Growth as a Description of Process

Damian W. Betebenner
Lynch School of Education
Department of Educational Research, Measurement, and Evaluation
Boston College
Damian.Betebenner@bc.edu

March 29, 2006

Abstract

At present, given the flood of annual student data emanating from state assessment programs, there is great enthusiasm for statistical models suitable for the analysis of longitudinal data. The Department of Education's solicitation of state growth models is indicative of this enthusiasm. Most of the research on the analysis of growth, particularly with regard to the family of "value-added" growth models, has focused on using the observational longitudinal data to estimate causal effects associated with teachers, schools, etc. Though the estimation of the contribution of different facets of the education system to the overall outcome of a child's academic ability is certainly worthy of study, the purpose of this paper is more modest: Setting issues of causal attribution to the side, this paper describes how growth can be employed to better understand process. To this end, a Markov Transition Model is employed, utilizing performance levels instead of scale scores, to analyze student growth.

Background

Today, students in the United States participate in large scale assessment to an extent never before seen. Numerous factors have contributed to this recent explosion in testing, but a major impetus has been the reauthorization of the *Elementary and Secondary Education Act of 1965*, commonly known as the *No Child Left Behind Act* (NCLB). Beginning with the 2005-2006 school year, NCLB requires states to test students in reading and mathematics from grades 3 though 8 and at least once in grades 10 through 12. By the 2007-2008 school year, states must also assess students in science achievement at least once in elementary, middle, and high school. Beyond the student assessment mandates, NCLB requires all states and districts receiving Title I assistance to produce report cards for each of their public schools that provide, among other things, information on how well students at the school have performed on state assessments.

Of particular relevance, Section 1111 of NCLB requires that the state's performance levels/standards be used for the school report cards. Specifically, for each school, student achievement at each performance level must be reported for all students as well as disaggregated by race, ethnicity, gender, disability status, migrant status, English proficiency and economically disadvantaged status. NCLB's goal is for universal proficiency of all students by 2014 in both reading and math. Pursuant to this, each state crafts a plan incorporating intermediate performance targets leading toward the 2014 universal proficiency goal. Schools which persistently fail to meet these performance targets will be designated as "in need of improvement" and are subject to corrective action or restructuring.

The goal of having all children reach a high level of academic proficiency is unimpeachable. However, numerous critics argue that such a goal is simply not attainable (Popham, 2004). Those sharing these sentiments suggest that alternate metrics of school quality need to be employed in order to better represent the performance of schools and determine adequate yearly progress (AYP) (Linn, 2003a). An approach that currently finds favor is to incorporate student growth into measures of AYP (Kingsbury, Olson, McCahon, & McCall, 2004; Hill, 2003; Thum, 2003). Along these lines, Linn (2004) has suggested a combination of longitudinal student analysis with an annual NCLB-like status measure as an attractive way to satisfy AYP.

The call to incorporate some form of student growth into NCLB should be reviewed carefully. In addition to questions concerning causal imputations made using the results (Rubin, Stuart, & Zanutto, 2004), there is a fundamental incongruence between NCLB and many of the longitudinal value-added measures currently in use. Whereas NCLB is constructed using performance levels, most value-added models are *normative* in nature. The Tennessee Value-Added Assessment System (TVAAS), for example, uses random effects' BLUPs to compare teachers or schools to one another within a given education agency (usually a district). There is a great deal of research being done to validate the results of such models (Ballou, Sanders, & Wright, 2004; Sanders, Saxton, & Horn, 1997; Raudenbush & Willms, 1995). However, it is far from clear that even if the models yield valid results, that they can be meaningfully incorporated into an accountability system articulated in terms of performance standards.

The purpose of this paper is to present a "third way" between NCLB performance mandates and the value-added growth models that currently find favor. This third way involves connecting the demands of accountability systems like NCLB with a model of student growth whose focus is on process. Undoubtedly, NCLB based school improvement measures and growth estimates derived by modeling student change each provide unique information about the status of the education system over time. However, finding concordance between these two sources of information is not trivial. The difficulty is largely alleviated if models are designed using the assessment's ordinal performance categories. As will be demonstrated, statistical models using these categories yield results in the metric of performance levels, the *lingua franca* for instruction, state assessment, and NCLB.

A Taxonomy of Different Measures

A great deal of confusion exists, even among experts, concerning growth. Much of this confusion resides in imprecise manner in which the term "growth" is employed. Most often, the term growth is used to connote student change over time. However, growth is often associated with school performance over time. The NCLB accountability system, for example, mandates increasing levels of achievement at the school level (i.e., school level growth) between now and 2014.[1] Growth/change associated with different units are not synonymous. A useful semantic distinction is to limit the use of the term *growth* to discussions of student change and employ the term *improvement* when discussing change at the institutional level.

Dale Carlson (2001) has suggested an amazingly insightful and simple 4 cell table, reproduced in Table 1, summarizing the different ways one can categorize the disparate measures, and associated questions, currently utilized to assess the quality of a school. All measures of school quality, as measured by assessment outcomes, currently in use today can be fit into one of these cells (or in some cases, into a combination of cells). The rows of the table mark two distinct qualities associated with judging the merit of schools: achievement and effectiveness. Neither quality is a necessary nor

[1] Throughout this paper school is used as the unit of analysis. However, the discussion is equally applicable to other levels including teacher, district, and state.

	How Good is this School?	Is it Getting Better?
Achievement	**A1:** What is the achievement level of students in this school? **Examples:** Percent proficient students, mean scale score, composite performance index.	**A2:** Is the achievement level of this school increasing? **Example:** Percent NCLB proficient over time.
Effectiveness	**E1:** Is this an effective school? That is, given the achievement level of students when they enter, how much do they learn or develop while they are in the school? **Examples:** Value-added residuals, transition matrices.	**E2:** Is this school becoming *more* effective? How much more, or less, are the students learning this year than they did the year before? **Examples:** None in widespread use (to my knowledge).

Table 1: Carlson's Taxonomy (2001) represents the four facets associated with measures of school quality

a sufficient condition for the other.[2]

- Cell A1 of Table 1 represents the achievement level (i.e., status) of a school at a given time.
- Cell A2 represents change in achievement level over time. This, for example, is the concern of NCLB: Is the percent of children who are NCLB proficient increasing over time towards the goal of 100 percent by 2014?
- Cell E1 is where much current attention is focused in terms of measuring student growth—A school is deemed effective if the students attending the school, no matter where they start, show significant improvement.
- Cell E2 represents one of the primary goals of education reform: the improvement of efficacy over time.

The columns Table 1 indicate the time frame through which the school is examined. Column 1 addresses school quality in the present while column 2 concerns school quality over time. That is, column 1 addresses likely concerns of most interest to parents while issues associated with column 2 are of most interest to policy makers and researchers interested in the education performance over time. This distinction is relevant with regard to school report cards, who they are written for, and what they should include. Table 1 also indicates how measures of growth associated with a school's students (Cell E1) are address different qualities than measures of improvement associated with a school (Cell A2). Measures of student growth utilize prior student information in an attempt to assess the quality/effectiveness of the education process the students were exposed to.[3]

[2] A measures inclusion in a cell merely suggests its intended purpose. It does not imply that all the measures are *equally* valid.

[3] The use of prior information/covariates to assess effect subtly weaves a causal web that is not well grounded. Lord's Paradox (1967) and its resolution by Holland and Rubin (1983) demonstrate the limits of using prior data to attribute and quantify causal effects.

Improvement in achievement for a school, however, is a cross-sectional examination of performance over multiple years.

Measures/issues of greatest relevance to education reform are situated in Cell E2. Because increases/decreases in effectiveness represent a change of a change, measuring such qualities is difficult with any degree of accuracy. NCLB's increasing levels of achievement over time (Cell A2) can be considered proxy measures of increasing effectiveness by increasing the percentages of students reaching proficient. Alternatively, instead of using the terms achievement and effectiveness to characterize different measures of educational quality, another way of conceptualizing the rows of Table 1 are in terms of *output* and *process*. Current achievement measures, such as those represented in Cells A1 and A2, are measures based upon outputs—what occurs as the result of education (and numerous other factors). If outputs increase over time (Cell A2), then one can likely infer that there is increasing effectiveness (Cell E2). The attractiveness of using the growth of students to evaluate the education system is that growth, as a measure, is consistent with the notion of *education as a process*. That is, the more effective the educational process, the better the growth.

A central thesis of this paper is that, given the model developed in the following section, it is possible to model the relationship between all four types of measures depicted in Table 1. The next section introduces a basic Markov transition model applicable to measuring student growth. This model produces an elementary relationship between growth and achievement. Because the model uses student performance level as the dependent variable, linking growth with achievement yields results based upon percentage of students in each of the performance levels. An outcome of particular relevance to many accountability systems. In line with this, the purpose of this paper is to provide better "descriptive measures" (Rubin et al., 2004, p. 113) of growth that will allow informed discussions to emerge from the mass of data.

Models for Longitudinal Data Analysis with Ordinal Outcomes

Diggle and Heagerty (2002) identify three approaches to model longitudinal data: marginal models, mixed-effects growth models, and Markov chain models.. The choice of model used to analyze longitudinal data largely depends upon the questions asked as well as the manner in which dependencies between observations are to be modeled. The most prevalent class of techniques currently used with longitudinal education data are mixed-effects models. Mixed-effects models have been extended to situations with ordered category responses under the guise of generalized linear mixed models or non-linear mixed-effects models (Hedeker, 2003; Pinheiro & Bates, 2000). When data is longitudinal, techniques used to analyze longitudinal data based upon continuous outcome measures have analogs when ordinal outcomes are of interest. Moreover, there is recent use of these techniques involving education data (Barbosa & Goldstein, 2000; Fielding, 1999). A purpose of this study is to add to the emerging body of literature regarding the use of ordinal outcomes in longitudinal growth models, through the articulation of a Markov model applicable to the analysis of change in education.

A criticism of using performance levels instead of scale scores is that performance levels lack the "nuance" that scale scores provide. Consequently, analyses using scale scores will provide a more granular analysis than those employing performance levels. It is certainly the case that a quasi-interval developmental scale is finer than a categorical ordinal scale, however:

1. In many instances, especially at scale score extremes, the precision afforded by scale scores is illusory. Measurement error is a significant issue that is currently not accounted for by current longitudinal analyses, including value-added models (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004).

2. Even with precision, the scale's arbitrary nature makes it difficult to communicate results and make them "actionable" by teachers, principals, and other stakeholders: A gain of 10 scale score points, even if precisely measured is difficult to interpret, especially if it occurs at different places on the scale.

Analyses using the ordinal outcomes represented by performance levels are a superior metric to scale score based analyses for a number of reasons. Foremost is that performance standards, when properly integrated with assessment, create an isomorphism between test performance and content mastery. For example, the list of concepts and tasks is available for a child designated proficient in 4th grade math. Moreover, by looking at the adjacent performance categories (e.g., below proficient and above proficient/exceptional), a refined understanding emerges of what the child is capable of doing well and what has still not been mastered. Over time, given a set of vertically moderated performance categories (Huynh & Schneider, 2004; Lissitz & Huynh, 2003), movement between these categories translates into content mastery growth. As results are aggregated upwards (e.g., to the school, district, and state level), gross movement between performance categories indicates what content is being successfully taught to students. Thus, an accountability system incorporating growth as measured by movement between performance levels provides a means of articulating change within the relevant framework of performance standards. The Markov transition model presented in this paper is a simple technique that suits this end.

Analysis techniques using longitudinal data with ordinal outcomes are not as well known as those utilizing continuous outcomes. That is not to suggest that there is not excellent research that does exist. A century ago Andrei Markov (1906) investigated a class of dependent discrete random processes which now bear his name. Markov models are attractive for the modeling of growth in ordinal education outcomes for a number of reasons:

1. The models are well suited for modeling growth using the ordinal performance standards present in state assessment systems.

2. The models are conceptually simple but flexible enough to adapt to complex situations such as those found in education.

3. The models have a long history and are well studied. Thus a large body of mathematical theory is immediately available for use.

4. The results of the models dovetail with current accountability mandates like those found in NCLB.

The remainder of this paper develops Markov models appropriate for the analysis of longitudinal data in schools and investigates their use with regard to standards based accountability systems.

Markov Models

The simplest Markov chains refer to a collection of random variables, $\{X_n\}$, $n \in \mathbb{N}$, called *states* that take on values in a set $E = \{E_1, E_2, \ldots, E_k\}$, called the *state space*, for which the probability of event X occurring at time $n+1$, $\Pr(X_{n+1})$, is dependent only on the what occurs at time n.[4] That is, for current results probabilistic dependence on the past depends only upon the most recent

[4] In the education context developed in this paper, E consists of the performance levels reported for an assessment.

past (Gnedenko, 1967; Brémaud, 1999). Formally, if X_n, $n \in \mathbb{N}$, denotes a sequence of discrete random variates, then

$$\Pr(X_{n+1}|X_n, X_{n-1}, \dots X_1) = \Pr(X_{n+1}|X_n). \tag{1}$$

Equation 1 is referred to as the Markov property. A Markov chain depending upon only its most recent observations is said to have order 1. In general, a Markov chain depending upon its j most recent observations is said to have order j. That is

$$\Pr(X_{n+1}|X_n, X_{n-1}, \dots X_1) = \Pr(X_{n+1}|X_n, X_{n-1}, \dots X_{n-j+1}).$$

This paper give details relevant to the use of first order Markov chains as a means of analyzing student growth.

For the present discussion, a Markov chain whose state space, E, consists of four performance levels (PL1, PL2, PL3 and PL4) is formulated. State spaces with fewer or greater number of levels generalize easily from the present discussion. Students in an education system with annual assessments are rated using the performance levels. As students proceed from grade to grade, a Markov chain is realized. Given two years of data, unbiased transition probability values (i.e., conditional probabilities) between time n and $n+1$ are easily produced using empirical frequency crosstabulation of the performance categories at time n against those at time $n+1$.

With state space $E = \{\mathsf{PL1}, \mathsf{PL2}, \mathsf{PL3}, \mathsf{PL4}\}$, the crosstabulation yields a 4×4 matrix, $\mathbf{P}$, called the *transition matrix*. Formally, $\mathbf{P} = \{p_{ij}\}_{i,j \in E}$, such that

$$p_{ij} = \Pr(X_{n+1} = j | X_n = i).$$

For example, p_{23} denotes the probability of a student transitioning from performance level 2 at time n to performance level 3 at time $n+1$. It is a simple matter to confirm that for each row of $\mathbf{P}$, the row entries must sum to 1. Matrices with this property, called the *stochastic property*, are called *stochastic matrices* or, more commonly, *transition matrices*.

Transitions between performance categories are of primary interest in the examination of growth using ordinal outcomes. Transition matrices encode the likelihood of the transitions and thus form the basis of our discussion of Markov chains and their use in investigating growth based upon performance categories. Using two years of state data derived from approximately 50,000 students, the empirical (or observed) transition matrix $\mathbf{P}$ between grades 3 and 4 in reading is given by

$$\mathbf{P} = \begin{array}{c} \\ \mathsf{PL1} \\ \mathsf{PL2} \\ \mathsf{PL3} \\ \mathsf{PL4} \end{array} \begin{array}{c} \begin{array}{cccc} \mathsf{PL1} & \mathsf{PL2} & \mathsf{PL3} & \mathsf{PL4} \end{array} \\ \begin{pmatrix} .66 & .34 & .00 & .00 \\ .10 & .59 & .31 & .00 \\ .00 & .09 & .82 & .09 \\ .00 & .00 & .46 & .54 \end{pmatrix} \end{array} \tag{2}$$

Here, rows represent the performance level in grade 3 and columns represent the performance level in grade 4. For example, note that Equation 2 depicts no chance of transitions between categories that are not adjacent or identical. For example, the probability of a student moving from PL3 to PL1, given by p_{31}, is 0. To simplify the discussion, this paper assumes that student movement across more than one performance category are so rare as to effectively have probability zero. Such movement yields transition matrices, called *tri-diagonal* matrices, with non-zero entries on only the sub-diagonal, super-diagonal, and the diagonal itself. All results, however, are applicable to other transition matrices as well.

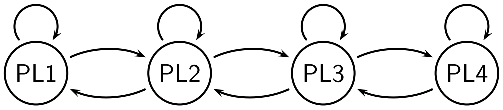

Figure 1: Markov diagram depicting transitions from time 1 to time 2 between performance levels

Markov diagrams are often employed to clarify the transition model under consideration. The Markov model that will be the basis for much of the subsequent analysis in this discussion is provided in Figure 1. Each arrow in Figure 1 depicts a transition from a performance level at time n to another performance level at time $n+1$. The probability of those transitions are given in the transition matrix $\mathbf{P}$. Notice that connections only exist between identical or adjacent performance levels (PL1, PL2, PL3 and PL4). That is, movement between levels two or more categories apart is assumed to have zero probability. This results in a tri-diagonal transition matrix that also reflects the empirical data presented in matrix $\mathbf{P}$. In addition, the model depiction is greatly simplified by not having two sets of arrows connecting all nodes to each other.

The most important feature of the model is that it permits the simple calculation of the state of the system at any time n. If ν_0 is the vector depicting the initial state of the system (i.e., the proportion of students in each of the four performance levels), then the subsequent state of the system, ν_1, is computed as

$$\nu_1 = \nu_0^T \mathbf{P}.$$

Where ν_0^T denotes the transpose of ν_0. Matrix multiplication allows the state of the system at time n, ν_n to be similarly computed:

$$\nu_n = \nu_{n-1}^T \mathbf{P_n} = (\nu_{n-2}^T \mathbf{P_{n-1}})\mathbf{P_n} = \cdots = \nu_0^T \prod_{i=1}^{n} \mathbf{P_i}. \tag{3}$$

Where $\mathbf{P_i}$ denotes the transition matrix between time $i-1$ and i.

In instances where the state space E contains four performance levels, ν_0 is a 4-tuple where each coordinate represents the proportion of students in the given performance category. ν_1 is thus a 4-tuple giving the proportion of students in the given performance category based upon the transition probabilities given in $\mathbf{P}$. The properties of the Markov chain yields two immediate benefits: Investigating different end results using various transition probabilities is accomplished easily through matrix composition and given pre-specified end results, it's a simple matter to investigate what transition probabilities are necessary so that the end results are achieved. The last benefit is of particular relevance since NCLB stipulates what the end state of the system will be in 2014. In the next section, the basic Markov chain tools to investigate NCLB and its mandates are applied.

Modeling NCLB Using Markov Chains

Among the many NCLB requirements, perhaps its most daunting is the 2014 universal proficiency goal for all children in grades 3 to 10. A number of authors have challenged the targets of NCLB as too ambitious (Popham, 2004; Linn, 2004, 2003a) and have suggested modifying NCLB with more reasonable expectations for improvement. The Markov transition model provides an ideal framework in which to investigate both what is necessary to meet the requirements of NCLB as well as what is possible based upon current results. That is, Markov chains, which employ the ordinal outcomes of performance standards, form a bridge linking targets based upon performance standards to longitudinal growth estimates. This section presents both analytic and empirical results for Markov chains with regard to NCLB. All calculations were performed using the R software package (R Development Core Team, 2006).

To investigate how Markov transition models can be employed to investigate NCLB performance mandates, recall Carlson's Taxonomy of measures associated with school quality given in Table 1. Cell A1 refers to levels of achievement. Using performance levels as the measure of achievement, percentages of students at the various performance levels provides the outcome of interest. Disaggregated by grade, the upper-left panel of Figure 2 shows column vectors depicting the proportion of students in each of four performance levels in a given grade in a given year. The upper entry of the column vector (color coded blue) refers to the proportion of students in the highest performance level, PL1, while the value of the bottom entry (color coded red) gives the proportion in the lowest level, PL4.[5] Percentages represented by these vectors are reported (perhaps aggregated across grades) on school report cards.

Cell A2 refers to changes in achievement over time. The upper-right panel of Figure 2 depicts how changes in achievement over time implies an examination of the percentages of the students in the different performance levels over time. This is indicated in the figure using light green rectangles to indicate which sets of quantities are being compared. This comparison represents the current focus of NCLB AYP mandates: The goal for schools is to have increasing levels of achievement leading toward universal proficiency by 2014. Note that the comparison of levels of achievement over time is done for *different* cohorts of students. Thus, for example, the grade 4 decrease in the percentage of students in the top two performance levels (65% in 2004 versus 63%), even if statistically significant, is difficult to interpret. The decrease does not necessarily represent lower effectiveness, since variability in the cohorts could well explain the observed difference.

An analysis of growth attempts to adjust for the different starting points of each cohort and is represented by the two transition matrices in the lower-left panel of Figure 2. Each transition matrix quantifies the growth of each cohort as they pass through the education from grade to grade, from year to year. The p_{11} entry for the lower transition matrix, equal to 0.33, represents the conditional probability of a student maintain the highest performance level, PL1, in grade 4 in 2005 given that they achieved PL1 in grade 3 in 2004. Thus, the panel depicts the third dimension, cohort, that exists along side grade and year.

The lower-right panel depicts increasing effectiveness over time. That is, increasing effectiveness represents changes in the transition matrices over time. Such changes, as mentioned previously, are likely to be minute and difficult to reliably detect. However, it is ultimately such changes that must take place in order that increasing achievement levels to occur (i.e., Cell A2, upper-right panel). This intuitive fact is often lost on those believing that growth analyses will make the universal proficiency requirements easier to achieve. A difficult requirement is difficult regardless how one measures progress towards it. Any benefit in using changes in growth to measure progress instead of changes in achievement is that growth, measured in the appropriate metric, is more closely aligned with the notion of education as a *process*.

If increasing effectiveness is the ultimate goal, then it is necessary to describe what increasing effectiveness means. When using scale score analyses, describing changes in effectiveness (e.g., changes in rates of growth) can yield statements that are difficult to interpret. For example, if the average scale score gain between grade 3 and 4 from 2004 to 2005 was 10 points, increasing effectiveness would imply an even large score gain—say, 13 points—between 2005 to 2006. A distinct

[5] Using the magnification function for pdf files allows one to view the details associated with the proportions in each grade in each year.

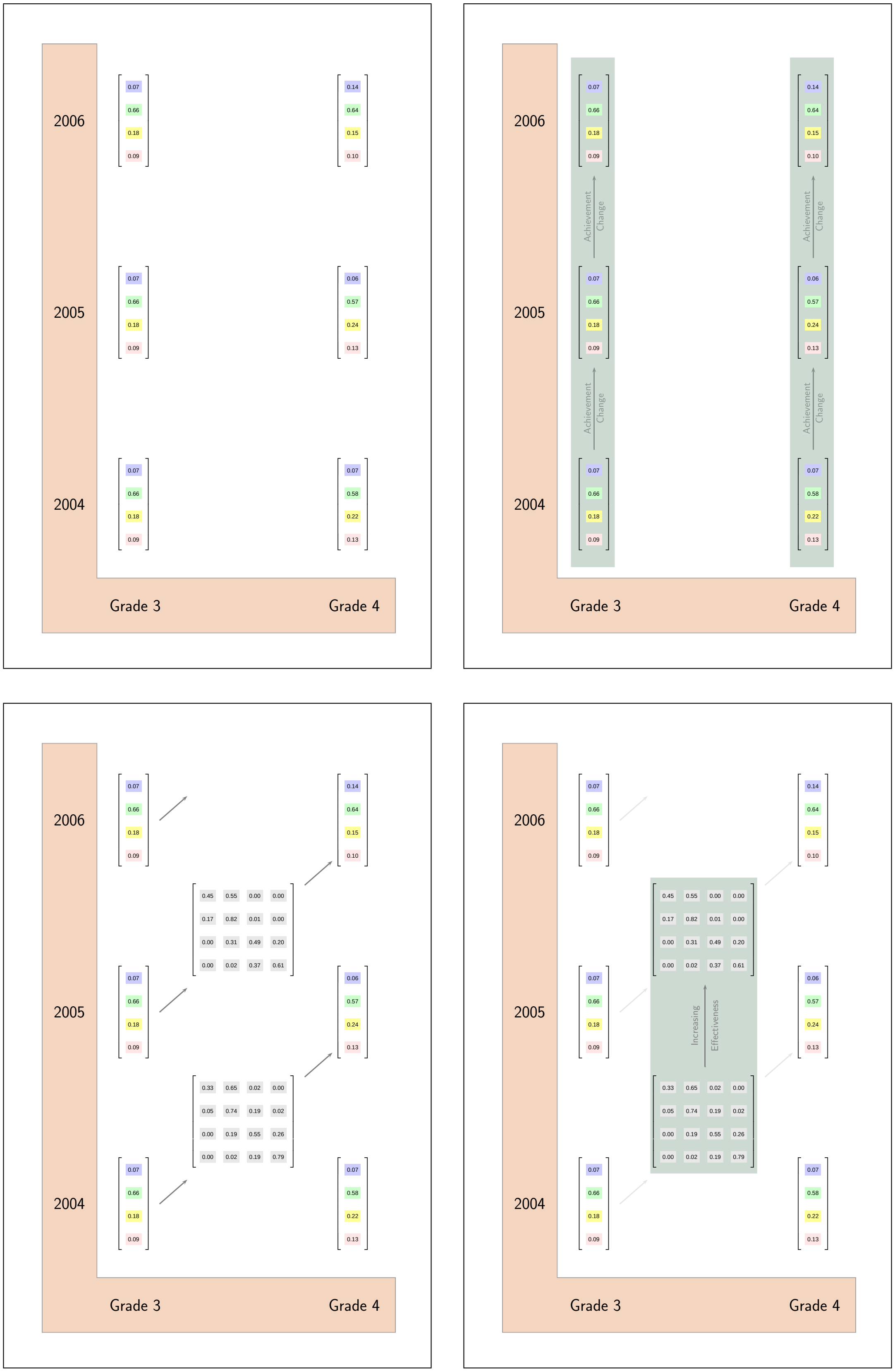

Figure 2: Carlson's four categories (see Table 1) with performance levels and transition matrices

benefit of Markov transition model is that change in effectiveness implies change in transition probabilities which is much easier to articulate. For example, the transition probability associated with student remaining in the lowest performance level in 2004 and 2005 between grades 3 and 4 is 0.79. That is 79 out of 100 3rd graders at the lowest achievement level are predicted to remain at that level in 4th grade. Improvement implies decreasing that rate. For example, a 10 year goal might be to reduce that rate to 50 out of 100 students, a probability of 0.50.

Modeling how changes in growth lead to changes in achievement is crucial if one wishes to understand the relationship between growth and achievement. The panels of Figure 2 suggest the connections that line growth and status across cohorts. Using successive matrix multiplication to derive status for successive cohorts (see Equation 3), one can expand this pattern to encompass numerous cohorts across grade 3 to 10 from now until the NCLB 2014 deadline. Figure 3 provides a graphical representation of this, simultaneously showing time (represented horizontally), grade (represented vertically) and cohorts (represented diagonally in blue/green as they pass from grade to grade, year to year). Examining the column vectors for a given grade across time (vertically) investigates achievement for successive cohorts over time—the current model of examining AYP for NCLB. Examining transition matrices for a given pair of grades over time investigates changing levels of effectiveness, and how changing levels of effectiveness impact outcomes.

Figure 3 is an interactive form allowing a user to enter different values into the transition matrices in order to examine how different levels of growth lead to different achievement level outcomes. This allows for a “birds-eye view” of the performance of an education system as measured by annual assessment results. The goal of the interactive spreadsheet is to allow users to investigate how much growth is necessary in order to bring out desired outcomes based upon percentages of students proficient. For example, if one takes the the NCLB guidelines for universal proficiency in 2014 seriously, then any student entering the system in 2013 who is below proficient, must have a 100% chance of moving to proficiency in a single year. If one take proficient to mean, for example, the upper two performance levels, then it is clear that the level of effectiveness necessary to produce such instantaneous results (a single year to proficiency no matter where you start) are extremely high and likely above anything demonstrated in schools today. By playing with different transition probabilities for the future—envisioning what one would like to occur—and then looking at current growth data to see what is currently occurring, one can rigorously enact Linn’s *existence proof* (Linn, 2003b).

The system presented in Figure 3 represents an over simplification of most education systems. In particular, the system depicted is closed—the system does not account for large perturbations to the performance levels based upon system inflow/outflow. For example, the influx of students associated with Hurricane Katrina into a state assessment system would likely alter the proportions of students across the different performance levels. This, in turn, would alter later results for a given cohort. As a first attempt to view the movement of millions of current and future students through an education system, however, it give perspective on the manner in which growth relates to effectiveness for multiple cohorts across an entire education system.[6]

Transition matrices affording universal proficiency

NCLB’s 2014 universal proficiency requirement places very rigid restrictions upon the form that a transition matrix can assume. These requirements allow the derivation of sharp analytic results with regard to what level of growth is necessary to achieve universal proficiency goals. Assume, based upon the four performance levels of Figure 1, that PL1 and PL2 represent proficient or

[6] The author has produced numerous posters based upon Figure 3. Viewing all the data using one large poster greatly enhances the ability to relate growth/effectiveness and achievement/status.

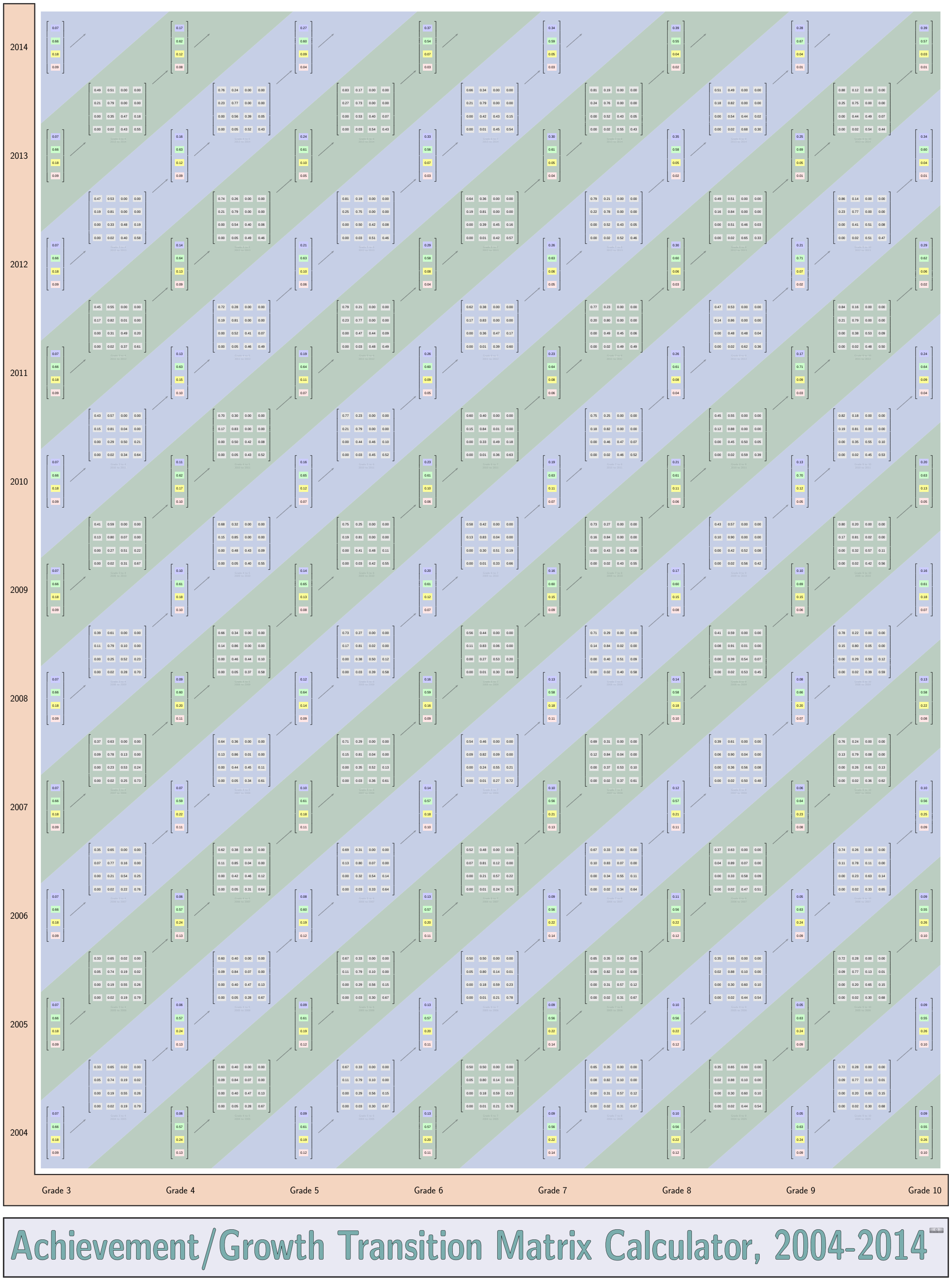

Figure 3: Depictions of multiple cohorts across grades and time using the Markov transition model.

above status. Then if universal proficiency is to exist in 2014 but doesn't exist in 2013, then all students not proficient in 2013 (i.e., students in PL3 and PL4 in 2013) must have a 100% chance of transitioning into either PL1 or PL2. Moreover, those student in PL1 and PL2 in 2013 must have a 0% chance of transitioning into either PL3 or PL4. Levels of growth such as these do not exist anywhere in school systems today and are unlikely to ever occur, thus demonstrating the difficulty in arriving at universal proficiency for all grades by 2014.

Another way of conceptualizing the relationship between growth and status is to investigate the steady state distribution associated with each transition matrix. Stochastic matrices, like those discussed in this paper, admit a steady state—a distribution, ω, that remains invariant under right multiplication by its associated matrix. That is:

$$\omega^T \mathbf{P} = \omega, \tag{4}$$

The distribution ω also represents the "long run" behavior of the matrix. That is, given any starting distribution, repeated right multiplication by $\mathbf{P}$ converges to ω. The existence of ω follows from the Perron–Frobenius Theorem (Brémaud, 1999, p. 197).

The requirement of universal proficiency can be formalized as a requirement on the value of ω. That is, one judges growth to be sufficient if and only if it would lead to an acceptable outcome were such growth experienced in perpetuity.[7] If ω represents the proportion of students in each of the four performance categories PL1, PL2, PL3 and PL4, with PL1 and PL2 representing proficient status or better, then NCLB requires that

$$\omega^T = (\gamma, 1-\gamma, 0, 0) \quad \text{where } 0 \leq \gamma \leq 1. \tag{5}$$

What are the properties of the transition matrix $\mathbf{P}$ such that ω represents the stationary distribution of $\mathbf{P}$?

The goals of NCLB, if met, define a situation, ideally, where Equation 5 represents a stationary distribution for the modeled system. That is, all students are exposed to growth that is sufficient to yield universal proficiency as a long term result. This situation, though unrealistic, yields a number of important results that are fundamental to understanding a Markov chain with Equation 5 and a probability distribution of some random variable X_n comprising the Markov chain. Consider the constraints imposed by the linear system given in Equation 4, where $\mathbf{P}$ is a tri-diagonal stochastic matrix. Solving the system yields the restriction that $p_{32} = 0$. That is, for an NCLB stationary distribution to exist, the probability of transition from the proficient category to the partially proficient category must be zero. Figure 4 provides the Markov diagram associated with such a transition matrix.

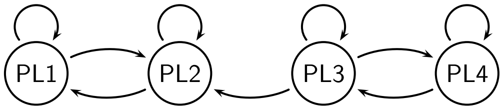

Figure 4: Markov diagram depicting transitions across grades between performance categories such that NCLB proficiency mandates are eventually met

[7] Steady states associated with transition matrices are an ideal metric with which to compare the growth represented by two transition matrices. Transition matrix $\mathbf{P_1}$ is said to be superior to transition matrix $\mathbf{P_2}$ if the steady state for $\mathbf{P_1}$ is preferable to that of $\mathbf{P_2}$. The notion of system equilibrium is easily modeled within the Markov framework but has no analog, to the author's knowledge, in other growth analyses.

Figure 4 is nearly identical to Figure 1 except for the removal of one arrow allowing for transitions from proficient to partially proficient. The elimination of the "bridge" prevents students who occupy the proficient and advanced categories from ever migrating back to unsatisfactory or partially proficient. In terms of policy, the consequences are immediately clear: In order to maximize the percentages of students in proficient or above categories, the percentage of students moving from proficient to lower categories must be kept to an absolute minimum. Stated colloquially, the key to getting all students ahead is to keep students from falling backwards. Thus, a corollary to the realization of NCLB's proficiency mandate is that no child *falls* behind.

Extensions to the Basic Markov Framework

Numerous extensions to the basic Markov transition model are possible (Langeheine & van de Pol, 1994). This section discusses, perhaps, the most important extension: adjusting for measurement error. Most growth analyses, including many of the popular mixed-effects analyses, fail to account for measurement error, confounding measurement error within the random effects employed in the model. With the Markov transition model, measurement error attenuates the transition probabilities located on the diagonal—probabilities quantifying the likelihood of remaining in the same performance level from year to year—overstating the probability of a transition out of a given performance level. Thus, in terms of the tri-diagonal transition matrices discussed in this paper, the diagonal elements are likely underestimated and the off-diagonal elements are overestimated. Latent Markov models (also referred to as Hidden Markov models) add a latent variable representing measurement without error. Such a model is well suited for the estimation of transition probabilities when the observed trait is subject to misclassification due to measurement error (Vermunt & Hagenaars, 2004; van de Pol & Langeheine, 1990; van de Pol & de Leeuw, 1986).

Though certainly of interest from a technical perspective, it is not clear how one might utilize measurement error adjusted growth within an accountability system. Accountability systems are implicitely built around observed scores. Were an accountability system built around observed transition probabilities, then some of the transitions attributed to a school, district, or state would be erroneous because of measurement imprecision. For example, numerous transitions from the lowest category, PL4, to the next to lowest category, PL3, would be false positives that might later result in false negatives when, in the following year, numerous transitions from PL3 to PL4 occur. Nonetheless, correction for measurement error in growth analyses provides a more complete picture of the process of the system over time. What follows is a description of how to augment the Markov transition model discussed thus far to account for measurement error, even in situations with non-uniform conditional standard errors of measurement.

The goal of accounting for measurement error is to quantify spurious transitions, with the ultimate goal being the estimation of transition probabilities associated with the true status of students. It is assumed that associated with each of the observed performance levels (also referred to as manifest performance levels) are classes associated with a corresponding latent variable. Markov models employing latent variables are referred to as latent Markov models (Wiggins, 1973). In standard applications of these models, it is often the case that the extent of the measurement error between latent and observed variables is unknown and subject to estimation. This is generally not the case with educational assessments where conditional errors of measurement are commonly available. In order to use this information to quantify the extent of these spurious transitions, a stochastic matrix called a misclassification matrix is employed (Kupermintz, 2004; Rogasa, 1994).

Let T denote latent variable associated with observed variable X. Assuming a one-to-one correspondence between performance levels and latent classes, the misclassification matrix, denoted $\mathbf{\Pi}$, is the matrix of conditional probabilities connecting latent classes (i.e., true status) with observed variables (i.e., observed status). That is,

$$\mathbf{\Pi} = \{\pi_{ij}\}_{i,j \in E} \quad \text{where} \quad \pi_{ij} = \Pr(X = j | T = i), \tag{6}$$

where E, denotes the set of performance levels. In most situations one assumes an equal number of performance levels and latent classes so the matrix $\mathbf{\Pi}$ is square.

Estimation of $\mathbf{\Pi}$ is not difficult if one has the cross-tabulated frequencies associated with observations in each class/level of the latent and observed variables. This approach is considered in Kupermintz (2004) using expert ratings associated with student performance. It is rarely the case, however, for one to have latent class information on students. Most state assessment systems utilize scale scores and covert student scale scores to performance ratings using cut-points on the scale. Given such a situation, estimation of the $\mathbf{\Pi}$ is not difficult given the conditional standard errors associated with each scale score and the proportion of students scoring at each scale level. The misclassification probability is simply a weighted average of the probabilities of misclassification across all scale score values in the given observed stated and true state. That is, suppose one wished to calculate the missclassification rate associated with students whose true class is PL1 but whose observed performance level is PL2. Then,

$$\begin{aligned} \pi_{\mathsf{PL1\,PL2}} &= \sum_{SS_t \in \mathsf{PL1}} \Pr(SS_t | \mathsf{PL1}) \cdot \Pr(SS_o \in \mathsf{PL2}) \\ &\approx \sum_{SS_t \in \mathsf{PL1}} \Pr(SS_o | \mathsf{PL1}) \cdot \Pr(SS_o \in \mathsf{PL2}) \end{aligned} \tag{7}$$

Where SS_o and SS_t represent observed and true scale scores, respectively. Note that Equation 7 uses observed frequencies associated with scale scores instead of the unknown true score frequencies. Thus, using Equation 7 one can estimate the missclassification matrix, $\mathbf{\Pi}$, of Equation 6 using commonly available data.

Taking account of true status, growth is now conceived as occurring along two parallel diminsions: the latent dimension and the observed dimension. Growth within each of these dimensions is modeled using a transition matrix, $\mathbf{P_l}$ and $\mathbf{P_m}$ for each of the two dimensions. The latent and observed dimensions are connected at each grade using the misclassification matrix for that grade, $\mathbf{\Pi_n}$. A graphical representation of this is provided in Figure 5. Figure 5 represents the latent and observed dimensions (each assumed to have 4 levels) as horizontal panels and the grade level dimension as vertical panels. The result is a commutative diagram linking growth across the two dimensions across time.

Discussion

Thus far the discussion of Markov chains and their application to modeling growth using performance levels has primarily focused on theory. To emphasize the relevance of Markov transitions to analysis of growth within current accountability systems, a more detailed example is presented. The investigations of this paper were motivated by a simple question: How much better must the education system have to be in order to sustain the results envisioned by NCLB? The Markov transition model developed in this paper allows one to analyze current performance and then project what level of student growth is necessary in order to achieve these outcomes based policy mandates. Once target levels of growth are calculated, it is a simple matter to assess whether any school or district currently sustains such levels of student growth. Thus, the analyses in this paper allow one to rigorously investigate Linn's notion of an *existence proof* (Linn, 2003b): Does there

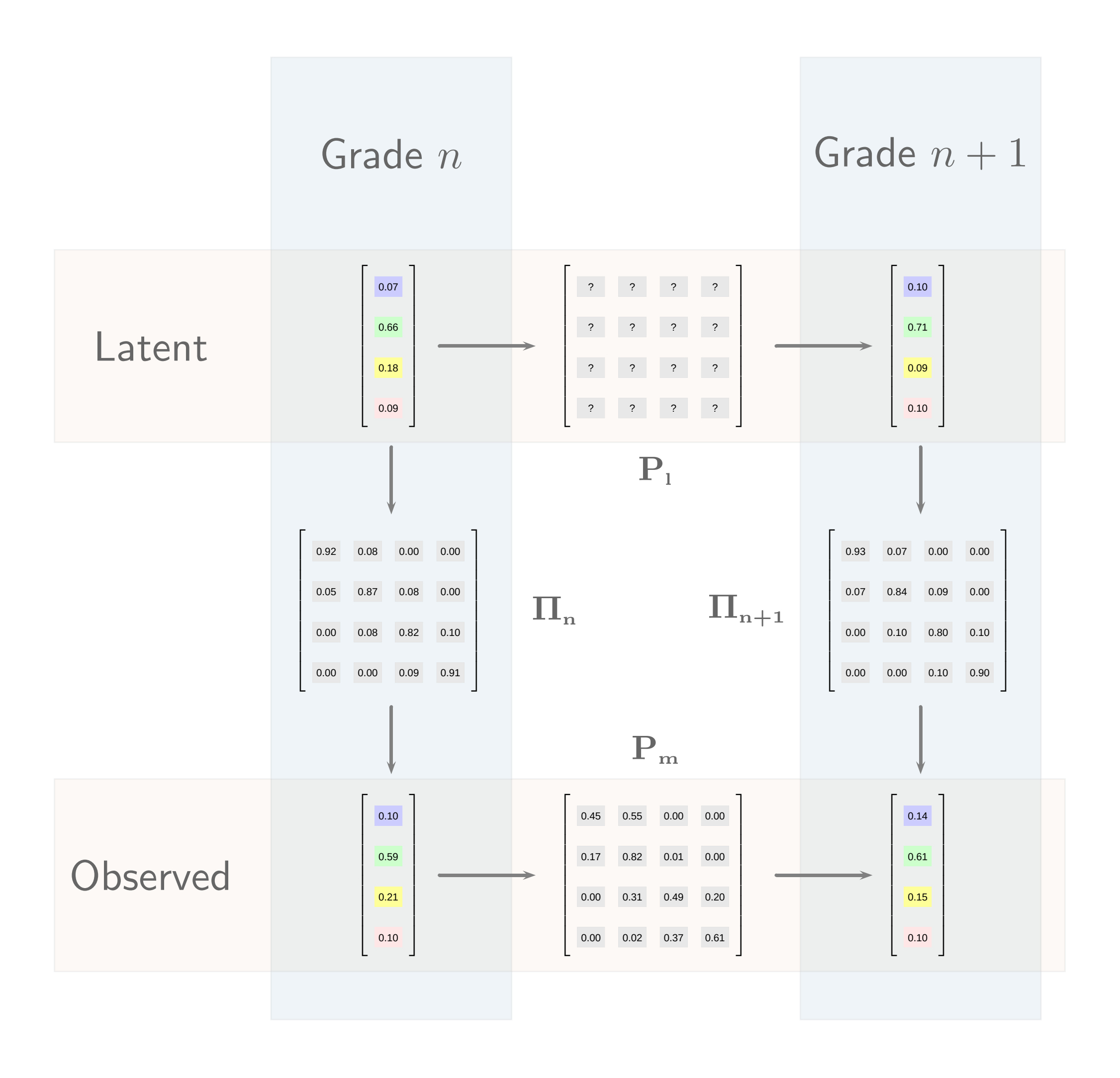

Figure 5: Commutative diagram depicting growth across latent and observed dimensions together with misclassification matrices by grade

exist a school that demonstrates the desired level of student growth (i.e. institutional effectiveness) modeled using the Markov transition analyses.

Recall matrix $\mathbf{P}$ from Equation 2. This matrix represents a snapshot of reading growth for a school between grades 3 and 4—a measure of *what is*. Associated with each school containing a fourth grade is a transition matrix. The distribution of transition probabilities for these schools provides a backdrop of what is possible—that is, what can be expected of schools. Table 2 provides descriptive statistics for the distribution of transition probabilities across schools with at least 15 students at the given performance level.

The transition probabilities expressed in Table 2 provide a continuum demonstrating a range of performance for schools. For example, with regard to maintaining performance at the advanced level, there exists a highly exceptional school for which 88% of its students classified in the highest performance level, PL1, in grade 3 were classified as PL1 in grade 4. In contrast is a school where only 31% of its students rated PL1 in grade 3 maintained that performance level in grade 4. Close

		PL1	PL2	PL3	PL4
PL1	Minimum	0.41	0.08	—	—
	25th Percentile	0.59	0.22	—	—
	50th Percentile	0.71	0.29	—	—
	75th Percentile	0.78	0.41	—	—
	Maximum	0.92	0.59	—	—
PL2	Minimum	0.03	0.12	0.06	—
	25th Percentile	0.06	0.40	0.32	—
	50th Percentile	0.11	0.49	0.40	—
	75th Percentile	0.16	0.56	0.50	—
	Maximum	0.40	0.80	0.80	—
PL3	Minimum	—	0.01	0.56	0.01
	25th Percentile	—	0.04	0.81	0.05
	50th Percentile	—	0.06	0.85	0.09
	75th Percentile	—	0.11	0.89	0.13
	Maximum	—	0.36	1.00	0.43
PL4	Minimum	—	—	0.12	0.31
	25th Percentile	—	—	0.27	0.57
	50th Percentile	—	—	0.35	0.65
	75th Percentile	—	—	0.43	0.73
	Maximum	—	—	0.69	0.88

Table 2: Minimum, maximum and quartiles of the distribution of school-level transition probabilities disaggregated by transition levels

examination of Table 2 reveals a number of interesting characteristics associated with performance category change over time as measured at the school level. It is the transitions involving the lowest category, PL4, that is of greatest concern. The probability of remaining within PL4 from grades 3 to 4 at an "average" school is approximately 0.7. That is, of 10 unsatisfactory students attending an "average" school in grade 3, 7 of these students are expected to remain at that performance level in grade 4. Even for a good school (i.e., a school at the 25th percentile with respect to students maintaining unsatisfactory status across the two years), nearly 6 in 10 students are expected to remain unsatisfactory. The descriptive statistics indicate that it is less difficult to transition from PL2 to PL3 than it is to move from the lowest category, PL4, to the next to lowest category, PL3.

Using the values in Table 2, one can construct a transition matrix that represents an achievable goal for schools across the state. For the purposes of this exercise, the values for schools doing better than three quarters of their peers are used. The transition matrix associated with this goal is given in Equation 8. The largest differences between the stochastic matrix of Equation 8 and that of Equation 2 are the probabilities for transition from PL2 to advanced. The steady state (i.e., the long term behavior) of this transition matrix is $(0.19, 0.74, 0.06, 0.01)$. That is, in the long run the system tends toward 7% of the students being categorized as either unsatisfactory or partially proficient. If the growth targets based upon what is currently observed in the system yield unsatisfactory results, then the system envisioned is a radical break from what currently exists.[8]

$$\mathbf{P} = \begin{pmatrix} .59 & .41 & .00 & .00 \\ .06 & .44 & .50 & .00 \\ .00 & .04 & .89 & .07 \\ .00 & .00 & .27 & .73 \end{pmatrix} \tag{8}$$

This paper has argued that current value-added models used to judge teacher and school effectiveness are limited because their results are normative and lack any connection to performance levels and their associated standards which permeate accountability systems nationwide. The data currently being collected by state assessment system has the potential to be an effective auditing system for education agencies, but to do so analysis techniques using the data must bridge the gap to the widely used performance levels. If evaluating teacher or school performance is a goal, then that evaluation should identify, at least broadly, deficiencies in performance failing to be addressed.

John Tukey made the point that it is better to have an approximate answer to the right question than a precise answer to the wrong question. This, it seems, is an appropriate rejoinder to much of the current value-added research associated with teachers and schools. There is a great push toward getting better and better estimates of teacher/school effects for the purpose of ranking teachers/schools. Ranking teachers accurately provides a precise answer to the wrong question. This paper has argued that a superior approach is to provide information relating to process grounded in performance levels to the relevant stakeholders: teachers, schools, administrators and policy makers. This information gives insight about what is necessary to transform the current education system into a more effective one—the one envisioned in NCLB, where all students are at least proficient in reading and math. This paper has outlined simple Markov chain analyses providing approximate answers to what, the author feels, are the right questions. Future research will focus on refining these techniques to address similar questions.

References

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment for teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–65.

Barbosa, M. F., & Goldstein, H. (2000). Discrete response multilevel models for repeated measures: An application to voting intentions data. *Quality & Quantity*, *34*, 323–330.

Brémaud, P. (1999). *Markov chains: Gibbs fields, Monte Carlo simulation, and queues.* New York: Springer.

Carlson, D. (2001). *Focusing state educational accountability systems: Four methods for judging school quality and progress.* (Unpublished manuscript made available to the author)

Diggle, P. J., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data.* Oxford: Oxford University Press.

Fielding, A. (1999). Why use arbitrary points scores?: Ordered categories in models of educational progress. *Journal of the Royal Statistical Society, Series A*, *162*(3), 303–328.

Gnedenko, B. V. (1967). *The theory of probability* (4th ed.; B. D. Seckler, Trans.). New York: Chelsea.

Hedeker, D. (2003). Multilevel models for ordinal and nominal variables. In J. de Leeuw &

[8] An example of such a situation where *what exists* should not be allowed to trump *what should be* is the performance associated with certain minority populations.

I. Kreft (Eds.), *Handbook for quantitative multilevel analysis* (pp. 300–345). Boston: Kluwer Academic Publishers.

Hill, R. (2003). *Using longitudinal designs with NCLB* (Tech. Rep.). Dover, NH: Center for Assessment.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement* (pp. 3–25). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Huynh, H., & Schneider, C. (2004, April). *Vertically moderated standards as an alternative to vertical scaling: Assumptions, practices and an odyssey through NAEP.* (Paper presented at the 2004 annual meeting of the American Educational Research Association, San Diego)

Kingsbury, G. G., Olson, A., McCahon, D., & McCall, M. (2004, July). *Adequate yearly progress using the Hybrid Success Model: A suggested improvement to No Child Left Behind* (Tech. Rep.). Portland, OR: Northwest Evaluation Association.

Kupermintz, H. (2004). On the reliability of categorically scored examinations. *Journal of Educational Measurement*, *41*(3), 193–204.

Langeheine, R., & van de Pol, F. (1994). Discrete-time mixed Markov latent class models. In A. Dale & R. B. Davies (Eds.), *Analyzing social & political change.* London: Sage.

Linn, R. L. (2003a, April). *Accountability: Responsibility and reasonable expectations.* (Presidential address to the annual meeting of the American Educational Research Association, Chicago, April 23, 2003)

Linn, R. L. (2003b, July). *Accountability: Responsibility and reasonable expectations* (Tech. Rep.). Los Angeles, CA: Center for the Study of Evaluation, CRESST.

Linn, R. L. (2004, July). *Rethinking the No Child Left Behind accountability system.* (Paper prepared for a forum on No Child Left Behind sponsored by the Center on Education Policy, Washington DC, July 28, 2004)

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

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*(5), 304–305.

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

No Child Left Behind Act of 2001. (2002). Public Law 107-110.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS.* New York: Springer.

Popham, J. (2004, May 26). Shaping up the 'No Child' Act: Is edge-softening enough? *Education Week*, *38*(23), 40.

R Development Core Team. (2006). *R: A language and environment for statistical computing.* Vienna, Austria. (3-900051-07-0)

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*(4), 307–335.

Rogasa, D. (1994). *Misclassification in student performance categories* (CLAS Technical Report). Monterey, CA: CTB/McGraw-Hill. (Appendix to CLAS draft technical report)

Rubin, D. B., Stuart, E. A., & 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., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. In J. Millman

(Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press, Inc.

Thum, Y. M. (2003). *No Child Left Behind: Methodological challenges and recommendations for measuring adequate yearly progress* (Tech. Rep.). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. (CSE Technical Report 590)

van de Pol, F., & de Leeuw, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods & Research*, *15*, 118–141.

van de Pol, F., & Langeheine, R. (1990). Mixed Markov latent class models. In C. C. Clogg (Ed.), *Sociological methodology* (pp. 213–247). Blackwell: Oxford.

Vermunt, J. K., & Hagenaars, J. A. (2004). Ordinal longitudinal data anaysis. In R. C. Hauspie, N. Cameron, & L. Molinari (Eds.), *Methods in human growth research* (pp. 374–393). Cambridge: Cambridge University Press.

Wiggins, L. M. (1973). *Panel analysis.* Amsterdam: Elsevier.