and 12th graders over time to state assessment programs that track achievement within grades over time.

[1] Instead of average scores, the cells could contain the more common summary statistic of the percentage of students who are proficient, that is, the number of proficient students divided by the total number of students in each grade times 100. An upcoming section reviews scales for reporting scores in greater detail, but interpreting differences in proficiency percentages across grades is rarely defensible, let alone interpreting these differences as growth. Not only are the students different in each grade-level, but there are likely to be arbitrary differences in proficiency cut scores across grades.

Table 1.3

Example of Within-Grade Improvement over Time

	Year					
Grade	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
3	320	380	350	400	390	420
4	400	450	420	450	480	500
5	510	550	600	650	620	620
6	610	620	630	620	650	660
7	710	780	750	750	800	800
8	810	810	820	820	810	840

A limitation of these within-grade comparisons, or the "improvement model," is that the students comprising the group do not stay the same from one year to the next. Thus, any observed changes in performance may be due to the changing composition of the group. This slice does describe a grade's performance over time and represents growth in a general sense. However, for the purposes of this guide, growth describes a particular student or group whose identity remains constant. In short, for growth, time varies, but the student or group does not. Because within-grade comparisons do not describe the same individuals or a group comprised of the same individuals, this guide does not refer to them as indicating or measuring growth.

3.3 The Diagonal Slice: Growth over Time

A diagonal slice through the data table as shown in Table 1.4 provides a representation of **growth over time**. The shaded cells represent the progression of a particular group of students over time and correspond with an intuitive definition of growth. The highlighted diagonal in the table below represents average scores from a single group of students from 3rd grade in 2007 to 8th grade in 2012.

In the case of Table 1.4, the diagonal represents averages from an unchanging cohort of students; it uses matched student data for students who have scores at all time points. Alternatively, these averages could include data from "mobile" students, who enter the cohort for some, but not all, years, and students whose data may be missing at one or more time points. This contrast is sometimes described as the longitudinal perspective (use data for only students with all matched scores over time) versus the cross-sectional perspective (use data for all students even those with missing values) on growth.

Table 1.4
Example of Growth

	Year					
Grade	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
3	320	380	350	400	390	420
4	400	450	420	450	480	500
5	510	550	600	650	620	620
6	610	620	630	620	650	660
7	710	780	750	750	800	800
8	810	810	820	820	810	840

Growth models often use complete data and either ignore incomplete data or make implicit or explicit assumptions about the missing data. An extensive review of missing data approaches is beyond the scope of this guide, but we include brief descriptions of the handling of missing data when models have particularly straightforward approaches. The remainder of this guide introduces different approaches to interpreting student data within two or more cells of diagonal slices that represent individual or aggregate growth over time.

4 - What is a Growth Model?

If growth describes the academic performance of a student or group over two or more time points, then what is a growth model? A growth model, like a region of a country in a guidebook, is best thought of as an entity with many components and features. A growth model can use a statistical model, but a growth model is not solely a statistical model. Moreover, some growth models are so statistically straightforward that they are best described as a collection of calculations and decision rules, rather than as a formal statistical model. This guide uses the following definition of a growth model.

> A **growth model** is a collection of definitions, calculations, or rules that summarizes student performance over two or more time points and supports interpretations about students, their classrooms, their educators, or their schools.

This definition is broad and likely to be counterintuitive to at least two audiences. First, to those with statistical training, modeling growth usually involves the estimation of a function that describes and predicts individual growth trajectories. Unfortunately, such a restrictive definition excludes many of the growth models in current practice and, more importantly, dramatically understates the scope of their complexity and ambition in educational

accountability contexts. Second, to practitioners with limited exposure to these models, a growth model may seem like a concise, perhaps even single-step procedure capable of achieving many desired goals and outcomes. Such a definition overlooks the multiple components of operational growth models and the complexity and judgment that are required as they increasingly attempt to serve multiple purposes.

Through the systematic characterization of growth models that follows, this guide provides an expansive perspective on the growth model landscape. However, this perspective is not intended as an exhaustive or "correct" way to classify and assess growth models. Growth models are quickly changing to meet the needs of local, state, and federal goals, reforms, and policies, and this guidebook, like real guidebooks, may require frequent revisions. However, the need for conscientious consideration of purpose, terminology, and defensible interpretations is relevant regardless of the growth model or the driving educational policy of the moment.

5 - Growth Models of Interest

The main chapters of this guide review seven individual growth models in turn. The ordering of the chapters is primarily pedagogical, beginning with more simple models and proceeding to more complex models. We attempted to select the most widely used growth models and label them by their most common names. However, some models (i.e., the residual gain model) are less commonly used but serve as a conceptual "missing link" between contrasting statistical foundations. A list of equivalent or closely related models is provided in each chapter. There is also an appendix relating these models to those associated with Council of Chief State School Officers (CCSSO) publications about growth models. The seven growth models of interest in this report follow:

- Gain Score
- Trajectory
- Categorical
- Residual Gain
- Projection
- Student Growth Percentile
- Multivariate

6 - Critical Questions for Describing Growth Models

In a guidebook to a foreign country, each region is described systematically through a series of questions or perspectives: Where are the best places to eat? What hotels offer the best value? Where are the best places to visit? This guide takes a similar approach by explaining each model through a series of critical questions:

1. What *Primary Interpretation* does the Growth Model Best Support?
2. What is the *Statistical Foundation* Underlying the Growth Model?
3. What are the *Required Data Features* for this Growth Model?
4. What Kinds of *Group-Level Interpretations* can this Growth Model Support?
5. How Does the Growth Model *Set Standards* for Expected or Adequate Growth?
6. What are the *Common Misinterpretations* of this Growth Model and Possible *Unintended Consequences* of its Use in Accountability Systems?

Before describing the growth models themselves, Sections 6.1 through 6.6 of Part I discuss these critical questions. Part II of the guide, Chapters 1 to 7, presents the seven growth models by answering these six critical questions for each of them.

6.1 Question 1: What *Primary Interpretation* does the Growth Model Best Support?

One of the central tenets of modern validity theory is that the target of validation is not a model but a use or interpretation of model results. A model suited for one interpretation may not be well suited to support an alternative interpretation. Thus, a natural starting point for growth model classification is the identification of the interpretations that particular growth models best support.

Growth models summarize — typically by quantifying — student performance over two or more time points. They result in metrics that describe individuals and/or groups. This guide identifies three fundamental interpretations that growth metrics can support:

1. **Growth Description**: How much growth? A growth metric may support inferences about the absolute or relative magnitude of growth for an individual or group.
2. **Growth Prediction**: Growth to where? A growth metric may support inferences about the future status of a student or group given current and past achievement.
3. **Value-added**: What caused growth? A growth metric may support inferences about the causes of growth by associating growth with particular educators (e.g., teachers or principals) and schools.

This guide classifies each growth model by the primary interpretation that the growth model supports best. Two caveats are essential here. First, a growth model may support a secondary or tertiary interpretation as well, and these are identified in the respective growth model chapters. Following the definition of a growth model as a collection of definitions, calculations, and rules, it is not surprising that some growth models have been extended to support multiple interpretations. Nonetheless, it is possible to identify a primary interpretation that the growth model supports most naturally.

Second, although a growth model may support a particular primary interpretation, it may not do so infallibly. A growth model whose primary interpretation is growth description may not describe growth in a manner that all users might find most useful. A growth model that primarily supports value-added interpretations may not in fact isolate the average value that a particular teacher or school adds to students. This is discussed further under Question 6 that concerns common misinterpretations of models and threats to their use in accountability systems.

An alternative approach to classifying models is by the more general purposes that the model might serve. Such general purposes include using growth models to inform classroom instruction, student learning, school accountability decisions, evaluations of educators, and evaluations of particular programs and interventions. These purposes are important but are farther removed from growth model output and therefore result in a less straightforward classification scheme. Clearer distinctions between models arise by focusing on the interpretations that growth model metrics support directly.

Table 1.5 provides examples of growth models classified column-wise by their primary interpretations. The models are also classified row-wise by their statistical foundations, which are presented in the next section. A brief description of each model is also included. When different facets of a model support different interpretations, the models are classified in more than one column.

Table 1.5

Classification Scheme for Growth Models

Primary Interpretation			
Statistical Foundation	**Growth Description**	**Growth Prediction**	**Value-Added**
Gain-Based Model Chapters 1-3: Based on score gains and trajectories on a vertical scale over time	• **Gain-Score** Chapter 1: Gains, average gains, slopes • **Categorical** Chapter 3: Changes and transitions between categories	• **Trajectory** Chapter 2: Extrapolation of gains into the future • **Categorical (a.k.a. Transition, Value Table)** Chapter 3: Implicit momentum toward higher categories in the future	• **Gains/Slopes as Outcomes** Chapter 1.4: Establishes links between average gains and classroom/school membership
Conditional Status Model Chapters 4-6: Expresses scores in terms of expectations based on past scores	• **Residual Gain** Chapter 4: Simple difference between status and expected status given past scores • **Student Growth Percentile (a.k.a the Colorado Model)** Chapter 6: Percentile rank of status given past scores	• **Projection (a.k.a. Prediction, Regression)** Chapter 5: Empirically predicted future score given past scores • **Student Growth Percentile (a.k.a. the Colorado Model)** Chapter 6: Continuation of current percentile rank into the future	• **Covariate-Adjustment** Chapter 4.4: Establishes links between average conditional status and classroom/school membership
Multivariate Model Chapter 7: Uses entire student score histories as an outcome to associate higher-than-expected scores with particular educators	• Generally not used for this purpose	• Generally not used for this purpose	• **Multivariate (a.k.a. EVAAS, Cross-Classified, Persistence Models)** Chapter 7

6.2 Question 2: What is the *Statistical Foundation* Underlying the Growth Model?

This guide also classifies growth models by their underlying statistical foundation. Although statistical methods can be intimidating and model descriptions can be opaque, we find that models can be classified into one of three categories: gain-based models, conditional status models, and multivariate models. These three categories make up the rows of Table 1.5, which cross-classifies growth models by Questions 1 and 2. This table represents a central conceptual framework for this guide. The following subsections briefly describe each statistical foundation in more detail and reference some of their corresponding models.

6.2.1 Gain-based models

The first type of statistical foundation underlies models that are based on gains, average gains, or score trajectories over time. We call these *gain-based models*. A gain or gain score is the simple difference between two scores at different points in time. The gain score can be extrapolated over future time points to support predictions. When there are more than two data points for an individual, the gain can be generalized over multiple time points by averaging and expressing progress as an average change per unit of time.

A common feature to all gain-based models is an implicit or explicit recognition of a *vertical scale*, a common scale that allows scores to be compared across different grade-level tests. Vertical scales support interpretable score differences over the time and grade range of interest. A gain-based statistical foundation is consistent with an intuitive definition of growth: the difference between where one was and where one is. However, vertical scales are difficult to design and maintain, and many useful questions about performance over time do not require vertical scales. This motivates a contrasting statistical foundation underlying a second class of growth models.

6.2.2 Conditional status models

The second type of statistical foundation supports interpretations about *conditional status*. The word "conditional" implies an "if" statement, a kind of dependence, and, indeed, conditional status recasts or reframes status with respect to additional information. Models that use this statistical foundation address the question: How well does a student perform with respect to *expectations*? These expectations are set empirically using the past scores of the student of interest and other students.

Using this past information, conditional status models use a two-step process. First, given a student's past scores, they establish expectations about his or her current score. Second, the student's actual status is compared to these "conditional" expectations given past scores. The use and differentiation of past and current scores allows this method to meet our definition of a growth model. The phrase, "conditional status," is a technical term arising from the models' referencing of student status in terms that are conditional upon past scores or, more simply, in terms that consider past scores or take past scores into account. This foundation is fundamentally distinct from models that have a gain-based foundation, where status is evaluated over time instead of compared to expectations based on past scores.

Notably, conditional status models can reference current status to other variables in addition to or in place of past scores, such as economic status, race and ethnicity, or participation in specific educational programs. It is entirely possible to use a conditional status model to describe status in terms of expectations set by less relevant variables like a student's height

or shoe size. This observation does not invalidate conditional status models as growth models but serves to emphasize how this statistical foundation supports a fundamentally different conception of growth: status with respect to expectations based on past scores and, potentially, other information.

A natural corollary of this definition of growth is that conditional status will change as expectations change. Setting expectations based upon two past scores will result in a different conditional status than setting expectations based on three past scores, and setting expectations based upon student demographic variables will change a student's conditional status score even further. In comparison, gain-based scores will also change under inclusion of additional time points. However, increasing previous time-points for *gain-based models* allows for better estimation of average gains, whereas using more past scores in *conditional status models* changes the substantive interpretation of the conditional status score. In sum, the output of conditional status models is interpreted most accurately with full appreciation of the variables that have been used to set expectations.

Conditional status scores can be reported on many metrics, from the test score scale to percentile ranks. As an example, consider a student whose high current status places her at the 80th percentile (among all students). In spite of this relatively high score, this student's past scores have been at even higher percentiles. Thus, her current percentile rank of 80 is somewhat below the empirically derived expectations given these past scores. One expression of conditional status is the simple difference between her actual current score and the score that is expected given her past scores. This describes the residual gain model in Chapter 4. Another approach expresses this low expectation in terms of a percentile rank. This latter approach is known as a Student Growth Percentile and is described in detail in Chapter 6. Table 1.5 displays conditional status models in its second row, cross-classified by the primary interpretations that these models support.

Chapters 4-6 review conditional status models and delve more deeply into the contrasts between gain-based and conditional status models. Understanding these contrasts is essential for accurate selection and use of growth models.

6.2.3 Multivariate models

The third type of statistical foundation is used primarily to estimate the "value-added" associated with classrooms and schools. Table 1.5 displays multivariate models in its third row and includes no models in the first two columns, as this statistical foundation is not well suited for growth description or growth prediction.

Multivariate models are distinguished by their complexity and their ability to use a large amount of data and variables in a unified approach. They require specialized and sometimes proprietary software and training in the interpretation of model output. The models are designed to

produce classroom- and school-level "effects" that may be associated with teachers and principals respectively. Formally, gain-based and conditional status models can be seen as special cases of a flexible multivariate model (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). However, in practice, it is useful to locate multivariate models as a separate statistical foundation. As Chapter 7 describes, the multivariate model has as its target of inference, not a student's gain or her conditional status, but her entire score history. This is clumsy for growth description and growth prediction, but is particularly well-suited to leverage maximal test score information for the estimation of classroom and school effects.

6.3 Question 3: What are the *Required Data Features* for this Growth Model?

The selection of a growth model can be motivated by both the advantages it offers and the constraints it satisfies. The selection of a desired model may necessitate alternative or additional data structures. In some cases, the cost of meeting data requirements may outweigh the benefits of the desired model.

In general, all growth models rely on the usual expectations for test reliability and validation. These are not trivial requirements, but this section focuses on requirements for growth, above and beyond the requirements for test score interpretations at a single time point. If low reliability threatens interpretations of test scores at a single time point, the problems will only compound as these scores are reconfigured to support growth inferences. Similarly, all the growth models in this guide require student data that is linked longitudinally over at least two time points.

This section reviews particular data requirements for the growth models considered in this paper, including vertical scales, proficiency cut scores articulated across grades, multiple cut scores articulated across grades, large student datasets, multiple prior years of data, and meaningful controls and covariates. Some requirements are more salient for some models than others. It is useful to note, however, that in many cases, the integrity of the interpretations from a growth model depends on the integrity of these data requirements. This is especially important to consider when the growth model requires cut scores or vertical scales as standard setting and even scaling, albeit to a lesser degree, involve judgmental decisions. The statistical model or calculations of a growth model do not compensate for poorly defined vertical scales or performance level categories. The principal data requirements for each model are reviewed in the model's respective chapter.

6.3.1 Vertical scales

Some assessments are scaled across grades with what is known as a "vertical scale." A vertical scale links the reporting test score scale across several grade levels so that a test score from one grade can be meaningfully compared to a test score in a subsequent or previous grade level. This type of scale contrasts with "horizontal" test score scales that support interpretations

for each grade level separately. Vertical scales are often more desirable than horizontal scales due to the growth interpretations they support, but vertical scales require more rigorous design specifications in test development to ensure a meaningful across-grade content continuum. Moreover, in many cases, vertical scales are not possible for the subject matter tested. For example, science classes may cover distinct topics in each grade and may not support an interpretable cross-grade continuum of "science" knowledge.

Vertical scales are necessary for gain-based models and are implicit in intuitive notions of growth. If a test has a defensible vertical scale, a user can take a simple difference of individual scores over time and interpret this as a gain regardless of the starting point on the continuum. In some cases, vertical scales are not formally supported but are implicit and loosely operationalized. An example of this is the categorical model where no vertical scale is claimed, but transitions across performance category boundaries are treated as gains, an interpretation that requires meaningful linkages in cut scores defining the performance categories across grades.

6.3.2 Proficiency cut scores articulated across grades

Some growth models afford growth predictions, often with inferences about trajectories toward some future standard such as "Proficiency" or "College and Career Readiness." These models proliferated under the Growth Model Pilot Program of 2005 (U.S. Department of Education, 2005) that required students to be "on track" to proficiency. Most growth models do not require a proficiency cut score to make a prediction, but the prediction is ultimately referenced to the cut score. In these cases, model predictions require *articulated* cut scores across grades, in other words, proficiency cut scores that maintain some consistent relative stringency or pattern of stringency across grades.

Such cut scores are determined through standard setting procedures in which a committee first defines what proficient students should know and be able to do and then sets cuts by taking into account characteristics of the test scale, item content and difficulty levels, and the qualitative description of proficiency. For many growth models, this process requires consideration of the definitions of proficiency in all other grade levels. Without articulated cut scores, nonsensical conclusions can arise, including a student who is on track to some future standard in one year and three years, but not in two years (Ho, Lewis, & Farris, 2009). Lack of articulation leads to unpredictable relationships between stringency of standards and the grade of entry, the time horizon to proficiency, and target year by which standards must be reached, respectively.

6.3.3 Multiple cut scores articulated across grades

Many accountability and evaluation policies focus primarily on students reaching a single achievement level, usually designated as "Proficiency." Some policies also operationalize performance levels that support finer grain distinctions at higher and lower score points.

Performance level descriptors may include Below Basic, Basic, Proficient, and Advanced, and some states include an even finer resolution of categories below proficiency. Standard-setting processes help to set these cut scores and elaborate on the descriptions for each category.

Categorical models, sometimes known as transition matrix models or value tables, use such ordered performance level categories to determine whether students are making adequate gains toward a standard. Such models rely heavily on the assumption that the performance level categories have been articulated within and across grades. Moreover, the same performance level category in different grades should reflect the same *relative* degree of mastery. As an extension of the previous argument for proficiency cut scores, any growth model that uses multiple cut scores to document growth must have well-articulated standards across grades to avoid counterintuitive results.

6.3.4 Large numbers of students

Some growth models require large numbers of students to produce reliable estimates. This is particularly essential for growth models that require estimation of several parameters, such as the Student Growth Percentile (SGP) model. The SGP model involves estimation of hundreds of parameters and thus requires large numbers of students to ensure that SGPs support appropriate interpretations. A rough, general guideline for a minimum sample size for SGP estimation is 5000 (Castellano & Ho, in press), but the requirement depends on the inferences that the model supports. Although 5000 is a comfortable size for many state-level datasets, some states may find instability if SGPs are calculated for particular districts, grades, or subgroups.

6.3.5 Multiple years

For growth models to support value-added inferences, they often need to accommodate several years of test score data for the same educator, ideally with large numbers of students for that educator. At the same time, students within each classroom require scores from many prior years. As the stakes associated with the use of the growth model results become higher, more data will be required to increase the precision of estimates.

6.3.6 Meaningful controls/covariates

Models that set empirical expectations based on selected variables, including all conditional status and value-added models, are interpreted most accurately when there is full awareness of the set of variables that have been used to set these expectations. In the case of value-added inferences, accurate interpretation requires an understanding of how many previous scores have been included and which additional student-, teacher-, and school-level variables have been incorporated, if any. Possible variables include the percentage of students from low-income families, the minority/ethnic composition of the school/classroom, and the percentage

of limited English proficiency students. Incorporating these factors can bolster the argument for the interpretation of teacher and school effects as "value added," but the primary goal should be adequate communication of the variables to understand the effects. By understanding that value added is more accurately interpreted as an average student status beyond expectations, the importance of understanding the variables that set these expectations becomes apparent.

6.4 Question 4: What Kinds of *Group-Level Interpretations* can this Growth Model Support?

Growth models use student-level performance data from two or more time points. Accordingly, a growth model can provide a number that characterizes a student's growth. However, practitioners are often more interested in *group-level* summaries of academic growth, especially in the context of accountability and evaluation. In most cases, group-level summaries are easily obtained by averaging student-level growth values for the students in a group of interest, such as averaging over the students in a classroom or school. In other cases, such as the case of the multivariate model, group membership is explicitly included in the model.

As policy, accountability, and evaluation decisions (such as for teacher effectiveness and school accountability) are so often associated with the group-level summaries, the validity of group-level interpretations is of paramount importance. Evidence supporting student-level growth interpretations is important, but this evidence does not ensure that an aggregate of a student-level metric can also be used for high-stakes purposes. In answering Question 4 for each model, this guide discusses the group-level interpretations that each model can support and describes the evidence needed for these interpretations.

6.5 Question 5: How Does the Growth Model *Set Standards* for Expected or Adequate Growth?

A growth model can be used to set standards for expected or adequate growth in different ways. All conditional status and value-added models set statistical standards for expected scores. However, these expectations may not be aligned with substantive and policy guidelines for adequate growth. In some cases, the choice of standard for growth performance can be based on norms or performance by a clearly defined group of peers. This can lead to judgmental decisions based on percentages, such as flagging the top or bottom 10 percent of students, teachers, or schools for further investigation.

Any standard-setting process involves subjective judgments. The necessity of these judgments to the use of operational growth models is a reminder that operational growth models are more than statistical models. Judgments are moderated by the stakes involved, the properties of the model itself, student performance and impact data, and the theory of action for the policy of interest. In each chapter, we review standard setting conventions in theory and practice.

Options include setting standards based on the test score scale for growth, standards based on a norm-referenced percentage, or standards based on an aggregate-level metric for group growth. All of these procedures support inferences about low, high, and adequate growth.

6.6 Question 6: What are the *Common Misinterpretations* of this Growth Model and Possible *Unintended Consequences* of its Use in Accountability Systems?

When visiting a new region, tourists frequently begin with preconceived notions of what they will encounter. These assumptions might be based on something they have heard, read, or experienced. A useful guidebook is one that understands common misconceptions and addresses them directly. As growth models are incorporated into educational policies, some impressions of models do not align with actual model function, and some common interpretations of model output may not be defensible. In answering Question 6, this guide clarifies common misconceptions of particular growth models that threaten the validity of the inferences derived from their use.

It is also well established that the validation of an evaluation system becomes difficult as the stakes of the evaluation rise. A metric that is initially designed for informing instructional decisions may be susceptible to corruption, inflation, and gaming when it is incorporated into a high-stakes system. A responsible guide is one that anticipates both positive and negative responses to growth models. In answering Question 6, this guide also explores how growth metrics can be gamed or distorted upon their adoption into a high-stakes accountability system.

This guide is about growth models, including, but not limited to, value-added models for school and teacher accountability. A full review of the issues surrounding the use of growth models for high-stakes accountability systems is not feasible here. Question 6 is an opportunity to identify some of the most obvious concerns that arise in common growth models. For a fuller discussion of teacher value-added models, we point to a number of other references that focus on this topic more specifically.[2] We comment on this issue only briefly here and in subsequent chapters.

Our first critical question makes it clear that we consider value added to be an inference, not a model. In the absence of a rigorous design where, among other requirements,[3] students are randomly assigned to classrooms, no model can support value added inferences on its own. The term is best considered to be a hypothesis that must be tested through the triangulation of multiple sources of evidence. Nonetheless, many models are used to support value-added inferences, and it is on this basis that we classify them, describe them, and, in this critical question, identify their strengths and weaknesses.

---

[2] See Reardon & Raudenbush (2009); Baker, Barton, Darling-Hammond, Haertel, Ladd, Linn, Ravitch, Rothstein, Shavelson, & Shepard (2010); and Glazerman, Loeb, Goldhaber, Staiger, Raudenbush, & Whitehurst (2010).

[3] See Reardon & Raudenbush (2009) and Rubin, Stuart, & Zanutto (2004).

7 - Alternative Growth Model Classification Schemes

This guide differs from many previous efforts at growth model classification. It is not intended as an authoritative classification scheme. It is instead, as its title suggests, a guide for practitioners, and it should not only aid understanding of growth models, but increase appreciation for alternative classification schemes. These alternatives are many, and we list them briefly in this section.

Some classification schemes are more concise than the one presented here. An example of this is CCSSO's *Understanding and Using Achievement Growth Data* brochure (Council of Chief State School Officers, 2011). Others are listed later in this section. These schemes tend to collapse categories across the critical questions we identify here, resulting in a simpler, one-dimensional summary. Table A.1 in the appendix maps the classification scheme from CCSSO's brochure onto the classification scheme of this guide.

Other classification schemes are focused on a particular critical question that we raise in this guide. For example, the CCSSO *Growth Model Comparison Study* (Goldschmidt, Choi, & Beaudoin, 2012) is an effort at comparing the empirical results of a number of different growth models, assuming that all models were reconfigured toward the goal of school "value-added"-type rankings. Table A.2 in the appendix also includes a mapping of that classification scheme onto that of this guide.

Still other classification schemes are more technical, including those comparing value-added models for teacher accountability (McCaffrey et al., 2004), and more specific in their primary interpretations, such as the final evaluation of the Growth Model Pilot Program that compared growth models for growth prediction (Hoffer, Hedberg, Brown, Halverson, Reid-Brossard, Ho, & Furgol, 2011). In contrast, this guide includes few empirical results. It represents a broader view of the growth model landscape and highlights the similarities and differences that might be most useful to practitioners.

This introductory chapter concludes with a list of comparative studies of growth models and alternative growth model classification schemes. Following this, a summary table reviews the question-by-model organization of this guide and briefly summarizes the answers to these questions. The remaining seven chapters of this guide in Part II review each of the seven growth models of interest.

References (Part I)

Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R., Ravitch, D., Rothstein, R., Shavelson, R.J., Shepard, L.A. (2010, August). *Problems with the use of student test scores to evaluate teachers* (Economic Policy Institute Briefing Paper #278).

Council of Chief State School Officers. (2011). Understanding and using achievement growth data. *Growth Model Brochure Series*. Retrieved September 19, 2012, from http://www.wera-web.org/links/Journal/June_Journal_2012/CC6_CCSSO_Growth_Brochures_jan2012.pdf.

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*.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., and Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: Brown Center on Education Policy at Brookings, from http://www.brookings.edu/research/reports/2010/11/17-evaluating-teachers.

Goldschmidt, P., Choi, K., and Beaudoin, J.P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance*. Washington, DC: Council of Chief State School Officers.

Ho, A.D., Lewis, D.M., and Farris, J.L.M. (2009). The dependence of growth-model results on proficiency cut scores. *Educational Measurement: Issues and Practice, 28*(5): 15-26.

Hoffer, T.B., Hedberg, E.C., Brown, K.L., Halverson, M.L., Reid-Brossard, P., Ho, A.D., and Furgol, K. (2011). *Final report on the evaluation of the growth model pilot project*. Washington, DC: U.S. Department of Education, from http://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/index.html.

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A., and Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1): 67-101.

Reardon, S.F., and Raudenbush, S.W. (2009). Assumptions of value added models for estimating school effects. *Educational Finance and Policy, 4*(4): 492-519.

Rubin, D.B., Stuart, E.A., and Zanutto, E.L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics,* 29(1): 103–116.

U.S. Department of Education. (2005). *Secretary Spellings announces growth model pilot, addresses chief state school officers' Annual Policy Forum in Richmond* (Press Release dated November 18, 2005), from http://www2.ed.gov/news/pressreleases/2005/11/11182005.html.

List of Comparative Studies of Growth Models

Beimers, J. (2008). *The effects of model choice and subgroup on decisions in accountability systems based on student growth.* Ph.D. dissertation, University of Iowa.

Buzick, H.M., and Laitusis, C.C. (2010). *A summary of models and standards-based applications for grade-to-grade growth on statewide assessments and implications for students with disabilities* (Educational Testing Service TS RR-10-14). Princeton, NJ: ETS. Retrieved March 29, 2012, from http://www.ets.org/Media/Research/pdf/RR-10-14.pdf.

Dunn, J.L., and Allen, J. (2009). Holding schools accountable for the growth of nonproficient students: Coordinating measurement and accountability. *Educational Measurement: Issues and Practice, 28*(4), 27-41.

Goldschmidt, P., Choi, K., and Beaudoin, J.P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance*. Washington, DC: Council of Chief State School Officers.

Hoffer, T.B., Hedberg, E.C., Brown, K.L., Halverson, M.L., Reid-Brossard, P., Ho, A.D., and Furgol, K. (2011). *Final report on the evaluation of the growth model pilot project*. Washington, DC: U.S. Department of Education. Retrieved April 27, 2012, from http://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/index.html.

Sanders, W.L., and Horn, S.P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*(3): 299-311.

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

Weiss, M. (2008). *Using a yardstick to measure a meter: Growth, projection, and value-added models in the context of school accountability.* Ph.D. dissertation, University of Pennsylvania.

List of Reports Overviewing and Classifying Growth Models

Auty, W., Bielawski, P., Deeter, T., Hirata, G., Hovanetz-Lassila, C., Rheim, J., Goldschmidt, P., O'Malley, K., Blank, R., and Williams, A. (2008). *Implementer's guide to growth models.* Washington, DC: Council of Chief State School Officers.

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A., and Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1): 67-101.

O'Malley, K.J., Murphy, S., McClarty, K.L., Murphy, D., and McBride, Y. (2011). *Overview of student growth models* (Pearson White Paper). Retrieved March 29, 2012, from http://www.pearsonassessments.com/hai/Images/tmrs/Student_Growth_WP_083111_FINAL.pdf.

Summary Table

Model / Characteristics	Gain Score	Trajectory	Categorical	Residual Gain	Projection	Student Growth Percentile	Multivariate
Brief Description	Describes growth with simple differences or average gains over time	Extends gains or average gains in a predictable, usually linear fashion into the future	Defines growth by transitions among status categories (e.g., Basic, Proficient, Advanced) over time	Describes growth as the difference between current status and expected status given past scores	Uses past scores to predict future scores through regression equations	Percentile rank of current status in a reference group of students with similar past scores	Uses entire student score histories, including other subjects and teachers, to detect higher than expected student scores associated with particular teachers
Aliases, Variants, Close Extensions	Growth Relative to Self, Raw Gain, Simple Gain, Slope, Average Gain, Gains/Slopes-as-Outcomes, Trajectory Model	Growth-to-Standards Model, Gain-Score Model	Transition Model, Transition Matrix Model, Value Table	Residual Difference Model, Covariate Adjustment Model, Regression Model, Percentile Rank of Residuals	Regression Model, Prediction Model	The Colorado Model, Percentile Growth Trajectories, Conditional Status Percentile Ranks	Sanders Model, EVAAS, TVAAS, Tennessee Model, Layered Model, Variable Persistence Model, Cross-Classified Model
Primary Question(s) Addressed	How much has a student learned on an absolute scale?	If this student continues on this trajectory, where is she likely to be in the future?	How has this student grown in terms of transitions through categories over time? In which category will she likely be in the future?	How much higher or lower has this student scored than expected given her past scores?	Given this student's past scores, and based on patterns of scores in the past, what is her predicted score in the future?	What is the percentile rank of a student compared to students with similar score histories? What is the minimum SGP a student must maintain to reach a target future standard?	Is this teacher associated with higher scores for his or her students than expected given all available scores and other teacher effects?
Q1: Primary Interpretation	Growth Description	Growth Prediction	Growth Description and Growth Prediction	Growth Description	Growth Prediction	Growth Description and Growth Prediction	Value Added
Q2: Statistical Foundation	Gain-Based	Gain-Based	Gain-Based	Conditional Status	Conditional Status	Conditional Status	Multivariate
Q3: Required Data Features	Vertical scale	Vertical scale	Articulated cut scores across years and grades. Values for value tables. Implicit vertical scale.	An interpretable scale. Assumptions of linear regression must be met.	Interpretable future scale or future standard.	Large sample sizes for reliable estimation.	For high-stakes value-added uses, many years of student data required for stable teacher effects.
Q4: Group-Level Interpretations	Average gain	Average trajectory or percentage of on-track students	Average across value tables or percentage of on-track students	Average residual gain	Average future prediction or percentage of on-track students	Median or average SGP, percentage of on-track students	Only group-level interpretations: Teacher- and school-level "effects"
Q5: Setting Standards	Requires judgment about adequate gain or adequate average gain. Requires understanding of the scale or can be norm-referenced.	Set by defining a future standard and a time horizon to meet the standard.	Set by defining cut scores for categories and values in value table. Requires judgmental cut scores to define adequacy of both individual and aggregate values.	Requires judgment about adequate residual gain. Requires understanding of the scale or can be norm-referenced.	Set by defining a future standard and a time horizon to meet the standard.	Requires judgment about an adequate SGP or median/average SGP. Predictions require a future standard and a time horizon to meet the standard.	Standards required to support absolute or relative distinctions among teacher/school effects, e.g., awards/sanctions to top/bottom 5%.
Q6: Misinterpretations and Unintended Consequences	Intuitive but dependent on vertical scales that can impart undesired dependencies between growth and initial status or socioeconomic status. Can be inflated by dropping initial scores.	Less of an empirical prediction than an aspirational and descriptive prediction. Requires defensible vertical scale over many years. Can be inflated by dropping initial scores.	Loss of information due to categorization of scores. Requires careful articulation of cut scores across grades and years: assumes an implicit vertical scale. Can be inflated by dropping initial scores.	Not a "gain" but a difference from actual and expected status. Violations of linear regression assumption can lead to distortions. Can be inflated by dropping initial scores.	The "projection" metaphor can be confused with "trajectory" when it is in fact a prediction. Maximizing predictive accuracy can diminish incentives to address low-scoring students.	Sometimes misinterpreted as the percentile rank of gain scores. Sometimes overinterpreted as supporting value-added inferences. Can be inflated by dropping initial scores.	Naming fallacy: calling a metric "value-added" does not make it so. Can be unreliable. Detached from theories about improving teaching. Can be inflated by dropping initial scores.

PART II
THE GROWTH MODELS

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.

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

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.

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

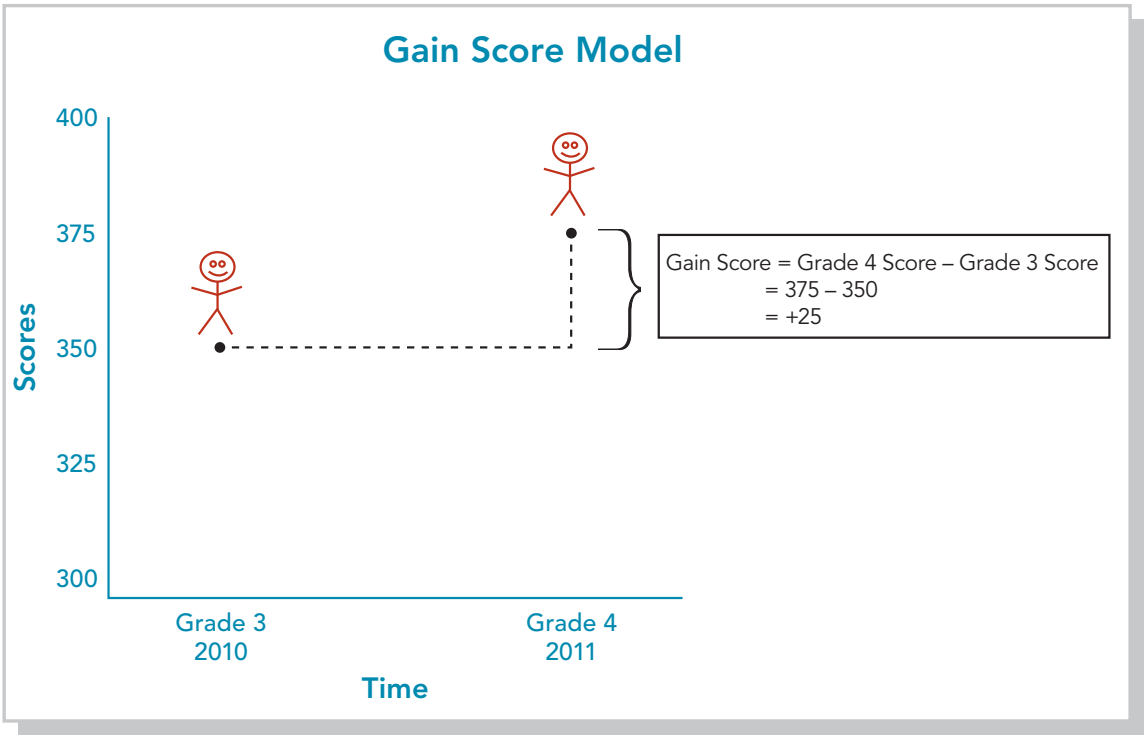


Figure 1.1 illustrates the gain score model calculation using data for a student in Grade 3 in 2010 and in Grade 4 in 2011 on a hypothetical mathematics test. The horizontal axis represents time, and the vertical axis represents the test score scale. For test scores from both the Grade 3 and Grade 4 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.

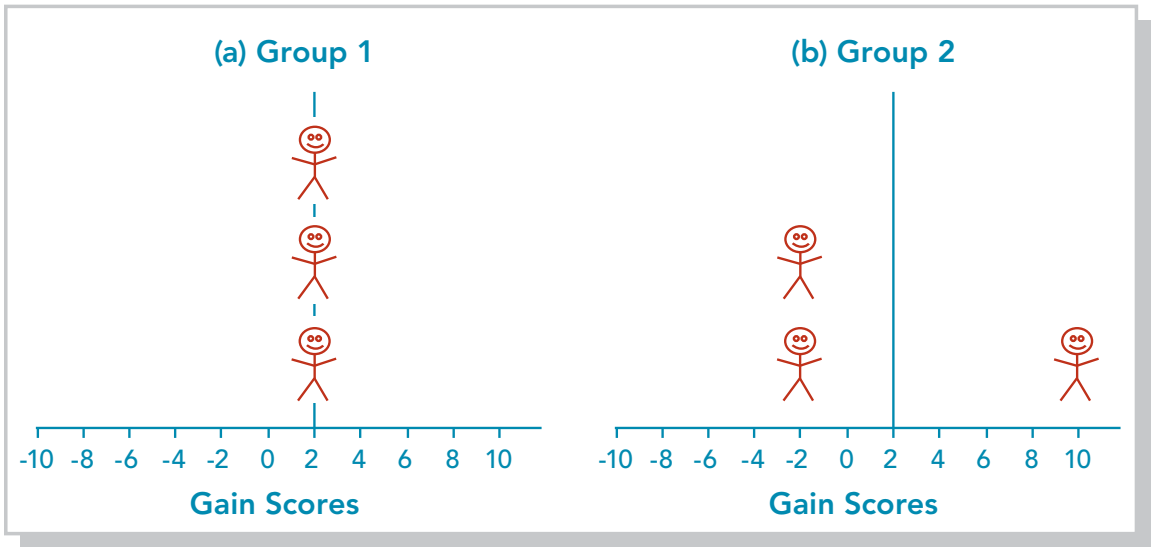
The gain score model supports simple calculations of group-level statistics. Most commonly, the group-level summary statistic for a set of students of interest, such as in a particular classroom, grade level, school, or district, is the average of their individual gain scores. This summary statistic 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 2
The Trajectory Model

The trajectory model is a natural extension of the gain score model. Like the gain score model, the trajectory model is **gain-based**, but instead of describing growth, the trajectory model is used primarily for **growth prediction**. The model uses student gain scores to predict student scores in some future year. The trajectory model, as the name suggests, assumes that a student will continue on his or her same trajectory, which is usually operationalized as an assumption of linear growth. That is, a student makes the same gains each year. For instance, if a student gained 3 points this year, the trajectory model predicts that he or she will gain 3 points in each subsequent year as well. The trajectory model answers the question

If this student continues on her trajectory, where is she likely to be in the future?

An additional and sometimes essential component of models for growth prediction is a determination of whether future predicted performance is satisfactory. Trajectory models can support this determination by providing an "on track" trajectory for each student into his or her future as well as a "predicted" trajectory based on the student's observed gains. The on-track trajectory is formed by determining the annual gains needed to meet a target score in *x* years. A comparison between a student's predicted and on-track trajectory can support a decision about whether a student is making adequate gains toward the future target score. This is discussed further in Sections 2.2 and 2.5.

TRAJECTORY MODEL

Aliases and Variants:
- Growth-to-Standards Model
- Gain-Score Model

Primary Interpretation:
Growth prediction

Statistical Foundation:
Gain-based model

Metric/Scale:
Predicted future score — on the common scale of inputted test scores

Data: Vertically-scaled tests from an initial time point to the final target time point and observed test scores from two time points

Group-Level Statistic:
Average slope, or percentage of students "on track" to reaching proficiency in x years or average trajectory

Set Growth Standards:
Define future standard and the maximum time until standard is reached

Operational Examples:
NCLB Growth Model (e.g., Alaska, Arizona, Arkansas, Florida, and North Carolina)

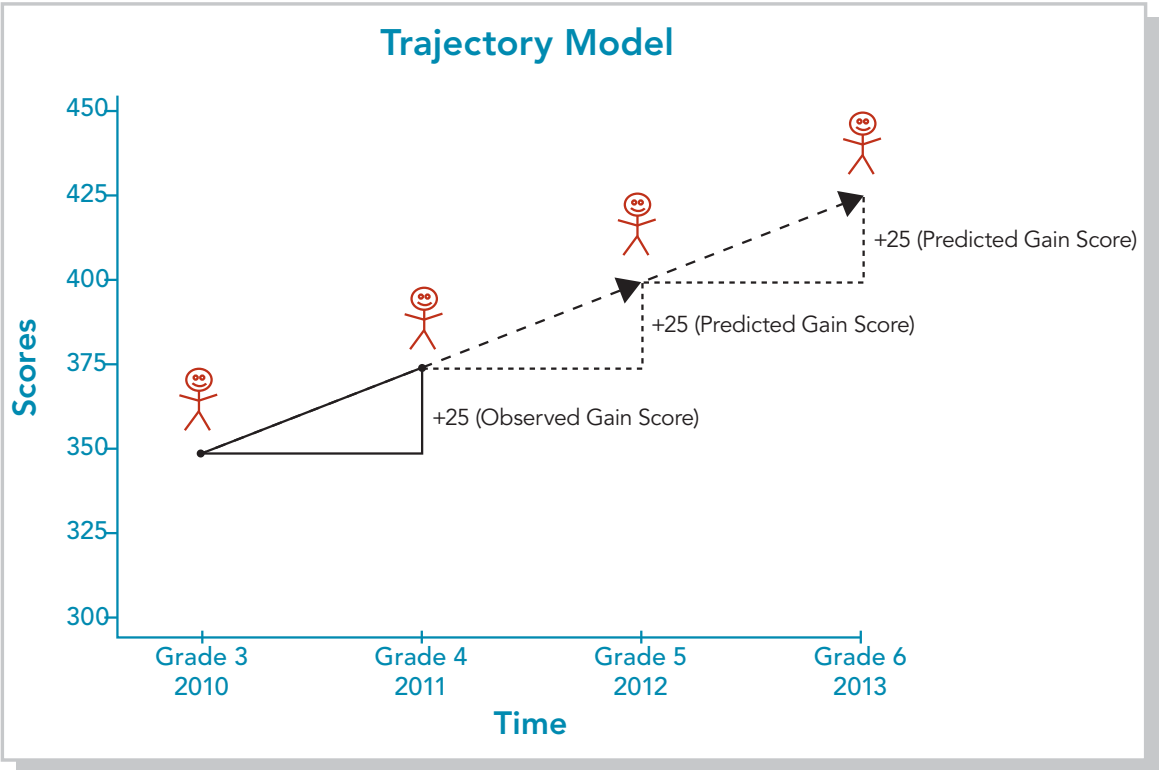
Question 2.1:
What *Primary Interpretation* Does the Trajectory Model Best Support?

By assuming that past gains will continue into the future, trajectory models provide predictions for future scores. They support growth predictions.

In the gain score model discussed in the previous chapter, the gain score — the difference between current and initial status — describes growth as the change from a previous time point to a current time point. The trajectory model uses this gain score as the basis for a growth trajectory extending into the future. Figure 2.1 illustrates this process.

Figure 2.1 uses the same hypothetical student's data from Figure 1.1, where the gain score was illustrated. As shown by the solid, black dots, this student earned a score of 350 in Grade 3 in 2010 and then a score of 375 in Grade 4 in 2011. The vertical distance between these scores corresponds to her gain score: 375 – 350 = + 25 from Grade 3 to Grade 4. For this gain score to be an interpretable quantity, the scores at Grades 3 and 4 must be expressed on a common vertical scale. If this scale also underlies tests at subsequent grade levels, gains through subsequent grade levels will also be interpretable quantities.

Figure 2.1
The Trajectory Model Makes Predictions about Future Student Performance, Assuming that Gains Will Be the Same over Time



From Grade 3 to Grade 4, Figure 2.1 displays the student's actual, or observed, gain. Accordingly, the gain score from Grade 3 to Grade 4 is labeled the "Observed Gain Score." These two points alone comprise the gain score model from the previous chapter. The trajectory mode requires the additional assumption that this student will continue to make

positive gains of 25 points each year. In this way, trajectory models support visualization of the student's achievement trajectory from now into the future, as illustrated in Figure 2.1. This line has a positive slope because the student made positive gains; if she had made negative gains, then the line would have a negative slope. The trajectory could be extending past Grade 6 by continuing in this way — adding 25 points to the student's previous score to obtain a predicted score in the subsequent grade — as long as the grade level assessments are all on the same vertical scale.

The vertical scale suggests that the difference of 25 points each year is comparable over time. This desired property is known as an equal-interval scale property, where differences, or equal intervals, share the same interpretation over the applicable range of the scale. Physical scales for height and weight generally support this property: a gain of 5 pounds is equivalent regardless of whether the individual originally weighed 120 pounds or 220 pounds. However, test score scales generally have weak arguments for equal-interval scale properties. It is difficult to argue that an achievement gain of 5 points in Grade 3 is the same as an achievement gain of 5 points in Grade 8, for example, because the material learned in the two grades can differ substantially. The argument becomes more difficult to support as the scale spans more grade levels. From this perspective, the trajectory is more defensible as a descriptive and aspirational prediction than it is as an empirical prediction.

Figure 2.1 helps to visualize how trajectory models answer the key question they address: If this student continues on her trajectory, where is she likely to be at some point in the future? Trajectory models are appealing because they predict growth along a linear trajectory, which is a straightforward way of extrapolating from an observed linear change. The intuition aligns with that of physical momentum or even Newton's First Law — an object in motion tends to stay in motion.

Question 2.2:
What is the *Statistical Foundation* Underlying the Trajectory Model?

Trajectory models are an extension of the gain score model that extrapolates from student gains to predict future performance. They are gain-based models.

Of the three statistical foundations presented in the introduction (gain-based models, conditional status models, and value-added models), trajectory models have a gain-based statistical foundation. Unlike the gain score model, which typically involves computing a single or average gain score over observed time points, the trajectory model extrapolates from observed gains to future time points.

The extrapolation of gains to support predictions is usually linear, as shown in Figure 2.1. However, in some cases, a nonlinear, curving predicted trajectory is warranted. If scales are

designed to support these nonlinear trajectories, then these nonlinear expectations can be built into the extrapolated trajectory. If, for example, there is a known acceleration in trajectories due to the design of the vertical scale, a gain can be algebraically accelerated in future years to match the assumptions of the vertical scale. In these cases, the statistical foundation is still fundamentally gain-based, as this accelerating factor is applied fundamentally to the observed gain. The key feature of gain-based models is the centrality of the gain to all calculations and inferences.

Another straightforward extension of the trajectory model is an averaging of the gain across multiple observed time points. The previous section noted that the gain-score model is capable of supporting average gains over more than two time points. These average gains can be extended in a linear fashion into the future to support predictions. These average gain or slope-based models use average gains over a given unit of time and extend them in a linear fashion. When the vertical scale supports this averaging of gains, these averages over multiple time points result in more robust estimates of student trajectories than simple gains over only two time points.

A contrasting use of the trajectory model involves "resetting" the trajectory after each year of data collection, using only the two most recent years of data to establish a gain-score and a linear extrapolation. This approach sacrifices robustness in the estimation of a linear trend for simplicity and ease in explanation. If the vertical scale properties do not hold over multiple grades, this approach can theoretically minimize the distortion imparted by poor vertical scaling; but in these cases, the best approach would be to select a model that does not require a vertical scale.

Question 2.3:
What are the *Required Data Features* for the Trajectory Model?

The trajectory model requires student test score data from at least two time points and a common, vertical scale that underlies all observed and predicted test scores from the initial observed score to the future unobserved prediction.

The trajectory model is a gain-based model whose primary supported interpretation is growth prediction. The only student data it requires are student test scores from two time points. The difference between the two test scores is the student's observed gain score, and this gain is extrapolated, usually linearly, into the future. Accordingly, this model requires that test scores from all observed and future time points of interest are linked to a common vertical scale.

Vertical scales facilitate comparison of scores from one year to the next. If the tests from the two time points are on different scales, then their score differences do not meaningfully

relate to changes in performance over time. Using different test scales is analogous to a scenario in which an individual takes the temperature one day in Fahrenheit and the next day in Celsius, and then takes the difference of the two temperatures on different scales. This simple difference is difficult to interpret and cannot indicate whether the temperature has risen or fallen due to the differences between the scales. In this simple case, the conversion of the scales is well known, and a simple linear conversion can locate them on the same scale. Vertical scaling is less simple, particularly when the nature of the achievement being measured changes fundamentally across grades. Calculating a gain score or trajectory is in this case more akin to subtracting temperature on a Fahrenheit scale from humidity on a percentage scale, where no simple conversion either exists or is reasonable.

Compared to the gain-score model in the previous chapter, trajectory models are generally more dependent on vertical scales. This is because vertical scales become more tenuous as the grade span increases. For a simple gain-score model with only two adjacent grades, the vertical scale may be well supported. In contrast, trajectory models extrapolate from observed gains to future status in even higher grades. There, the argument for a common scale can be more difficult to support, particularly if the achievement measured in the higher grades cannot be mapped meaningfully to achievement measured in the lower grades. Depending on the uses of growth predictions, trajectories across particularly large grade spans may warrant caveats. Evaluation of the vertical scale is necessary across the entire range of grade levels through which the trajectory model extends.

Question 2.4:
What Kinds of *Group-Level Interpretations* can the Trajectory Model Support?

The average gain score for a group can be extrapolated as if it were for an individual, supporting group growth prediction. Alternatively, each student may be classified as "on track" by his or her individual trajectory. This can be aggregated to a group-level interpretation about the percentage of students who are on track.

The trajectory model supports group-level interpretations in at least two ways. One approach concerns average gains and average predictions. This requires calculation of the average gain score of the group. The gain is extended into the future to illustrate as if it were an individual trajectory, but it can be interpreted as the predicted average trajectory of all students in the group. A second approach begins with a straightforward standard setting approach described in the next section. This approach classifies a student as "on track" to a future target cut score if the student predicted status exceeds the cut score at the grade of interest. These student classifications can be aggregated into a "percentage of on-track-students" statistic.

Figure 2.2

Illustration of the Trajectory Model at the Aggregate Level for Three Students (A, B, and C).

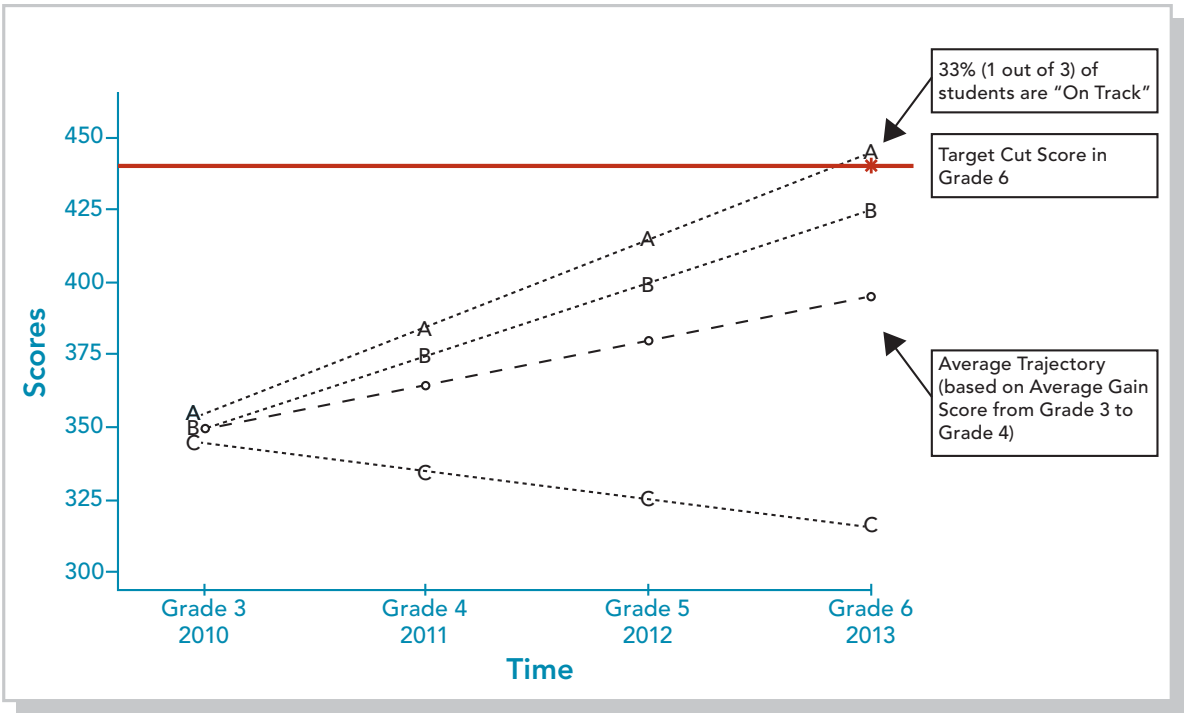


Figure 2.2 illustrates both of these group-level methods for the trajectory model. In this figure, the trajectory for the student used in Figure 2.1 is displayed as the line with score points denoted by the letter "B." Two additional students, A and C, are part of this student's group. Groups are a general construct and can be formed by students with a common teacher, school, school district, or demographic subgroup. For simplicity, assume that these three students comprise all fourth graders in a particular school. The average trajectory for these three students is shown by the thicker black line with open, black dots denoting the average scores at each time point. For both the students and the average trajectory lines, the first line segment connecting the scores from Grade 3 to Grade 4 is solid because it corresponds to observed gain, whereas the line segments between Grades 4 and 6 are dashed because they correspond to predicted gains.

Calculation of the average trajectory proceeds by taking the simple average of the three scores at each of the four time points, then simply connecting the dots. An alternative and algebraically equivalent formulation involves 1) taking the average scores of the observed time points in Grades 3 and 4; 2) connecting these two points to depict the average observed gain (the solid, bold line); and 3) extending this gain in a linear fashion through Grades 5 and 6 (the dashed, bold line). The average observed gains for students A, B, and C are +30, +25, and -10, respectively. The average gain of +15 is the group-level average gain, and the

trajectory shown in Figure 2.2 is the visual representation of this gain of +15 extrapolated in a linear fashion over the next two grades.

Figure 2.2 also shows a target cut score, set through a process described in the next section, that is established at 440 in Grade 6 and marked by a gray asterisk. The location of each student's predicted Grade 6 score can be compared to this line, and it is clear that only student A's predicted score exceeds this future standard. An alternative description of group-level growth prediction is that 1/3 or 33 percent of students are on track to the future standard. In practice, because students are either proficient, on track, or not on track, the percentage of on track students is either added to the percentage of proficient students or expressed as a percentage of nonproficient students who are on track Hoffer, Hedberg, Brown, Halverson, Reid-Brossard, Ho, & Furgol, 2011). The sufficiency of these percentages can be compared to minimum required percentages of proficient and on-track students (for example, Annual Measurable Objectives) that are set by other policy committees. The importance of standard setting is emphasized in this next section.

Question 2.5:

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

The adequacy of predicted student (or group) growth can be determined by the slope of the student trajectory or the student's predicted future status. At the group level, expectations can also be set on the average slope, the average predicted future status, or the percentage of students predicted to be on track to meeting a target future status. To identify any particular target future status, a time horizon must also be designated.

The trajectory model can support a variety of standards for expected growth. At the individual level, the slope of the trajectory can be compared to a standard, but this is equivalent to setting a standard on gain-scores, and this is described in the previous chapter. A more common approach, related to the model's primary interpretation of growth prediction, is to compare an individual's predicted future status to a standard. For any individual trajectory, this comparison requires two pieces of information: the time horizon to meet the standard and the cut score at that time horizon. Following the previous section, this could be expressed as 440 by Grade 6.

These standards follow from policies that might dictate, for example, student proficiency, or that students should be college and career ready by high school graduation. Proficiency in lower grades may take the form of college readiness cut scores in Grade 12 that are articulated down through earlier grades. Trajectories can be compared with target cut scores to evaluate whether students are on track.

Here, it is worth noting that there are two seemingly different but actually equivalent approaches to evaluating on-track status. First, the gain score can be extrapolated and compared to the future target cut score. In Figure 2.2, for example, this results in a statement like, "Student B has a gain of 25 points and is on track to a score of 425, which is below the target score of 440." Alternatively, the required gain could be calculated by comparing the future cut score with the initial status, calculating the required gain, and comparing this to the student's actual gain, resulting in a statement, "Student B gained 25 points this year but needed 30 to be on track to a score of 440." These two formulations are algebraically equivalent and should not be considered to be different models.

For trajectory models, the selection of the time horizon to meet a cut score is just as consequential a standard setting decision as the selection of the cut score itself (Ho, Lewis, & Farris, 2009). A longer time horizon to reach proficiency is generally more lenient and realistic, and a shorter time horizon is generally more stringent. Time horizons can be set by a fixed number of years from the time a student enters the data system. In Figure 2.2, the student must be proficient within three years of entering the system. If proficiency is required before exiting a school, the horizon can be set, for example, as "three years from entry into the system or by graduation, whichever is sooner." As a student progresses through grade levels, an additional decision must be made about whether to have a fixed time horizon for each student or allow the time horizon to shift and effectively reset, always staying, for example, two years ahead of the student's most recent completed grade.

Whenever cut scores in different grades serve as targets for a trajectory model, these cut scores must be articulated, that is, they must share a common meaning and, ideally, a similar level of relative stringency across grades. Without this articulation, counterintuitive results follow, including students who are on track to proficiency in Grades 4 and 6 but are not on track to proficiency in Grade 5. The issues of time horizons and articulated cut scores arise in any model for growth prediction that sets standards in terms of a future cut score.

Expectations can also be set on adequate growth at the group level. A group's average trajectory can be extrapolated to determine if, on average, the students in the group are predicted to meet/exceed the future target score. This was illustrated in Figure 2.2. The average trajectory in this illustrative example results in a predicted average Grade 6 score that is lower than the target Grade 6 score. Groups whose averages are not predicted to meet the target future score could be deemed as "not making adequate growth," and groups whose averages are predicted to meet the target could be deemed as "making adequate growth."

In contrast, standards can be set on the percentage of students who are predicted to be "on track." In practice, this percentage can be combined or cross-referenced with the percentage of proficient students. Each student can be classified into one of four mutually exclusive

categories: 1) proficient and on track, 2) proficient and not on track, 3) not proficient and on track, and 4) not proficient and not on track. Under a status model such as the original incarnation of the No Child Left Behind Act (NCLB), only the first two categories counted positively for a school. A growth model can count the first three categories positively, or it may count only categories 1 and 3. The former approach, one that takes the union of status and growth, was a popular strategy among states using the trajectory model for revised NCLB purposes (Hoffer et al., 2011).

Question 2.6:

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

The trajectory model is aligned with user intuition about growth over time. However, it is deeply dependent on the underlying vertical scale, and the model can create unusual incentives to artificially lower initial scores, inflating gain scores and thus trajectories.

Trajectory models are intuitively appealing because they allow for growth predictions that follow an assumption of linear growth over time. However, extrapolated predictions based on linear growth are not empirical as much as descriptive and aspirational, and the prediction requires thoughtful construction of an underlying vertical scale. Just as gain-score models can be distorted by vertical scales, trajectory models with poorly developed scales can have ceiling effects, floor effects, and spurious relationships between initial status and growth.

The equal-interval property assumed of vertical scales, where a gain in 25 points from Grade 3 to Grade 4 is assumed to be equivalent to a gain in 25 points from Grade 7 to Grade 8, can be more salient here than in gain-score models due to the extension of trajectories across a large grade span. In extreme cases, the predictions from trajectory models can extend to future score points that simply do not exist. Student C in Figure 2.2 is predicted to have an extremely low Grade 6 score that may not even be possible on the Grade 6 test. A nonsensical trajectory does not invalidate trajectory models but motivates thoughtfulness in reporting and use of model results.

Finally, as in the gain-score model, trajectory models that function in isolation can motivate not only increases in current scores, but decreases in past scores, as both will augment gains and increase predicted trajectories. A simple approach to diminishing this "fail-first" incentive is the application of a status model in conjunction with a growth model, where the artificial deflation of earlier scores is only an advantage if the scores do not fall below the status-relevant cut score.

References

Ho, A.D., Lewis, D.M., and Farris, J.L.M. (2009). The dependence of growth-model results on proficiency cut scores. *Educational Measurement: Issues and Practice, 28*(5): 15-26.

Hoffer, T.B., Hedberg, E.C., Brown, K.L., Halverson, M.L., Reid-Brossard, P., Ho, A.D., and Furgol, K. (2011). *Final report on the evaluation of the growth model pilot project*. Washington, DC: U.S. Department of Education, from http://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/index.html.

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.

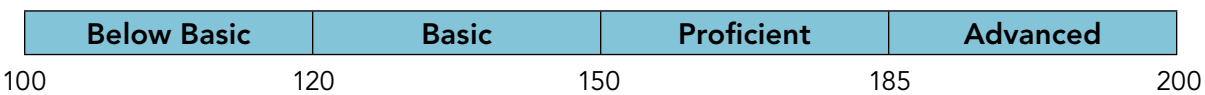
Question 3.1:

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				

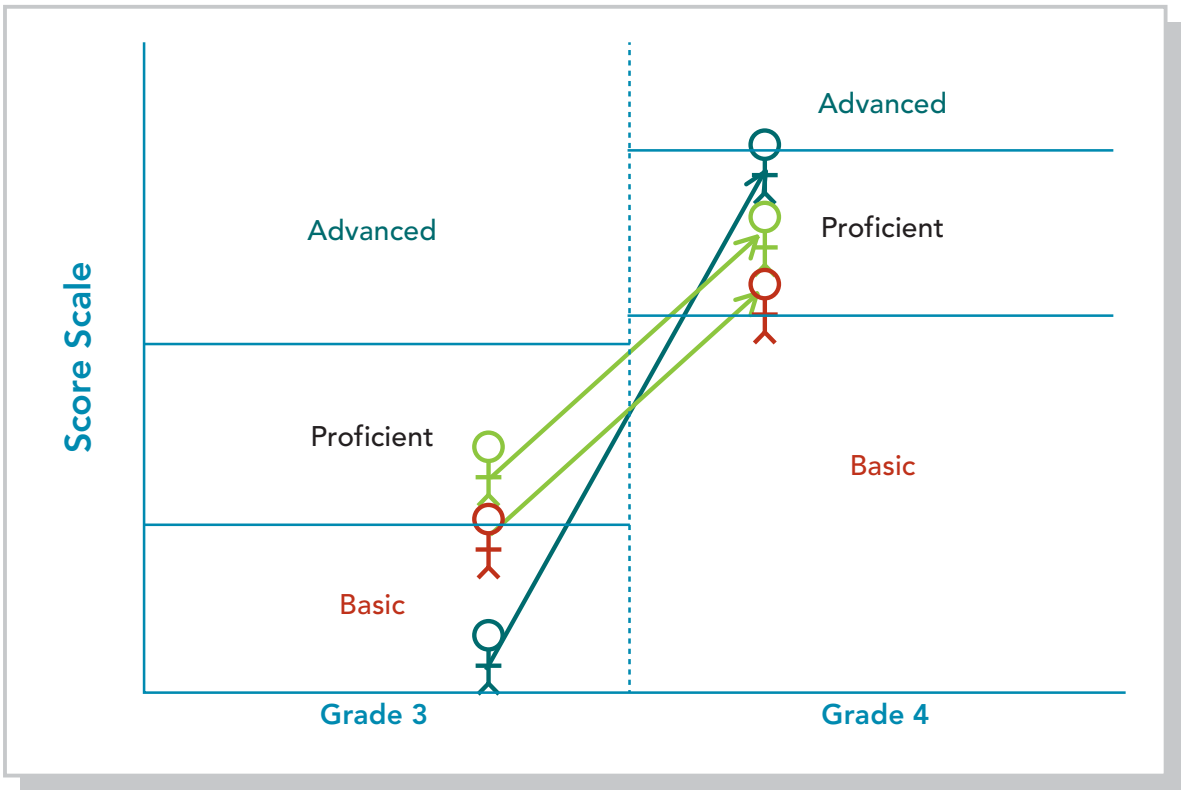
Table 3.1 illustrates the use of categorical models for growth description. This simple table shows the student of interest increased one performance level category. Within the Grade 3 domain of mathematics, the student only had a Below Basic understanding and mastery of the material. However, in Grade 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 4
The Residual Gain Model

The residual gain model can be motivated by concerns about the gain scores used in the gain-based models, particularly the purported low reliability of gain scores and ceiling effects for high-scoring students. The residual gain model uses linear regression to determine expected current status for students at different initial scores. These expectations are derived empirically given past scores. The residual gain is simply each student's observed current status minus his or her expected current status. This difference between observed and expected outcomes is commonly referred to as the "residual" in regression terminology. Residual gain scores represent the amount students scored above or below what was expected given their past performance.

Residual gain scores support **growth description** by answering the question

How much higher or lower has this student scored than expected given her past scores?

Because residual gain scores are the differences between observed and expected current status, they are also on the same scale as the current test score. They report current status in terms of, or "conditional upon," past scores, making them a **conditional status model** instead of a gain-based model.

Although the statistical model used in computing residual gains sets a statistical expectation for growth, residual gain models may require additional judgmental standards to determine what amount of residual gain represents "adequate" growth. This is described in Section 4.5. The following subsections address each of the six questions of interest to further elaborate on this model, particularly as it stands in stark contrast to the gain score model.

RESIDUAL GAIN MODEL

Aliases and Variants:
- Residual Difference Model
- Covariate Adjustment Model
- Regression Model
- Percentile Rank of Residuals

Primary Interpretation: Growth description

Statistical Foundation: Conditional status

Metric/Scale: Difference score (on the current-grade score scale)

Data: An interpretable scale, linearly related test scores

Group-Level Statistic: Average residual gain

Set Growth Standards: Setting expected or adequate residual gain score

Operational Examples: Evaluating a treatment

Question 4.1:

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

The residual gain model supports growth description by describing how much higher or lower a student scored than what was expected given her prior year's score.

The simplest form of the residual gain model involves setting expectations for current scores based on only one set of previous scores. In this case, the residual gain model and the gain score model can use the exact same data but describe growth in a fundamentally different way. Instead of describing how much a student changed this year from last year as the gain score model does, the residual gain model describes how much higher or lower a student scored this year than expected given last year's scores.

The residual gain model uses a statistical model known as linear regression to set empirical expectations for current scores given past scores. It is useful to note here, however, that linear regression in the residual gain model is for *describing* current scores given past scores and not for *predicting* future scores given current and past scores. This distinction is apparent when contrasting the residual gain and projection models in the next chapter.

Question 4.2:

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

Although the name "residual gain model" suggests that this growth model is gain-based, it is actually a conditional status model. Gain-based models involve taking a difference between current and past performance. In contrast, the residual gain model takes the difference between current performance and expected current performance given, or conditional upon, prior performance.

The residual gain model uses linear regression to calculate expected current scores given past scores. These expectations are statistical and empirically derived. Unlike the gain score model, scores from each included grade level do not need to be from vertically scaled assessments. This section explains the statistical model underlying the residual gain model for the simplest case of using data from only one prior grade level as a predictor in the linear regression model. However, it is straightforward and common to include greater numbers of previous grade scores, and the regression model is also fully capable of incorporating demographic variables to establish expectations as well.

Linear regression is a useful statistical method that supports prediction of an *outcome variable*, in this case, the current score, using one or more other *predictor variables*, in this case, one or more past scores. The choice of predictors is generally motivated by

associations between the predictor and the outcome, so that knowing a value on the predictor variable provides information about the value of the outcome variable. In this case, because relationships between past and current scores are generally moderate to strong and linear, the model often fits the data well. Linear regression provides expected values for the outcome variable by finding the line that best fits the averages of the outcome variable at each level of the predictors. This is most readily understood with an example and a graph, which follow for the residual gain model context.

The following example assumes a small group of students currently in Grade 4 with test scores from the current grade and the previous grade, Grade 3. For purely illustrative purposes, suppose there are only 8 fourth graders in the group of interest. Figure 4.1(a) provides a scatterplot of these students' Grade 3 and Grade 4 scores. The 8 students are represented by 8 solid dots. The horizontal position of the points is determined by the student's Grade 3 score and the vertical position by the student's Grade 4 score. This plot shows that students earned scores of 345, 350, or 355 in Grade 3, but earned scores ranging from 335 to 385 in Grade 4. The solid black line in Figure 4.1(a) represents the output of the linear regression model, a line that predicts Grade 4 scores given Grade 3 scores.

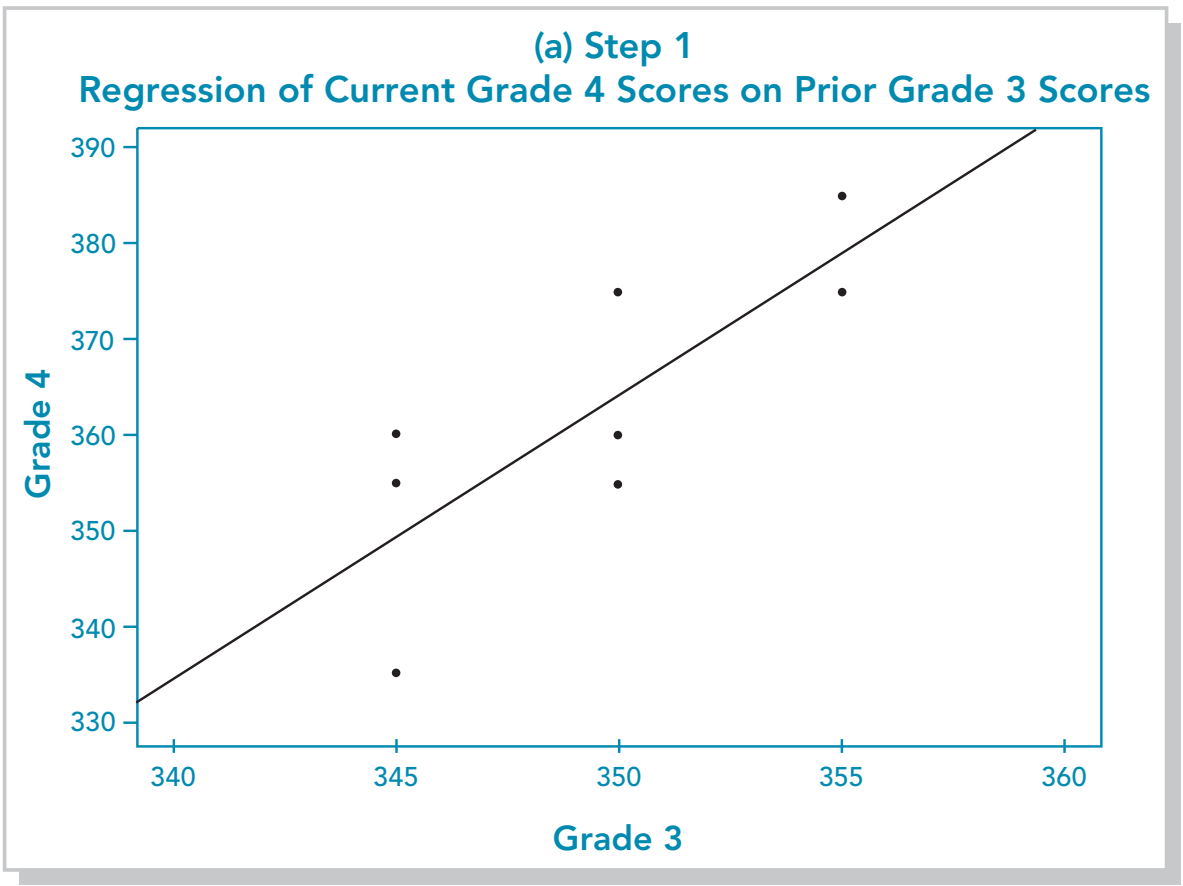
This line represents the best fit of the average Grade 4 score across all Grade 3 scores, in this case, all 3 of them. Unsurprisingly, the line goes roughly through the middle of each of the three vertically aligned sets of points at the Grade 3 scores of 345, 350, and 355. The line therefore represents the expected Grade 4 score at each possible Grade 3 score. For instance, in Figure 4.1(b) a dashed horizontal arrow from the linear regression line shows that at a Grade 3 score of 350, the expected Grade 4 score is 364. This result supports an interpretation like the following, "Students who earn a score of 350 in Grade 3 are expected, on average, to earn a score of 364 in Grade 4."

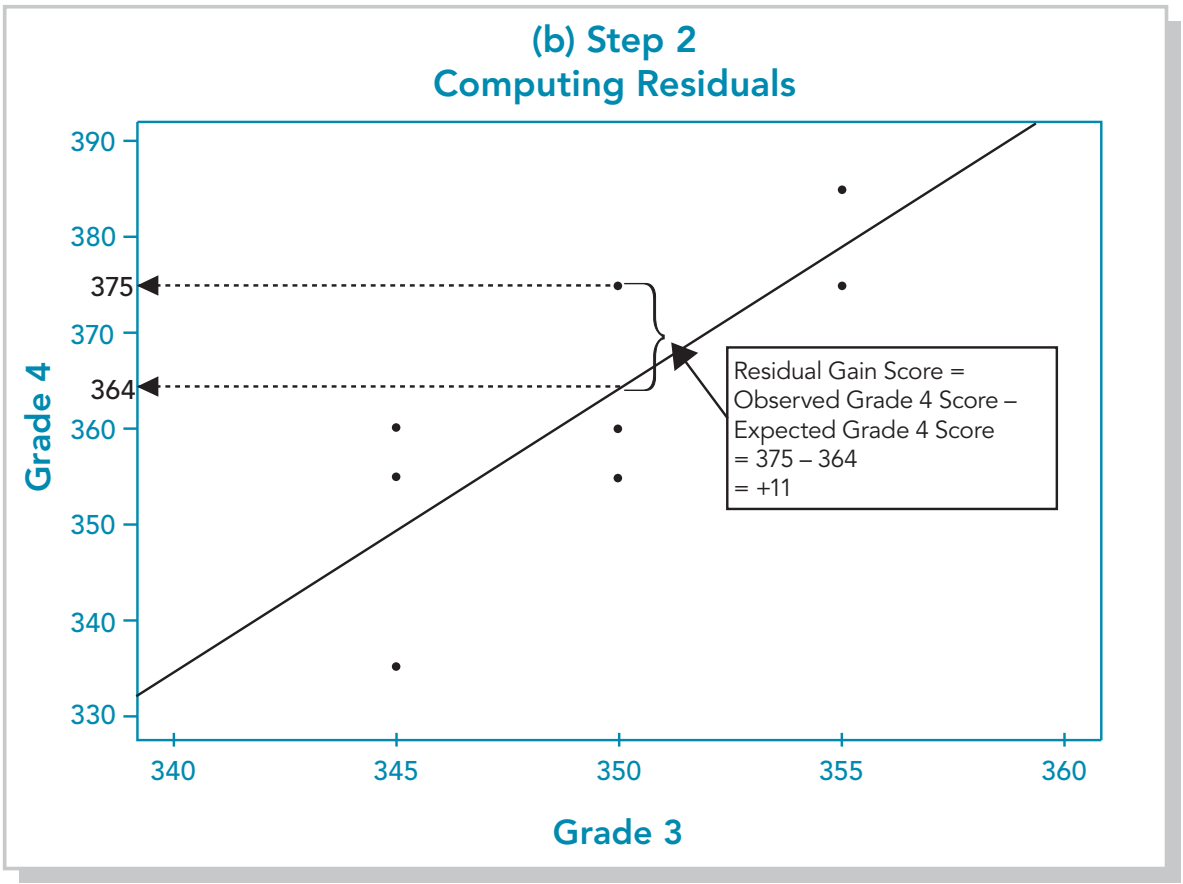
Fitting the linear regression line is only one step in the residual gain model. Figure 4.1(b) illustrates the next step that results in residual gain scores. The residual gain score is found by taking what is commonly called the residual, or the difference between the observed score on the outcome variable and the expected score on the outcome variable. In this example, this difference is between students' observed and expected Grade 4 scores. Figure 4.1(b) shows this difference for a particular student who earned a score of 350 in Grade 3 and a score of 375 in Grade 4. This student's expected score is empirically derived from the regression line as 364. The student's residual gain is the simple difference between the observed and expected score as follows:

Residual Gain Score = Observed Grade 4 Score – Expected Grade 4 Score
= 375 – 364
= +11

The student's residual gain score of +11 indicates that he scored 11 points higher on the Grade 4 test than expected given his Grade 3 score of 350. A negative residual gain score indicates that a student scored *below* his/her expected score. Graphically, the residual gain is visually represented by the vertical distance between any point and the regression line. Students above the regression line have positive residual gains, and students below the regression line have negative residual gains. This illustration demonstrates that the residual gain score does not truly represent a gain, a change in points from one grade to the next, as in the gain score, trajectory, and categorical models. Instead, it is achievement beyond expectations given past scores.

Figure 4.1
Illustration of the Residual Gain Model





Question 4.3:

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

Residual gain models, and conditional status models in general, do not require test score scales to be linked across grades. This is due to their emphasis on conditional status, that is, status beyond expectation, instead of growth over time. Like any growth or status model, residual gain models require appropriate within-grade scales. The assumptions of linear regression must be met, including linear relationships between current and past scores and similar amounts of variation in current scores for any particular past score. When these latter assumptions are not met, more flexible regression models can be used.

By framing growth in terms of conditional status, the residual gain model is applicable to a broader range of test score data than gain-based models. The scores of interest do not need to be linked on a common vertical scale across grades, and the model can easily

accommodate more than one prior year of data if desired. Although the various grade level test scores do not need to be linked on a vertical scale, they do need to be linearly related. One approach to evaluating this is through plots of the current grade level scores against each of the prior grade level scores, where the relationship should look linear, roughly like Figure 4.1. When there are nonlinear relationships, inaccurate expectations and thus inaccurate residual gains can result.

An additional requirement of regression models is that the conditional variability of outcome scores should be similar across different levels of the predictors. In Figure 4.1, this can be visualized in terms of the spread of points around the regression line at each vertical slice, 345, 350, and 355. At each level, the overall variation should be similar. In the case of Figure 4.1, it may seem as though the variability at the score level of 355 is smaller, that is, the points are clustered closer to the line, but the sample size is far too small to make such a determination. However, in a large sample situation, when the variability is not equal across predictor values, higher scoring students may have far more or less variable residual gain scores than lower scoring students. This may be an observation that reflects reality, but if it is instead an artifact of the scaling of the test, an alternative regression model, like those used in Student Growth Percentiles, may be warranted.

To understand why vertical scaling is not required of conditional status models, it is most helpful to reframe the nature of the growth that these models measure. This growth is less a fixed quantity that is being estimated and more a comparison between status and a key concept: *expectations*. These expectations can be based on prior year scores from a single grade, as in Figure 4.1, or a collection of prior year scores from multiple grades. However, the regression model does not consider these prior grade scores as a trajectory over time, but an unordered combination of facts that generate an empirical expectation.

In the context of a newborn growing over time, the gain-based approach tracks the weight over time, from 8 pounds to 9 pounds to 10 pounds at one, two, and three months, respectively, for example. The conditional status model asks instead, given that the newborn was 8 pounds at one month and 9 pounds at two months, how much heavier is she than expected at three months? We could also add, given that this newborn is a girl, and breast-fed, and from the United States, how much heavier is she than expected at three months? Each variable that is added, or conditioned upon, changes the expected weight at three months, and it is clear the variables that set these expectations need not be on the same scale. For example, it is clear that the sex and nationality of the newborn are not on the same scale as the outcome. The regression model is a tool for setting expectations, and, as such, it does not require the variables that set these expectations to be on the same scale as the outcome or each other.

Question 4.4:

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

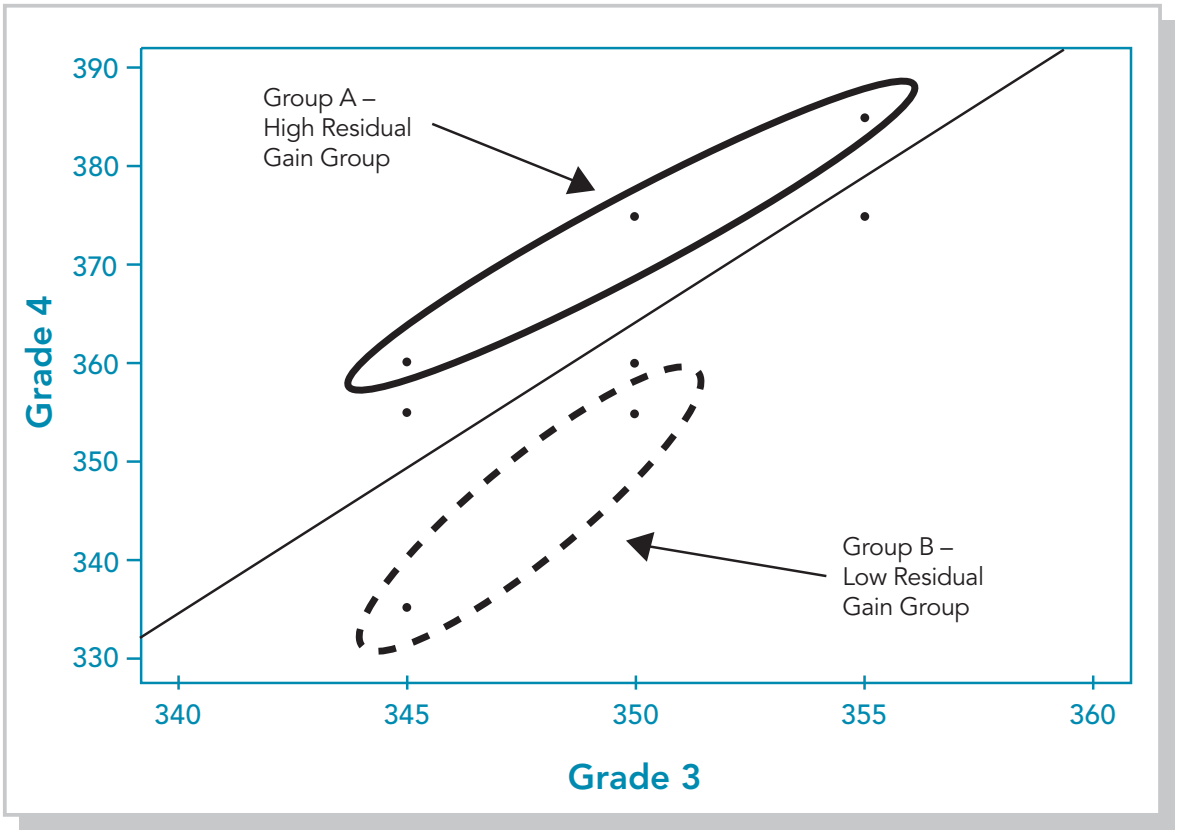
The residual gain model supports simple group-level statistics through averaging. The residual gains of a group can be averaged within a grade, although comparison of averages across grade levels requires a common across-grade scale. The average residual gain score represents the average amount students in the group scored above or below expectations given their past performance.

Several group-level statistics can be computed for the residual gain model to summarize the performance of all students in a group. The most typical summary statistic is the average residual gain for the students in a group. As a technical point of reference, it is worth remembering that, across the entire dataset to which the regression is applied, the average residual is always zero. In Figure 4.1, with a hypothetical group of 8 fourth grade students, the mean residual gain score across all 8 students is zero. This should be intuitive. If the regression model is working properly, the average expected value should be the same as the average observed value. However, for any subgroup of the 8 students, the mean residual gain score is not necessarily zero.

The sign and magnitude of the average residual gain score reflects the average status of students in the group of interest, above and beyond expectations. Figure 4.2 helps to illustrate group-level performance as measured by the residual gain model. Figure 4.2 is a reproduction of Figure 4.1(a), but, in this case, there are circles around some collections of points to indicate different groups, in this case, hypothetical small classrooms of students. One set of students is labeled as "Group A" and another as "Group B." The three students in Group A have varying prior Grade 3 scores, but all have points above the regression line, indicating that all of these students have Grade 4 scores greater than expectations based on Grade 3 scores. Their residual gain scores are about 10.51, 10.64, and 5.77 from left to right in the figure. The simple average of these three residual gain scores is around 9.

This average residual gain of 9 can be interpreted as, "Students in Group A, on average, scored nine points higher than expected given their prior year scores." In other words, given their Grade 3 scores, on average, these students exceeded expectations for their Grade 4 test by 9 points. Group A is thus labeled as a "High Residual Gain" group in Figure 4.2. In contrast, Group B's two students performed worse than expected given their initial scores. Both of these students have points that lie below the regression line and thus have negative residual gain scores. These residual gain scores are about -14.49 and -9.36, which results in an average residual gain score of around -12. On average, Group B's students scored about 12 points below expected on the Grade 4 test given their Grade 3 scores. Relative to other students with the same prior Grade 3 scores, these students performed worse on the Grade 4 test than expected, making them a "low residual gain" group.

Figure 4.2
Group-Level Interpretations from the Residual Gain Model



This example is rather simplified as it involves extremely small groups comprised of students who had either all negative residual gain scores or all positive residual gain scores. In practice, groups will likely have a mixture, but summary statistics like the mean, median, and standard deviation of residual gains can summarize the patterns of student status beyond expectations for groups.

A more formal statistical approach to simple averages of residual gains is known as **the covariate adjustment model**. Instead of growth description, the covariate adjustment model primarily supports *value-added* interpretations. It is called a covariate adjustment model because it adjusts expectations about current status using various predictor variables, just as the residual gain model does. It contrasts with the residual gain model by providing formal group-level estimates of group status compared to a baseline by explicitly incorporating group membership variables in the model. These group-level estimates can support discussions about whether group membership, whether it is to a classroom or school, predicts student test scores above and beyond past scores.

The intuition behind the covariate adjustment model is nearly identical to that supporting Figure 4.2. Classroom and school estimates from covariate adjustment models are in fact strikingly similar to averages of residual gain scores in practice. However, the covariate adjustment model fits separate regression lines for each group and compares these lines to each other, where higher lines imply higher status beyond expectation. This is a statistical improvement over the ad hoc, two-step approach of averaging residual gains after the regression model has been fit.

The underlying similarities between the residual gain model and the covariate adjustment model allow for deeper insight into the use of these models for value-added interpretations. The residual gain model is used for growth description. This growth is best described as status above and beyond expectations set by other variables. At the group level, an average residual gain is a statement about a group's average status beyond expectations. The covariate adjustment model supports both a statistical and substantive extension of the averaged residual gain approach. The statistical extension is an improved method for estimating average status beyond expectations. The substantive extension is the assumption that this average status beyond expectations is the value that the educator or school adds to the average test scores in the group.

Question 4.5:

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

The residual gain model references expected status given past performance. Such expectations are statistically defined and do not relate to what amount of growth is "adequate" in an accountability setting. Value judgments can be made by an informed committee about thresholds for adequate student-level and group-level (average) residual gain scores for particular grades and subjects.

As a linear regression model, the residual gain model sets statistical expectations for current performance given past performance. Accordingly, this model allows for computations of how much students deviate from an expected level of performance, resulting in residual gain scores. However, the residual gain score in and of itself does not indicate whether improvement was "good enough" in the settings of accountability or evaluation. Such judgments require additional input by invested stakeholders.

One approach involves selecting a standard and operationalizing it as a cut score on the residual gain metric. The cut score can be set on the scale itself if there is clear understanding of what 5, 10, or 50 points above expectations actually means on the score scale.

Alternatively, the standard can be set normatively, such as defining the top 30 percent of residual gain scores as exceeding expectations. Alternatively, the residual gains can be sorted and reported as percentile ranks, resulting in percentile ranks of residuals. In practice, these percentile ranks of residuals are very similar to Student Growth Percentiles (Castellano & Ho, in press). This normative approach can support comparisons of residual gains across different grades and subjects.

The residual gain model sets expectations empirically for a particular group of interest. By definition, for this group, approximately half of the residual gain scores will be positive and the others negative. Setting standards on a fundamentally relative metric may be undesirable as, ironically, growth over time will be difficult to measure. An alternative approach involves assuming that residual gains will persist over time into the future, and comparing these future scores to future cut scores. This extension shifts the primary interpretation of the residual gain model from growth description to growth prediction, but it allows for standards to be set on the residual gain metric that are free from the "tyranny of averages" where approximately half of students will always be below average.

Question 4.6:

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

The residual gain model is something of a misnomer, as it is less a gain than it is status beyond expectations given past scores. When assumptions of the linear regression model, including linearity and common outcome variance across prior scores, do not hold, residual gains can be systematically distorted for higher or lower scorers.

The residual gain model is not a central feature of any active state accountability systems, although it serves as a basis or helpful contrast for many active models, including its close cousin, Student Growth Percentiles. Its most natural extension, the covariate adjustment model, is one of the most common models supporting value-added interpretations. The model is often used in experimental research where there is interest in the effectiveness of a treatment in a pretest/posttest design.

An obvious misinterpretation of the residual gain model would be to assume it describes growth over time in a similar manner as the gain-score model. As this section has demonstrated, the residual gain is a fundamentally distinct quantity from the gain score. It is a difference between an actual score and an expected score. The expected score is derived empirically from past scores and will change if different combinations of variables are used to establish expectations.

If residual gain scores were used in a high-stakes system, the model assumptions — linearity and common outcome variance across prior scores — become more important. Violations will lead to systematic relationships between initial status and the average and variability of residual gains. More generally, residual gain models, like gain-based models, share the property that may incentivize "gaming" the system by artificially decreasing students' initial scores so as to increase their residual gains. This can be visualized in Figure 4.1b, where points that shift to the left, that is, declining in initial scores while maintaining current scores, will have a larger residual gain. Of course, unlike gain-based models, the shifting of points changes the empirical expectations, thus this strategy only works if these shifting points have a negligible effect on the regression line.

The empirical derivation of expected scores using extant student data is a reminder that residual gains are based on the performance of their peers. It follows that expectations will change if different students were included in the regression analysis. This is a property of all conditional status metrics. The word "conditional" emphasizes that any growth interpretation is conditional on prior performance — not just of the student of interest, but all students in the cohort of interest. Returning to the example presented in this chapter, a student who is in fourth grade next year could earn the exact same Grade 3 and Grade 4 scores as a student in this year's cohort, but receive a different residual gain score if the students in general performed differently. In particular, if the relationship between current and prior scores is distinct from the one presented in Figures 4.1 and 4.2, the fitted regression line will be different, resulting in different expected scores and, in turn, different residual gain scores.

Reference

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*.

CHAPTER 5
The Projection Model

The projection model, sometimes known as the prediction or regression model, is primarily used to project or predict scores in a future grade, supporting **growth prediction**. It fundamentally answers the question

Given this student's observed past scores, and based on patterns of scores in the past, where is she likely to score in the future?

The projection model relies on linear regression to answer this question. The model uses test score data from a past cohort of students who have already completed the future grade of interest to estimate a prediction equation. This equation is then applied to the data for a current cohort of students to predict their future scores. A necessary step in establishing a projection model is the determination of a time horizon to which the model will predict future status.

The predicted future status can be evaluated with respect to a future standard such as "Proficiency." Predicted status above this standard can support the judgment that the student is "on track" and making "adequate growth."

PROJECTION MODEL

Aliases and Variants:
- Regression model
- Prediction model

Primary Interpretation: Growth prediction

Statistical Foundation: Conditional status model

Metric/Scale: Score scale of test in the future target grade level

Data: Interpretable future scale or future standard

Group-Level Statistic: Average future prediction or percentage of on-track students

Set Growth Standards: Define future standard, minimum time until standard is reached

Operational Examples: NCLB Growth Model (e.g., Ohio and Tennessee)

Question 5.1:
What *Primary Interpretation* Does the Projection Model Best Support?

The projection model uses a statistical technique to predict future scores from current and prior year scores. It is specifically designed to support growth prediction.

The projection model is designed to predict student test scores in a future grade. Relying on the statistical tool of linear regression, this model allows for interpretations like, "On average, students with a score of 110 on the Grade 3 mathematics test and 250 on the Grade 4 mathematics test have a predicted Grade 5 mathematics score of 275." The predicted scores can be compared against a target score, such as the future grade's proficiency cut score, to support interpretations about adequate growth.

The projection model and the trajectory model both support growth prediction; however the projection model operates under fundamentally different assumptions and data requirements than the trajectory model. A simple way to describe the contrast is that the projection model is more data-driven, whereas the trajectory model is more scale-driven. The projection model uses regression to maximize the predictive accuracy of the model. If a variable does not contribute to the prediction of future status, the regression model will assign it a lower weight. In this way, the projection model is informed by the data and results in an equation that maximizes predictive accuracy.

In contrast, the trajectory model is scale-driven. It relies on the construction of a vertical scale and the assumption that a linear extrapolation of observed trajectories is defensible. Because it is less reliant on data-driven predictions, it is, as noted in Chapter 2, more of a descriptive and aspirational model than an empirical model.

The projection model approach to growth prediction can be taken to a mercenary extreme. Any available variable can be used to increase predictive accuracy, extending beyond previous test scores in the same subject to test scores from different subjects, demographic variables, and classroom- and school-level variables. If predictive accuracy is the primary goal, inclusion of these variables can be well motivated even as it becomes detached from an intuitive idea of growth. If the model is intended to create incentives to maximize student growth, prediction may be less important than communicating information that supports educator efforts to increase student growth.

Question 5.2:

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

The projection model is an example of a conditional status model. Given current and past scores, the model predicts a future status. Unlike gain-based models, growth is not defined as an increase in some quantity over time. Instead, current and past scores are used as unordered inputs to a weighted prediction equation for future status.

The projection model, like the residual gain model, uses linear regression for prediction and the setting of expectations given past scores. Unlike the residual gain model, the outcome variable is not the "current" year score but a future score for which a prediction is desired. Although both the residual gain model and the projection model use linear regression, the differences between the models are more substantial than, for example, the difference between the gain-score and the trajectory model. The projection model is not an "extension" of a residual gain score in the same way that the trajectory model is an extension of a gain score. The residual gain model describes the difference between current status and an empirical expectation for current status. The projection model establishes an empirical expectation for future status, period. The next paragraphs review the example used in the previous chapter and adapt it for the primary goal of growth prediction.

As was noted in the previous chapter, the residual gain model provides a score for each student that denotes how much a student scored beyond expectations given past scores. In the simplest case, only one prior year score is included in the regression. The current year score is the outcome variable and the prior year score is the predictor. Figure 5.1 illustrates this scenario by reproducing Figure 4.1(b), where a small group of eight Grade 4 students has their Grade 4 scores plotted on their Grade 3 scores. Each point in Figure 5.1 represents a student, where the horizontal location of the point is determined by the Grade 3 score, and the vertical location is determined by the Grade 4 score.

In the residual gain model, the prediction of the outcome variable is an intermediate step on the way to the residual gain score calculation. The predicted outcome is for the current year score, which has already been observed for this set of students. The interest is in the distance between the observed outcome and this predicted or, more specifically, *expected* outcome. This difference between the observed scores and expected Grade 4 scores is called a "residual" in the context of regression and a "residual gain score" in the context of this guide. The *projection model*, in contrast, focuses on the prediction itself, but for a different set of students who have not yet taken the Grade 4 test.

Figure 5.1
Illustration of the Residual Gain Model: Regression of Grade 4 Scores on Grade 3 Scores

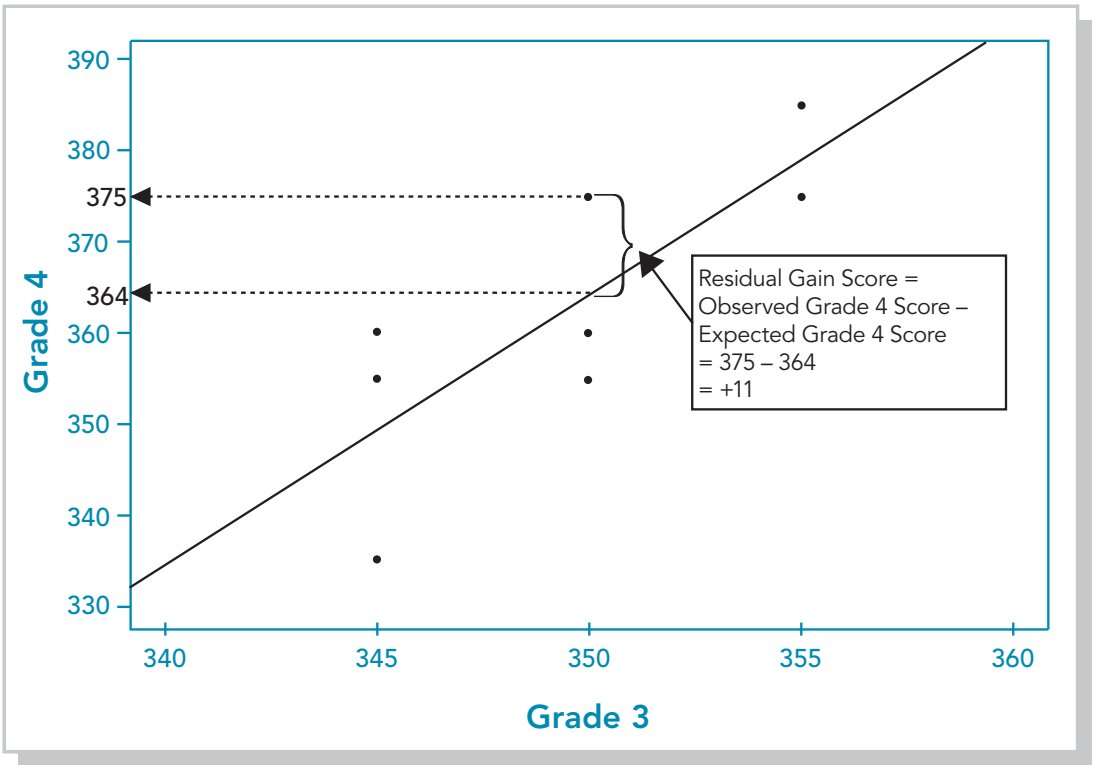
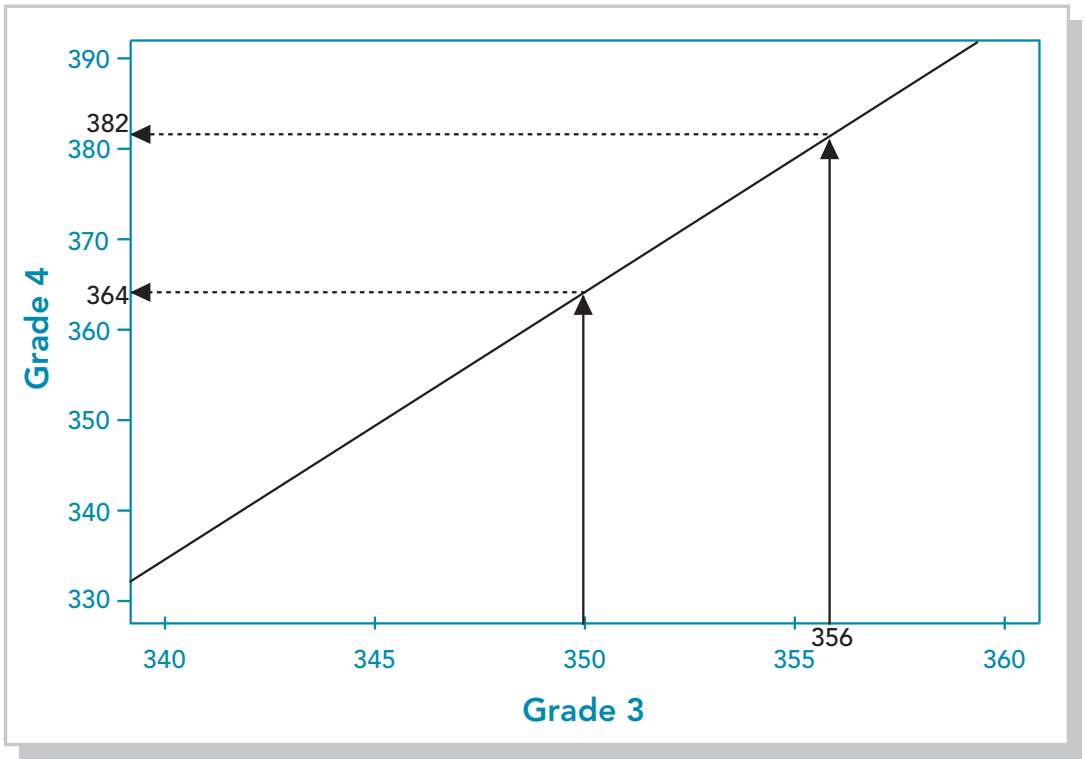


Figure 5.1 illustrates the computation of the residual gain score for a particular student. This student earned a 375 in Grade 4 and a 350 in Grade 3. It is clear from the graph that the student scored higher than the other two students who also scored a 350 in Grade 3. The regression line establishes the student's expected Grade 4 score at about 364, 11 score points below the observed score of 375. The residual gain score for this student is +11, indicating a score that is 11 points higher than expected given past performance.

The residual gain model allows for growth description for the students used to fit the regression. The projection model, on the other hand, takes the linear regression fitted for one cohort of students and applies it to another set of students who have yet to reach the future grade of interest. Using the same example, we may switch our primary interpretation from growth description for current Grade 4 students to growth prediction for current Grade 3 students. The current Grade 3 students will not enter Grade 4 until the next academic year. Their Grade 4 scores are not known, but their Grade 3 scores are. However, the prediction line in Figure 5.1 can be estimated from the current Grade 4 students who do have data. Then, this line, which was used to provide *expected* Grade 4 scores for the current fourth graders in the residual gain model, can be used to *predict* the future Grade 4 scores of the current third graders.

Figure 5.2
The Projection Model: Using a Prediction Line Estimated from one Cohort to Predict Grade 4 Scores for another Cohort



To illustrate this prediction process, Figure 5.2 reproduces the exact same prediction line estimated in Figure 5.1. Although this line is estimated using the scores of the students shown in Figure 5.1, these students are no longer of interest and are not shown. Instead, their prediction line is used to predict Grade 4 scores given current Grade 3 scores. Figure 5.2 illustrates predictions for students earning Grade 3 scores of 350 and 356. From the previous discussion, the expected or predicted Grade 4 score is 364 for students who scored 350 on the Grade 3 test. This is illustrated by the solid arrow going from the Grade 3 score of 350 to the regression line and then from the regression line to the vertical axis at the Grade 4 predicted value of 364.

The regression line allows for Grade 4 score predictions based on any possible Grade 3 score, not just for students at the score values of the cohort from which the line was derived. For instance, Figure 5.1 contains no students in the current Grade 4 cohort who scored a 356 on the Grade 3 test. However, a student in the current Grade 3 cohort may have a score of 356, and this student will still have a prediction, 382, as shown in Figure 5.2. This calculation is supported by a prediction equation that is the output of the regression model. In this example, the prediction equation is

$$\text{Predicted Grade 4 Score} = -677.667 + (2.974)*(\text{Observed Grade 3 Score})$$

where -677.667 is the intercept and 2.974 is the slope or regression weight for the prior observed Grade 3 score. Any student with an observed Grade 3 score can be entered into this equation to find a predicted Grade 4 score. For instance, entering 350 and 356 into this equation for the "Observed Grade 3 Score" will return the predicted values shown in Figure 5.2. It is clear that this regression equation can only be estimated using data for students who already have Grade 4 scores. The projection model thus requires longitudinal data from a past cohort of students that have test scores in all predictor and target grades.

Figures 5.1 and 5.2 are the simplest versions of the projection model where there is only one predictor. In practice, projection models make predictions much farther into the future than one year and use more than one year of data as a predictor. With a large enough longitudinal dataset that spans 6 grades, a prediction equation can be estimated to support predictions for current Grade 5 students on the future Grade 8 test. In such a scenario, the current Grade 5 cohort may use scores in Grades 3, 4, and 5 to support their predictions. The prediction equation takes the following form:

$$\text{Predicted Grade 8 Score} = \text{Intercept} + [a * (\text{Observed Grade 3 Score})] + [b * (\text{Observed Grade 4 Score})] + [c * (\text{Observed Grade 5 Score})]$$

Here, *a*, *b*, and *c* are simply placeholders for the estimated regression weights. The intercept is the predicted Grade 8 score when the Grade 3, 4, and 5 scores are all zero, which does not mean that zero must be a possible score for each grade-level test. The intercept is needed to anchor the regression line and is usually not an interpretable value in a practical setting. In this

case, and any case in which there is more than one predictor, it is no longer possible to graph the relationships in two dimensions, however the intuition of fitting a model to set expectations and maximize predictive accuracy still applies.

Question 5.3:

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

The projection model does not require vertical scales underlying different grade level tests and can accommodate as many predictor variables as are available. The model does rely on regression assumptions, such as linear relationships between predictors and the outcome, for predictive accuracy. The projection model also requires longitudinal data over a significant grade span. To obtain a prediction equation for a future target grade, the model must use a previous cohort of students with longitudinally linked data from the earliest grade that supports prediction to the target grade of interest.

The projection model is flexible in the types of variables it can accommodate, but is demanding in terms of the data required to produce growth predictions. The model is more flexible than gain-based models in not requiring a vertical scale, and many prior years of data can function as predictors along with non-test-score variables, if desired. However, with greater numbers of grade-level and subject area tests included as predictors in the model, the percentage of students with missing data will be higher and may need to be addressed through "imputation" of missing values, where missing data are estimated according to assumptions.. Missing data will be an issue with both the current cohort that requires prediction and the previous cohort that supports the prediction equation.

The projection model requires selection of predictor variables and the future target outcome of interest. Once these are selected, a cohort must exist that has longitudinally linked data for all of these variables. In the example in the previous section, where three recent grades of data are used to predict an outcome three years into the future, the model requires longitudinal data spanning six years. This past "reference" cohort will generate the prediction equation. There is also a requirement that this reference cohort be substantively similar to the current cohort. Substantive differences between the cohorts may result in an irrelevant regression equation and poor prediction.

The use of the regression model requires attention to regression model assumptions. Like the residual gain model, the projection model assumes a linear relationship between the outcome variable and the predictors. If there are nonlinear relationships, this will degrade the overall predictive accuracy of the model and may lead to inaccurate predictions for students with particular patterns of scores.

Finally, if the projection model's predicted scores are compared to standards in that particular grade, some articulation of standards across grades is necessary to prevent counterintuitive

findings. For both the trajectory and the projection model, highly variable standards across grades can lead to nonsensical results where, for example, students are on track to proficiency in Grades 6 and 8, but not Grade 7.

Question 5.4:
What Kinds of *Group-Level Interpretations* can the Projection Model Support?

Projection models result in predicted scores that can be aggregated to average predicted scores. Alternatively, individual students can be classified as satisfactory or "on track" to some future standard based on their predicted future score, and a group-level statistic can be the percentage of students who are on track to reach the future target score.

The projection model can produce two useful group-level statistics — an average predicted future score and a percentage of students "on track" to some future standard. The projection model uses the estimated prediction equation to provide predicted scores for all students. These may be averaged for a group of interest. Other summary statistics, like the median and standard deviation, can be used to describe the central tendency and variability of the predicted scores of a group. Using the example from Figures 5.1 and 5.2, if a particular group of interest has three students with Grade 3 scores of 350, 350, and 356, these can be readily inserted into the prediction equation. The Grade 4 predicted scores are 364, 364, and 382 respectively, and the average predicted Grade 4 score is 370.

This average can be interpreted as, "Based on their Grade 3 performance, the students in this group have an average predicted Grade 4 score of 370." This average predicted score can be compared against a future standard, such as the Proficient cut score in Grade 4. If the average predicted score is above the target score, then, on average, the average student in the group is predicted to exceed the standard. Standard setting committees could also determine cut points for which average predicted scores might correspond to "low," "typical," or "high" group growth.

If an individual's growth to a standard is the primary focus of accountability, the predicted status of each individual can be compared to the future standard. If a student's predicted status is higher than the future standard, that student can be considered to be "on track." Group performance can be summarized by the percentage of students in the group who are predicted to meet or exceed the future standard. If, in our example, the Grade 4 standard of interest is a proficiency cut score of 375, then only one of the three students is predicted to exceed this target, resulting in the group having 33 percent (1/3) of its students on track. Additional standards could be set for gauging whether this percentage is adequate.

Question 5.5:

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

The projection model returns a predicted future score for each student. This score can be compared to a target cut score or otherwise evaluated for adequacy. Similarly, the aggregation of predicted scores for a particular group, for example, into an average predicted score, can be compared to a group-level standard, and the percentage of students on track to the target cut score can be compared against some desirable threshold.

The trajectory and projection models both support growth prediction and offer predicted scores on the scale of the test at the target grade. These scores can be compared to the relevant cut score at the target grade. This may be a cut score that has been previously set for another purpose, or it may be an alternative cut score established with explicit attention to the role of growth prediction. The decision rule is then as simple as deeming students as "on track" if their predicted score exceeds the standard. Finer grain categorical distinctions are also possible. There may be multiple standards for both students' predicted scores and for groups' average predicted scores. These additional cut scores could distinguish among different levels of growth, such as "low," "typical," and "high."

Like the trajectory model, growth predictions can be updated each year that new data become available. Students transitioning to a new grade may use the prediction equation that includes the most recent grade as an additional predictor. A decision also needs to be made about whether the time horizon for prediction should be a moving window of, say, three years, or if it should diminish with each year the student is in the growth model. This might, for example, require a student to actually reach a standard (instead of merely being on track) within three years or before graduation from the school, whichever is sooner. As with the trajectory model, the number of years to the target time horizon of interest is a consequential standard setting decision.

As each year brings new data, the prediction equations themselves may be updated. It may be more desirable to fix prediction equations for multiple year windows instead of recalculating them annually. In spite of a possible degradation in prediction accuracy, fixing prediction equations keeps two students with identical score patterns from having different predictions from one year to the next. Instability in prediction equations is akin to instability in standards and may be minimized to allow standards to gain consistent meaning over time.

Question 5.6:

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

The metaphor of "projection" can imply an extension from a current trend, thus the projection model is often incorrectly assumed to function like the trajectory model. Pursuing a goal of prediction can lead to diminishing returns for the goal of incentivizing growth.

The word "projection" is consistent with both prediction and the extrapolation of a line, thus the projection model is often assumed to work the same as a trajectory model. Instead, the two contrast starkly, and no trajectory over time is modeled or even recoverable from the construction of the projection model.

When the cohort that estimates the prediction equation differs from the cohort whose scores are predicted, poor prediction and systematic distortions can be introduced into the model. The prediction equations will also tend to degrade over time as the relationships between grade-to-grade scores change with shifting instruction and accountability structures. More generally, violations of the linear regression model, including nonlinearity of relationships between target and predictor grades, will have similar negative effects on prediction accuracy.

Finally, strict adherence to the goal of predictive accuracy is likely to diminish the formative potential of this particular model. First, maximizing prediction motivates the incorporation of ancillary predictor variables that may have weak substantive justification, like including scores from other subjects or demographic variables. These will improve prediction but are poorly aligned with intuition about classroom learning. Second, teacher response to a student with low predicted growth does not follow from the model, particularly when so few of the variables are under the teacher's direct control. Trying to maximize the accuracy of future predictions seems at odds with the classroom goal, which is, ideally, rendering predictions for low-scoring students inaccurate. When multi-predictor prediction equations show that no score on any single test is sufficient to raise a low-projection student to an on-track designation, the predictive accuracy of the model seems to diminish the incentives to teach these students. Although a status model layered over a projection model can provide more hope for these "condemned-by-prediction" students, gain-based alternatives like trajectory models may allow for improved incentives while preserving a reasonable level of predictive utility (Hoffer, Hedberg, Brown, Halverson, Reid-Brossard, Ho, & Furgol, 2011; Ho, 2011).

References

Ho, A.D. (2011). *Supporting growth interpretations using through-course assessments*. Austin, TX: Center for K–12 Assessment & Performance Management, ETS, from http://www.k12center.org/rsc/pdf/TCSA_Symposium_Final_Paper_Ho.pdf.

Hoffer, T.B., Hedberg, E.C., Brown, K.L., Halverson, M.L., Reid-Brossard, P., Ho, A.D., and Furgol, K. (2011). *Final report on the evaluation of the growth model pilot project*. Washington, DC: U.S. Department of Education, from http://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/gmpp-final.pdf.

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

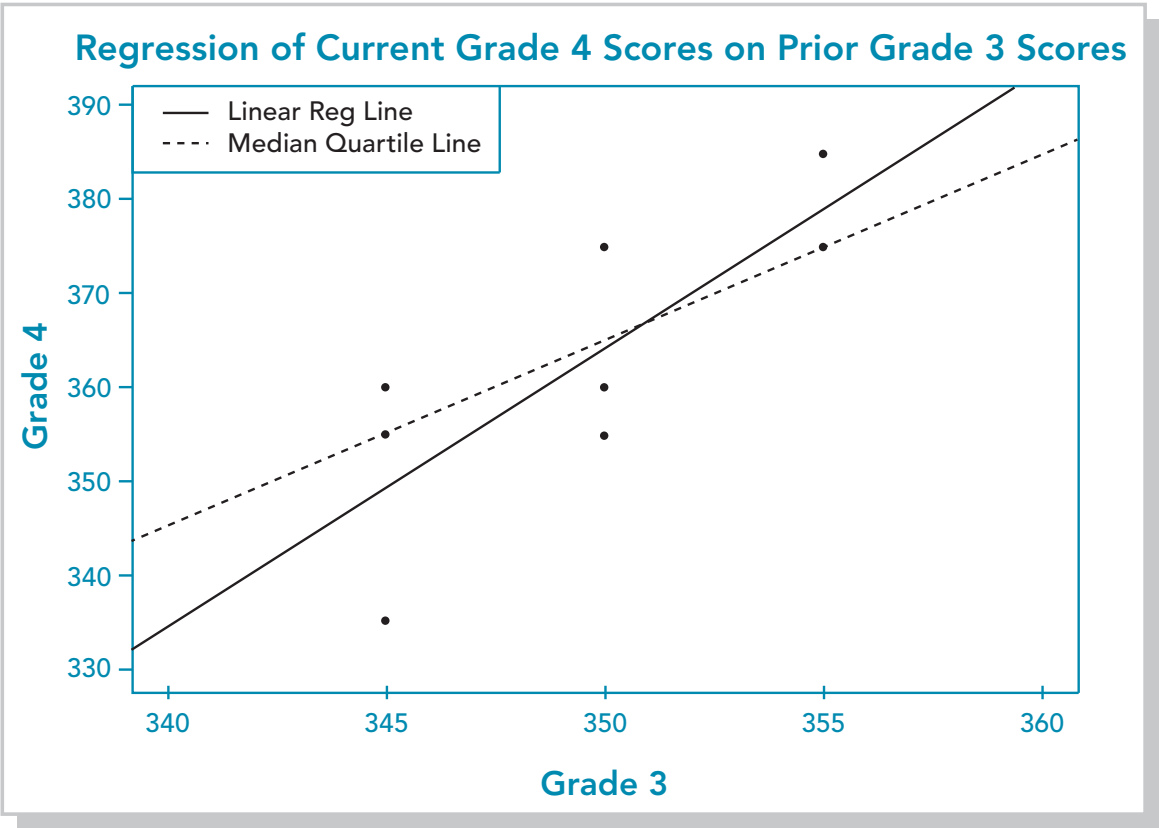
Group-Level Statistic: Median/mean SGP – describes the average/typical status of students relative to their past performance, or percentage of students on-track (to a future standard)

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 Quantile 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 x 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?

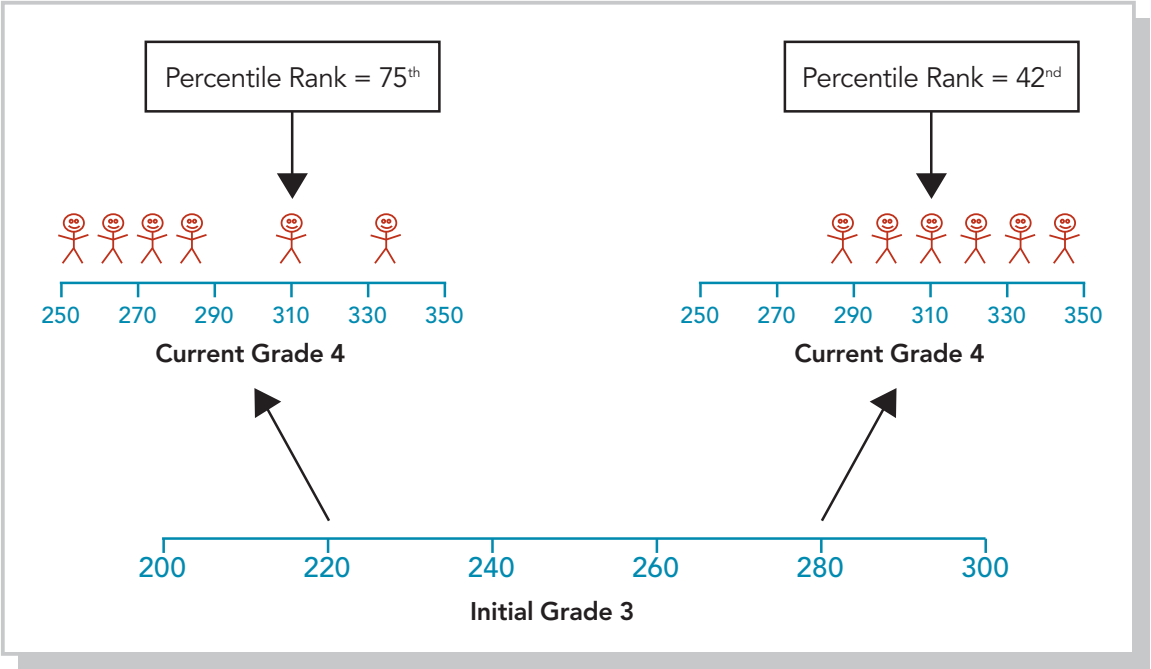
The SGP model is a conditional status model.

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

$$\text{Percentile Rank} = \frac{\text{Number of students at or below 310} + (.5 * \text{Number of students at 310})}{\text{Number of students in the academic peer group}} \times 100$$

$$= \frac{4 + (.5 * 1)}{6} \times 100$$

$$= \frac{4.5}{6} \times 100 = 75$$

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

$$\text{Percentile Rank} = \frac{\text{Number of students at or below 310} + (.5 * \text{Number of students at 310})}{\text{Number of students in the academic peer group}} \times 100$$

$$= \frac{2 + (.5 * 1)}{6} \times 100$$

$$= \frac{2.5}{6} \times 100 \approx 75$$

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.

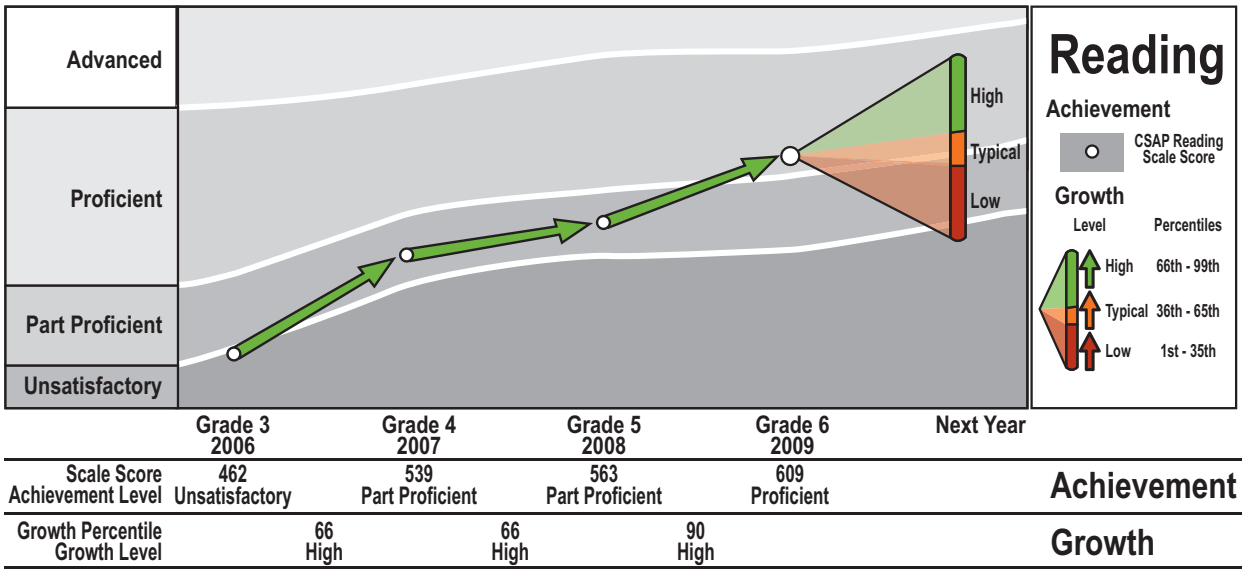
A higher-level standard setting approach arises from an extension of SGPs to support growth predictions. These "percentile growth trajectories" can support inferences about student trajectories toward a particular standard, such as Proficient or College and Career Ready. Percentile growth trajectories 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.

CHAPTER 7
The Multivariate Model

The multivariate model is designed for the primary purpose of supporting value-added inferences for teachers and schools. It supports answers to questions such as

How much better or worse did the students in a particular classroom perform when compared to expectations given

1) students' scores in other grades and subjects,

2) average district scores for each grade-subject combination, and

3) other teachers who are previously or currently teaching the same students?

The term "multivariate," meaning multiple variables, arises from the model's consideration of all student score variables, past and current, as a simultaneous target for modeling. Through this complex web of students moving through classrooms, schools, and school districts over time, statistical expectations for student performance are set. Higher or lower than expected performance can be directly related with students' particular teachers or schools, resulting in estimates for each teacher or school.

These estimates are often interpreted as causal effects — the teacher or school's direct contribution to average student performance. These inferences are generally difficult to support using model results alone. For simplicity, we will explain the underpinnings of the multivariate model using classrooms and their teachers as the target of inference. In many models, including the popular Educational Value-Added Assessment System (EVAAS) (Sanders & Horn, 1994) that we will use in this chapter as our prototypical multivariate model, these teacher associations are assumed to persist undiminished into the future. This persistence suggests that the student performance attributable to a student's third grade teacher persists into fourth grade, fifth

MULTIVARIATE MODEL

Aliases and Variants:
- Sanders Model
- EVAAS
- TVAAS/Tennessee Model
- Layered Model
- Variable Persistence Model
- Cross-Classified Model

Primary Interpretation:
Value-Added

Statistical Foundation:
Multivariate

Metric/Scale:
Usually a standardized (standard deviation unit) scale

Data: Generally no vertical scale is required; multiple years of data are recommended for teachers and students

Group-Level Statistic: Teacher "Value-Added"

Set Growth Standards:
Standards required to support absolute or relative distinctions among teacher/school effects, e.g., awards/sanctions to top/bottom 5%.

Operational Examples:
Ohio and Tennessee

grade, and so on. This is sometimes called a *layered model*, in a reference to the layering of estimated teacher "effects" onto a particular student over time. It is possible to relax this assumption using a "variable persistence" model (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004).

The EVAAS model sets expectations for any particular teacher's students by considering all of these students' scores, usually in a five-year window, both before and after the students enter and leave the teacher's classroom, and including all scores in other subjects. In addition, the district's average scores are factored into the expectation, as well as the teacher estimates from all of the students' other teachers over time. The EVAAS model and multivariate models in general are capable of incorporating other student-, teacher-, and school-level demographic or structural variables, although this is not done operationally (Ballou, Sanders, & Wright, 2004). The EVAAS model is complex, requires highly specialized and proprietary software, and is difficult to explain without reducing teacher estimates to a simplistic "value added" (causal) inference.

Question 7.1:
What *Primary Interpretation* Does the Multivariate Model Best Support?

The multivariate model supports value-added interpretations by expressing a teacher's students' performances in terms of their average distance from expectations. These expectations are set by considering students' other test scores, average district performance, and the other teachers that the students have had.

The primary outputs of interest from the EVAAS model are teacher-level, not student-level estimates. These estimates are found using equations for each grade and subject test that are connected through the covariance matrix, a summary of the interrelationships between test scores over grade levels. The multivariate model improves upon the covariate adjustment model (see Section 4.4), which also models "effects" for groups, by incorporating more information: over time, across subjects, and across other teachers. The intuition underlying the multivariate model is that a student's entire score history can be affected by membership in a particular teacher's classroom. As a heuristic device, imagine that we wish to estimate the added value associated with being in a particular classroom at a particular grade. We can take all the students who passed through that classroom and compare them to students like them, taking into account scores on other tests and the other teachers that they have had. Average differences between the score histories of students with this particular teacher and the score histories of other students can be described as a "teacher effect." This is only a heuristic that understates the complexity and assumptions of the multivariate model considerably, but it illustrates how

this model can support interpretations about the contribution of teachers to student test scores. The EVAAS model can be applied to multiple cohorts, and a more stable estimate for a teacher in a particular grade and subject area can be calculated by pooling teacher estimates from different cohorts together (Braun, 2005).

These estimates can support value-added interpretations. This assumes a causal attribution of the difference between actual and expected classroom performance to the particular teacher for that grade and subject. It is best to supplement these estimates with other sources of information when evaluating the teacher's effectiveness. For instance, the EVAAS model does not take into account the specific strategies and lesson plans that teachers utilize, preventing understanding of the mechanisms that might underlie added value (Braun, 2005). Although the EVAAS teacher estimates undergo a great deal of scrutiny and may have higher reliability than, say, classroom observations, triangulation of multiple sources of information is always desirable when making high-stakes decisions.

Question 7.2:
What is the *Statistical Foundation* Underlying the Multivariate Model?

As the name suggests, these models use a multivariate statistical foundation that allows for simultaneous consideration of many years of student scores as well as scores in other subjects.

From a more advanced statistical perspective, the gain-based and conditional status models are actually restrictive special cases of the multivariate model, which in its most unspecified form represents a useful unifying framework. From a practical perspective, and as the model is operationalized, the multivariate foundation is a stark contrast to the foundations underlying gain-based and conditional status models, which result in substantially more interpretable output. The advantages of the multivariate statistical foundation include the opportunistic use of data, not only over time but also across subjects and for students with missing data, to maximize information about the students in teachers' classrooms. The model is also flexible enough to allow for the layering of teacher estimates onto any given student's scores in a way that simpler models cannot. Alternative forms of the model can include an estimate of the fading out of teacher associations over time in what is known as a *variable persistence model* (McCaffrey, et al., 2004).

To help visualize the mechanics of the EVAAS model, the following layering of equations demonstrates how each student's grade-level score is decomposed for the simplest case of a single school system, a single subject, and a single cohort of students with Grade 3 to Grade 6 scores:

Student *i*'s Grade 3 Score = Average Grade 3 Score + Grade 3 Teacher Estimate
+ Individual Student Error for Grade 3

Student *i*'s Grade 4 Score = Average Grade 4 Score + Grade 3 Teacher Estimate
+ Grade 4 Teacher Estimate
+ Individual Student Error for Grade 4

Student *i*'s Grade 5 Score = Average Grade 5 Score + Grade 3 Teacher Estimate
+ Grade 4 Teacher Estimate + Grade 5 Teacher Estimate
+ Individual Student Error for Grade 5

Student *i*'s Grade 6 Score = Average Grade 6 Score + Grade 3 Teacher Estimate
+ Grade 4 Teacher Estimate + Grade 5 Teacher Estimate
+ Grade 6 Teacher Estimate
+ Individual Student Error for Grade 6

These equations demonstrate the persistence of a teacher's estimate into each subsequent grade-level—that is, the Grade 3 teacher estimate is carried over to Grades 4, 5, and 6, and, similarly, the Grade 4 teacher's estimate is carried over to Grades 5 and 6, and so on. A variable persistence model would allow the magnitude of a prior grade-level teacher's estimate to decrease over time.

It is not easy to deduce from the above equations precisely how the teacher estimates are estimated. A detailed explanation of this statistical model is beyond the scope of this chapter. However, it is useful to note that in the first grade-level included in the model, the teacher estimate is not adjusted for any prior performance or other "historical factors," such as demographic or economic variables. Thus, these historical factors are confounded with the Grade 3 teacher estimate and should therefore be interpreted cautiously (McCaffrey, et al., 2004).

Disadvantages to this statistical approach include a lack of parsimony and clarity in model interpretation. Gain-based models align with intuitive notions of growth over time. Conditional status models align less well to intuitive conceptions of growth, but it is not difficult to imagine an expected score empirically determined from past scores and a referencing of actual performance to expected performance. The conditional interpretation from the multivariate model is aggregated to the level of teachers or schools, and the expectation is based on 1) not only past scores but future scores after students leave a teacher's class, 2) not only same-subject scores but all available scores, and 3) a layering of other teacher associations from all teachers who have ever had each student in their class. Although it is easy to casually abstract these scores to "value added," the more rigorous interpretation considers the variables that set the expectations, and these variables are numerous with complex interrelationships.

Question 7.3:

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

The multivariate model is very flexible in terms of the data it can utilize. Generally, it can accommodate a large amount of test score data from multiple grade-levels and subjects. Moreover, as this model is primarily for producing group-level estimates (e.g., for teachers or schools), students not only need to have unique identifiers but also identifiers for all of their teachers, schools, and districts over time so that these associations can be tracked in the model.

Without the need to report student-level growth results, the sample sizes of interest pertain to the number of test scores for students in each teacher's classroom over time. The efficiency of the model in using available data usually results in a substantial improvement over covariate-adjustment models, although this can also sacrifice interpretability of model results. A vertical scale is not required for most uses of the multivariate model, but standard deviation units are assumed to hold consistent meaning across grades and subjects. Due to the assumption of persistent teacher effects, their magnitudes, expressed in standard deviation units, are assumed to stay constant across the test score scales of different grades and subjects.

Question 7.4:

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

The multivariate model is designed for group-level interpretations, particularly at the classroom level, although school and district level interpretations are also possible through minor reconfigurations of the model.

Generally, teacher or school estimates from the EVAAS model are most appropriate for identifying teachers who may benefit from additional professional development and for identifying schools for further investigation as they may be underperforming. In these cases, the group-level estimates serve as a screening tool that selects teachers or schools that may need additional resources (Braun, 2005). Value-added interpretations of the group-level estimates should be triangulated with other sources of information, such as teacher portfolios and classroom observations. Given that the entire focus of this chapter is on group-level interpretations, we do not expand on this topic further here.

Question 7.5:

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

The "value-added" scores are most often interpretable in terms of standard deviation units with respect to a baseline average centered on zero. Relative comparisons of value-added scores are possible, such as flagging a certain top and bottom proportion for further investigation.

The multivariate model results in a distribution of educator or school estimates. These are not interpretable on an absolute scale and must be interpreted normatively. Standards may be set by selecting a top or bottom proportion or identifying a number of standard deviation units away from a reference point. Additionally, statistical significance tests can be conducted to support inferences about an educator's estimate being higher or lower than a particular target cut score to a degree of statistical significance.

Question 7.6:

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

The interpretation of value-added scores as actual value that a teacher has added is an example of a naming fallacy — naming a metric "value added" does not necessarily make it so.

Ascribing causal effects to teachers is generally not warranted by educational data designs. It is more precisely a deviation from expectations associated with the class of students, where the expectation is set by student scores and students' past and future teachers from other classrooms. This more disciplined interpretation can allow for an interpretation of the "teacher effect" in context and a deeper exploration of plausible alternative explanations for high or low scores. Moreover, some studies have found that the most extreme ranks — those at the very top and bottom — are unreliable (Lockwood, Thomas, & McCaffrey, 2002), which could have substantial implications for high-stakes decisions focused on the very top and bottom ranked teachers. In addition, often only a small fraction of teachers, 33 percent or less, are found to be reliably different from the average teacher in a district (Braun, 2005).

Like conditional status models, multivariate models do not allow for intuitive growth interpretations but instead represent an enhancement of status interpretations by incorporating a reference point, an expectation based on other information. Like incentives for gain-based models, a teacher is incentivized to maximize the scores of the students in his or her class. The teacher also benefits if the scores of his or her students are artificially deflated in every other classroom except that teacher's own.

References

Ballou, D., Sanders, W., and Wright, P. (2004). Controlling for student background in value-added assessment for teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65, from http://web.missouri.edu/~podgurskym/Econ_4345/syl_articles/ballou_sanders_value_added_JEBS.pdf. doi: 10.3102/10769986029001037

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service, from http://www.ets.org/Media/Research/pdf/PICVAM.pdf.

Lockwood, J.R., Thomas A.L., and McCaffrey, D.F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics, 27*(3): 255-270.

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A., and Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1): 67-101.

Sanders, W.L., and Horn, S.P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*(3): 299-311.

APPENDIX A
CROSS-REFERENCING GROWTH MODEL TERMS

Growth model classification systems, like growth models themselves, serve multiple purposes. Two documents associated with this guide deserve special attention for their growth model classification systems, the CCSSO *Growth Model Comparison Study* (Goldschmidt, Choi, & Beaudoin, 2012)) and the CCSSO *Understanding and Using Achievement Growth Data* brochure (Council of Chief State School Officers, 2011). CCSSO's growth brochure was intended as a concise review of growth model principles, and the *Growth Model Comparison Study* is more empirical, more technical, and focuses primarily on school-level accountability metrics. In contrast, *The Practitioner's Guide to Growth Models* represents a middle ground, an in-depth overview of the growth model landscape. The distinct purposes of these three documents lead to different growth model classifications. This appendix summarizes the contrasting growth classification schemes.

The CCSSO brochure identified five basic types of growth models: Categorical, Gain-Score, Regression, Value-Added, and Normative. These five growth model types are listed and related to this guide's terminology in Table A.1 below. For instance, this guide also reviews Categorical and Gain-Score models but emphasizes that the Categorical model is a type of gain-based model that creates an implicit vertical scale. This is elaborated fully in Chapter 3 on the Categorical model.

The *Practitioner's Guide* treats the Regression model as a statistical approach that underlies many models. Regression is essential for all models that use the conditional status statistical foundation, from Projection Models to Student Growth Percentiles. Regression, as a statistical technique, also supports Multivariate models. Although a Regression model refers in practice to Projection models for growth prediction and Covariate Adjustment models for value-added inferences, this guide uses "regression" in reference to the statistical technique.

Finally, this guide uses "normative" to refer to the referencing of scores to a norm group, that is, a reporting technique, and not a particular model. Although Student Growth Percentiles report scores on a norm-referenced metric, other growth models are also capable of reporting different conceptions of growth in a norm-referenced fashion.

Table A.2 below presents the 9 growth models reviewed in the *Growth Model Comparison Study*. One of the uses of this guide is to help to contextualize and explain the observed differences between growth models when they are applied to real data. An important conceptual distinction between the *Practitioner's Guide* and the *Growth Model Comparison Study* is that the latter focuses on a single purpose, a "value-added" type of ranking, at a single level of aggregation — the school level. In contrast, this guide includes multiple purposes, including growth description and growth prediction, and multiple levels, including the student, teacher, and school levels.

Table A.1

Mapping Growth Model Terminology from CCSSO's *Understanding and Using Achievement Growth Data* to those in this *Practitioner's Guide*

Understanding and Using Achievement Growth Data[8]		***Practitioner's Guide to Growth Models***
Categorical	→	Categorical Model and Type of Gain-Based Model
Gain-Score	→	Gain Score Model and Type of Gain-based Model
Regression	→	A statistical approach that supports many models, Residual Gain, Projection, Student Growth Percentiles, Covariate Adjustment, and Multivariate
Value-Added	→	A purpose associated with many models, particularly Covariate Adjustment and Multivariate Models
Normative	→	A reporting metric associated particularly with Student Growth Percentiles, but more broadly applicable

Table A.2

Mapping Growth Model Terminology from the CCSSO *Growth Model Comparison Study* to those in this *Practitioner's Guide*

Growth Model Comparison Study[9]		***Practitioner's Guide to Growth Models***
Simple Gain	→	Gain Score Model and Type of Gain-Based Model
Fixed Effects Gain	→	Type of Gain-Based Model
True Score Gain	→	Type of Multivariate Model
Covariate Adjustment with School Fixed Effects	→	Covariate Adjustment Model
Covariate Adjustment with School Random Effects	→	Covariate Adjustment Model
Simple Panel Growth	→	Type of Multivariate Model
Layered Model	→	Type of Multivariate Model
Student Growth Percentile	→	Student Growth Percentiles (in the Student Growth Percentile Model)
Growth to Standards	→	Trajectory Model

[8] See CCSSO (2011).

[9] See Goldschmidt et al. (2012).

BIBLIOGRAPHY

Auty, W., Bielawski, P., Deeter, T., Hirata, G., Hovanetz-Lassila, C., Rheim, J., Goldschmidt, P., O'Malley, K., Blank, R., and Williams, A. (2008). *Implementer's guide to growth models.* Washington, DC: Council of Chief State School Officers.

Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R., Ravitch, D., Rothstein, R., Shavelson, R.J., Shepard, L.A. (2010, August). *Problems with the use of student test scores to evaluate teachers* (Economic Policy Institute Briefing Paper #278). Retrieved March 29, 2012, from http://epi.3cdn.net/724cd9a1eb91c40ff0_hwm6iij90.pdf.

Ballou, D., Sanders, W., and Wright, P. (2004). Controlling for student background in value-added assessment for teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65. Retrieved March 30, 2012, from http://www.epi.org/publication/bp278/.

Beimers, J. (2008). *The effects of model choice and subgroup on decisions in accountability systems based on student growth.* Ph.D. dissertation, University of Iowa.

Betebenner, D.W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51.

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. Retrieved March 29, 2012, 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. Retrieved March 29, 2012, from http://ccsso.confex.com/ccsso/2011/webprogram/Session2199.html.

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service. Retrieved March 30, 2012, from http://www.ets.org/Media/Research/pdf/PICVAM.pdf.

Buzick, H.M., and Laitusis, C.C. (2010). *A summary of models and standards-based applications for grade-to-grade growth on statewide assessments and implications for students with disabilities* (Educational Testing Service TS RR-10-14). Princeton, NJ: ETS. Retrieved March 29, 2012, from http://www.ets.org/Media/Research/pdf/RR-10-14.pdf.

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.

Council of Chief State School Officers.(CCSSO). (2011). Understanding and using achievement growth data. *Growth Model Brochure Series*. Retrieved September 19, 2012, from http://www.wera-web.org/links/Journal/June_Journal_2012/CC6_CCSSO_Growth_Brochures_jan2012.pdf.

Delaware Department of Education. (2010). For the 2009-2010 school year: State accountability in Delaware. Retrieved on April 29, 2012, from http://www.doe.k12.de.us/aab/accountability/Accountability_Files/School_Acct_2009-2010.pdf.

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. Retrieved on March 29, 2012, from http://www.nciea.org/publications/MeasuringGrowthMCASTests_CD06.pdf.

Dunn, J.L., and Allen, J. (2009). Holding schools accountable for the growth of nonproficient students: Coordinating measurement and accountability. *Educational Measurement: Issues and Practice, 28*(4), 27-41.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., and Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: Brown Center on Education Policy at Brookings. Retrieved March 29, 2012, from http://www.brookings.edu/~/media/Files/rc/reports/2010/1117_evaluating_teachers/1117_evaluating_teachers.pdf.

Goldschmidt, P., Choi, K., and Beaudoin, J.P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance*. Washington, DC: Council of Chief State School Officers.

Ho, A.D. (2011). *Supporting growth interpretations using through-course assessments*. Austin, TX: Center for K–12 Assessment & Performance Management, ETS. Retrieved April 21, 2012, from http://www.k12center.org/rsc/pdf/TCSA_Symposium_Final_Paper_Ho.pdf.

Ho, A.D., Lewis, D.M., and Farris, J.L.M. (2009). The dependence of growth-model results on proficiency cut scores. *Educational Measurement: Issues and Practice, 28*(5): 15-26.

Hoffer, T.B., Hedberg, E.C., Brown, K.L., Halverson, M.L., Reid-Brossard, P., Ho, A.D., and Furgol, K. (2011). *Final report on the evaluation of the growth model pilot project*. Washington, DC: U.S. Department of Education. Retrieved April 27, 2012, from http://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/index.html.

Kolen, M.J., and Robert L. Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer Science+Business Media, Inc.

Lockwood, J.R., Thomas A.L., and McCaffrey, D.F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics, 27*(3): 255-270.

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.

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A., and Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1): 67-101.

O'Malley, K.J., Murphy, S., McClarty, K.L., Murphy, D., and McBride, Y. (2011). *Overview of student growth models* (Pearson White Paper). Retrieved March 29, 2012, from http://www.pearsonassessments.com/hai/Images/tmrs/Student_Growth_WP_083111_FINAL.pdf.

Reardon, S.F., and Raudenbush, S.W. (2009). Assumptions of value-added models for estimating school effects. *Educational Finance and Policy, 4*(4): 492-519.

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.

Rubin, D.B., Stuart, E.A., and Zanutto, E.L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29(1)*: 103-116.

Sanders, W.L., and Horn, S.P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*(3): 299-311.

Sanders, W.L. (2006). *Comparisons among various educational assessment value-added models*. SAS white paper presented at The Power of Two—National Value-Added Conference, Columbus, OH, on October 16, 2006. Retrieved March 29, 2012, from http://www.sas.com/resources/asset/vaconferencepaper.pdf.

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

U.S. Department of Education. (2005). *Secretary Spellings announces growth model pilot, addresses chief state school officers' Annual Policy Forum in Richmond* (Press Release dated November 18, 2005). Retrieved March 29, 2012, from http://www2.ed.gov/news/pressreleases/2005/11/11182005.html.

Weiss, M. (2008). *Using a yardstick to measure a meter: Growth, projection, and value-added models in the context of school accountability*. Ph.D. dissertation, University of Pennsylvania.

ABOUT THE AUTHORS

Katherine E. Castellano, Ph.D.
University of California, Berkeley

Katherine E. Castellano is an Institute of Education Sciences postdoctoral fellow at the University of California, Berkeley. Dr. Castellano manages operational assessment projects for the Berkeley Evaluation and Assessment Research Center and conducts research related to multilevel models and student growth models. She is particularly interested in the properties of student growth percentiles at the individual- and aggregate-level and how this popular metric compares with other regression-based approaches. Dr. Castellano has consulted for the National Opinion Research Center on the *Final Report on the Evaluation of the Growth Model Pilot Project* and has interned at Westat, the United States Department of Energy, the National Center for Education Statistics, and the Educational Testing Service (ETS). She received the 2010 ETS Harold Gulliksen dissertation fellowship and was awarded the L.B. Sims Outstanding Master's Thesis award in 2008. She earned her Ph.D. in educational measurement and statistics from the University of Iowa, where she also completed her master's degree in statistics.

Andrew D. Ho, Ph.D.
Harvard Graduate School of Education

Andrew Ho is an Assistant Professor at the Harvard Graduate School of Education. He is a psychometrician interested in educational accountability metrics, an intersection of educational statistics and educational policies. He has studied the consequences of proficiency-based accountability metrics, the validation of high-stakes test score trends with low-stakes comparisons, and the potential for alternative accountability structures — such as "growth model" and "index systems"— to improve school- and classroom-level incentives. He has his Ph.D. in Educational Psychology and his master's degree in Statistics from Stanford University. Dr. Ho has been a postdoctoral fellow at the National Academy of Education and the Spencer Foundation. He is also a recipient of the Jason Millman Promising Measurement Scholar Award from the National Council on Measurement in Education.