# Measuring Student Growth with the Conditional Growth Chart Method

Author: Yi Shang

# BOSTON COLLEGE
## Lynch School of Education

Department of Educational Research, Measurement, and Evaluation


# MEASURING STUDENT GROWTH WITH THE CCONDITIONAL GROWTH CHART METHOD

Dissertation
by


**YI SHANG**


submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

December, 2009

# Abstract

The measurement of student academic growth is one of the most important statistical tasks in an educational accountability system. The current methods of measuring student growth adopted in most states have various drawbacks in terms of sensitivity, accuracy, and interpretability. In this thesis, we apply the conditional growth chart method, a well-developed diagnostic tool in pediatrics, to student longitudinal test data to produce descriptive and diagnostic statistics about students' academic growth trajectory. We also introduce an innovative simulation-extrapolation (SIMEX) method which corrects for measurement error-induced bias in the estimation of the conditional growth model. Our simulation study shows that the proposed method has an advantage in terms of mean squared error of the estimators, when compared with the growth model that ignores measurement error. Our data analysis demonstrates that the conditional growth chart method, when combined with the SIMEX method, can be a powerful tool in the educational accountability system. It produces more sensitive and accurate measures of student growth than the other currently available methods; it provides diagnostic information that is easily understandable to teachers, parents and students themselves; the individual level growth measures can also be aggregated to school level as an indicator of school growth.

# Chapter 1 Introduction

## *1.1 Background*

The No Child Left Behind (NCLB) Act of 2001 resulted in the implementation of a nationwide accountability system whereby schools are held accountable on the basis of student achievement test results. Similar accountability testing systems had been put in place by a number of states in the 1990s (Haertel and Herman, 2005; Linn, 2005). These policies represent a hoped for political solution to an educational problem which is characterized by inadequate and unequal student academic achievement (McDonnell, 2005). In order to reach the goal that all students reach proficiency on State academic achievement standards by the year of 2014, annual large-scale testing of students in grades 3 through 8 in reading and mathematics is carried out, and schools are first evaluated then rewarded or penalized according to the results of the tests. These specific provisions reflect policymakers' belief that sanctions and rewards to schools will boost teachers' and school administrators' level of effectiveness (McDonnell, 2005).

The problem with this rationale is that sanctions and rewards are justified only when causal effects of schools on student achievement can be accurately estimated. In other words, we should at least be certain that a school *caused* the academic lag in its students before we penalize it. And yet credible causal inferences are not easily made on the basis of observational data. Various types of mixed-effects models have been proposed to identify and estimate school and teacher contributions to student achievement gains based on longitudinal test data (see, for example, Ballou et al., 2004; McCaffrey et al., 2004; Tekwe et al., 2004; and Lockwood et al., 2007). Yet many statisticians argue that causal

inferences about school effectiveness based on results of these models are not completely scientifically defensible (Rubin et al., 2004; Raudenbush, 2004; Braun, 2005). In the rare cases where the school effect on student achievement can be isolated from all non-school effects with observational data, the estimated school effectiveness is still a composite notion which includes factors not controllable by school personnel such as peer effect (Raudenbush, 2004). Since schools should not be held accountable for things not in their control, results of these models should not be used as direct evidence for school sanctions or rewards. And the present practice of evaluating schools based on comparing aggregate student achievement to an absolute standard (i.e. making the Adequate Yearly Progress AYP) regardless of where the students started is much less valid (Linn, 2005a). Moreover, even if causal inferences can be soundly made, it is doubtful whether penalizing the ineffective schools would help school personnel to become more competent.

Due to these problems, Linn (2006) calls for a change in perspective in the accountability system—instead of making explicit or implicit causal statements based on large-scale testing results, the test scores should be treated as a source of descriptive information and could be used to form hypotheses for further studies. "Treating the results as descriptive information and for identification of schools that require more intensive investigation of organizational and instructional process characteristics could be of considerable value .... It is unlikely that such a change in perspective would be politically acceptable at the present time. The change, however, would make the use of accountability results more consistent with the tenets of scientific reasoning and

research" (Linn, 2006).

*1.2 The Conditional Growth Chart Method*

The main purpose of this thesis is to propose a model or a family of models which evaluate student and school progress based on testing results from a perspective consistent with that of Linn (2006). The model proposed is called the conditional growth chart method. It has proved to be an effective tool in pediatrics and other areas of public health. One of its major applications has been to generate descriptive information about children's development based on longitudinal height and weight growth data, which helps in the early diagnosis of various diseases (see, for example, Cole, 1994; Carey et al., 2003; and Wei et al., 2006). With some revisions in model specification, this method could be used with student longitudinal test data to produce easily interpretable descriptive and diagnostic information about the growth trajectory of each student and school, and to flag those who need further investigation and assistance.

The basic idea behind the conditional growth chart is that, by comparing a child's measurement with an appropriate set of norms or reference values, we can decide whether the child lies outside of the normal range and needs further tests. An unconditional reference growth chart usually has the dependent variable (such as height or weight) plotted against age. It consists of a group of smooth curves each of which represent a chosen percentile of the population over time obtained from cross-sectional data. Once the child's results are located at each age point, it becomes apparent which percentile groups she belonged to, whether she has changed her rank in the population dramatically over time, or whether she showed typical trajectories for her age group.

The method becomes a more powerful diagnostic tool when the unconditional growth chart is replaced by a set of conditional growth charts. In this process, longitudinal data sets are collected instead of cross-sectional ones, and relevant variables such as children's growth history are added as conditioning covariates. A special regression technique called quantile regression is a major component in the conditional growth chart method. Quantile regression is similar to the familiar Ordinary Least Square (OLS) regression, except that it does not estimate the conditional mean of the dependent variable as does OLS regression. Instead, it estimates outcome values that correspond to a set of preselected percentiles in the conditional distribution of the outcome variable given the covariates. With these values as a reference frame, it is possible to determine the relative location of an individual in the conditional distribution of the outcome, which is called the conditional percentile.

In the context of educational measurement, we use the conditional growth chart method to reconstruct the conditional distribution of each year's test score given students' historical growth patterns. The quantile regression method helps us to describe the distributions more accurately than OLS regression, since the latter only estimates the mean and the variance of the conditional distribution which is assumed to be normal, while the former estimates a series of percentiles of the distribution without an assumption about its shape. The conditional percentile score estimated for each student answers the question—is she growing faster or slower than her peers who started from the same place? The student-level information can also be aggregated to school level. Schools in which most of the children have low conditional percentiles in their respective

groups merit further investigation to decide what resource they lack or what other assistance they need. On the other hand, it may also be necessary to study schools where most children grow at a higher rate than their peers conditioning on past scores, since it would help educators and researchers to understand what factors (of the school, the neighborhood, the student peer group etc.) contribute to higher academic growth rate and whether (or how) these factors could inform policy.

*1.3 How the Growth Chart Method Addresses the Limitations of the Current Educational Accountability System*

In order to illustrate the benefit of applying the conditional growth chart method to the field of educational assessment, it is necessary to probe further the limitations of the current school-based educational accountability system. Perhaps one of the most visible drawbacks of the system lies in its unreasonable expectations for schools. Linn (2003, 2005a, 2005b) has repeatedly called attention to this problem. He points out that the AYP targets set in the early years of the NCLB implementation have already been unrealistic for many schools that started with low performance, and they "will become increasing so, not only for those schools but for all schools as the increases in AYP targets start kicking in, especially in 2005 and 2008 when many states will have big jumps in their AYP targets" (Linn, 2005a, p.19). This prediction has been confirmed. By the year of 2007, 1000 of California's 9500 schools were branded chronic failures. State officials predicted that all 6063 public schools serving poor students will fail to meet the universal proficiency target by 2014. In Florida, 441 schools have failed the AYP target for 5 years consecutively and are candidates for closing. In Maryland, Baltimore alone has 49

schools falling in this category (Schemo, 2007). After AYP results for the academic year of 2007-08 were released in the summer of 2008, the report of National Education Association summarizes that "the number of schools failing to make AYP has increased, dramatically so in many cases. In several states the rate at which schools are failing AYP doubled, tripled, and even quadrupled (from that of the previous year)" (NEA, 2008). On the one hand, these facts underline the inadequacy of the current U.S. educational outcome and the urgency for educational reform. On the other hand, they clearly show that the AYP targets have been unrealistic for these schools. If states strictly follow the law and take over all these failing schools, which they apparently have not (Schemo, 2007), their resources would be depleted. Ambitious and realistic goals could inspire and motivate educators and students, but when the goals are evidently unattainable, they do no more than demoralize everyone involved in the system.

To set reasonable expectations, Linn (2005a) calls for an existence proof. That is, the goals should be grounded in past experience. If the highest performing schools were not able to achieve it in the past, then such targets should be called unrealistic. The standard for "realistic" also changes from school to school. Although ultimately the same expectation applies for every student/school to achieve proficiency in reading and math, at the present stage it must be acknowledged that the same goal clearly involves different amounts of effort for students/schools that started from different places. A target that is reasonable for one school may be completely impractical for another.

The conditional growth chart method proposed in this thesis comes as a handy tool to help setting reasonable objectives for schools and students based on their starting

points. The conditional distributions of academic achievement of previous cohorts are estimated, and they help policy makers and educators to understand the probabilities that a specific goal was achieved in the past by students/schools with different academic backgrounds, so that they have a reference frame when setting goals for the present cohort. Suppose that, with data pooled over the past few years, test scores associated with the 50$^{th}$, 75$^{th}$, and 95$^{th}$ percentiles of fourth graders given a certain third grade score are found to be A, B, and C respectively. It would then be reasonable to expect the fourth graders in the present year who shared the same third grade score to at least perform at the level of A. The level of B may be a somewhat ambitious goal, considering that only a quarter of students who shared this academic background were able to make it, but it is still arguably obtainable with sufficient effort. If the official target is set at or higher than the level of C, policy makers and educators would know that it is probably an unrealistic goal given the small historical probability of reaching it. The rationale of using the conditional growth chart method to set objective for academic progress is discussed in further details in the next chapter.

Setting ambitious but obtainable goals for schools and students would be the first step toward building a more functional accountability system, and there are other ways to strengthen the current system as well. O'Day (2004) presents an argument concerning several features of the NCLB program that inhibit organizational improvement. A central limitation of the system concerns the nature and quality of the information provided. The new accountability has its attention focused on schools. Since intervention happens at the school level, most of the interpretable information generated in the system is used to

build school-level results. Yet one has to acknowledge that, no matter what kind of incentives and sanctions a school receives, the actions required for higher achievement have to ultimately come from individual students. Improvement of student academic performance depends largely on the feedback that they receive. With the school-centered accountability scheme, most efforts in the current educational statistics community have been concentrated on estimating school growth and school effectiveness, which are not very informative and applicable for individual students' development (O'Day, 2004). Test scores and achievement levels of each student are reported but no additional tool is provided for teachers, students, and parents to evaluate the academic growth patterns contained in these numbers.

Is a student keeping pace with her peers who share similar academic background? What kind of growth is typical for a student at her level? And does this typical rate of growth set students at this level on track to ultimately achieving the proficiency goal? The conditional growth chart method directly addresses these questions and provide helpful feedback in individual-level learning and instruction process.

One of the other important limitations of the new accountability observed by O'Day (2004) lies in its maladaptive incentive structure. All the negative incentives that are placed on “failing” schools serve to inspire fear in school administrators rather than motivation from students and teachers. As a result, escaping punishment might become a higher priority for some highly disadvantaged schools than improving learning, schools and other stakeholders may use part of their resource to “game the system” instead of helping students, and the educational system as a whole may become even less efficient

than it was before the reform. Familiar strategies that disadvantaged schools may employ to bolster aggregate test performance include teaching to the test, increasing special education placements, and preemptively retaining students etc. (Jacob, 2002; Figlio and Getzler, 2002; Cullen and Reback, 2006). Figlio (2005) also shows that with the same type of misbehavior, students with lower academic performance is more likely to be suspended than those with higher academic performance, and the gap of punishment is substantially widened during the high-stakes accountability regime. Needless to say, these gaming practices would only put the already disadvantaged students more in harm's way.

Seven years after the NCLB Act was signed into law, there is still no compelling evidence that public schools in America are systematically improving at a greater rate than in the pre-NCLB era. While some studies show that state accountability systems in the 1990's positively affected the rate of change in student test performance (Hanushek and Raymond, 2003; Carnoy and Loeb, 2002), other studies demonstrate that gains on states' high-stakes tests typically shrink or disappear in low-stakes national and state tests (Jacob, 2007; Figlio and Rouse, 2005; Jacob, 2002). Fuller et al. (2007), using assessment data from both state tests and the National Assessment of Educational Progress (NAEP) spanning the 1992-2006 period, find that growth in fourth grade reading in 12 diverse states flattened out after enactment of NCLB, and growth of fourth grade math was slower post-2003 than before enactment of NCLB. The authors also find that no further narrowing of achievement gaps has occurred since 2002. In contrast, the Center for Education Policy (2007) reports that states have generally seen substantial gains in

reading and math since 2002 and narrowing of achievement gaps based on states' high-stake testing results. However, Fuller et al. (2007) point out that state testing programs are unstable due to changes of tests and cut points, and could produce biased results due to test design problems, and therefore should not be used to draw inferences about trend of growth. Their conclusion, based on the more stable and consistent results of NAEP, "raises the crucial question as to whether standards-based accountability is sufficient to advance more effective and equitable schools".

As much as we may acknowledge that schools with substandard instructional practice (if we could prove it) should be held accountable for their students' low performance outcome, we believe accountability should be implemented with much more caution than we see at present. Schools should not be judged based on a few numbers that are not scientifically defensible as estimates of their effectiveness. Schools should not be held accountable for things beyond their control, such as community effects. And above all, rewards and sanctions for schools should always be a lower priority than diagnoses and treatments. Finding out the reasons for unsatisfactory student growth requires carefully-designed quantitative and qualitative studies. It is important work but not the most urgent one. The first and foremost task is to find out exactly *where* the problems are —in which schools, in which grades, and for which individuals. The task is not as easy as it sounds, since achievement and growth are different concepts. Schools with adequate achievement are not necessarily growing at satisfactory rates. Making accurate diagnoses of students/schools' rates of growth helps educators to detect problems at their early stages.

The conditional growth chart method is designed for diagnosis of growth rates. It aims at *describing* the growth process as accurately as possible rather than finding the *cause* of differential growth rates. It does not claim to have the capacity to estimate school effectiveness, but it can easily identify schools where most students deviate from the "normal" growth rates. The results of the conditional growth chart model can be used to direct resource distribution to achieve more efficiency in the educational system. Indeed, the redistribution of resources and assistance, instead of reward and punishment, is the real force behind the improvement of education quality and the closing of achievement gaps (O'Day, 2004).

*1.4 Methodological Significance of This Study*

Besides significant policy implications, the model proposed in this thesis also introduces methodological innovations. The conditional growth chart based on quantile regression is a cutting-edge analysis technique in the field of biostatistics (Wei & He, 2006). The flexibility of the model and its attractive large-sample properties suggest that it could become a very promising member of the growth model family that is currently employed for educational accountability purposes. In this thesis, we also adopt an innovative SIMEX method to correct for measurement error-induced bias in quantile regression. The SIMEX method, combined with quantile regression, has not been utilized to model student academic growth before, and we believe they are powerful tools to help us understand the tremendous amount of test data that are generated in the educational accountability system in recent years.

*1.5 Outline of This Thesis*

Chapter 2 provides a literature review about some of the currently available methods to measure and evaluate student performance and growth, as well as school performance and growth. The basic elements of the growth chart method and quantile regression are also introduced. In chapter 3, we briefly describe our data and present the models that will be used in this thesis to estimate student growth. A specific section will be devoted to the measurement error problem and possible ways of adjusting for measurement errors. The SIMEX method, which will be used to corrects for measurement-error-induced bias, is explained in detail. We then proceed to empirical data analysis and a Monte Carlo study in chapter 4, where results of the models are presented and discussed. Chapter 5 concludes this thesis with its major findings, a discussion of the possible limitations of the models, and directions for future research.

## Chapter 2 Literature Review

There are four important types of outcomes in the new accountability system—student performance, student growth, school performance, and school growth. How to

define, estimate, and report these outcomes is one of the central issues in the implementation of NCLB policies. In this chapter, the concepts related to student and school performance and growth will be clarified, and the commonly used methods of measuring or estimating them will be discussed. We start with a brief review of student performance indicators and theories of scaling and linking. The pros and cons of using vertically-linked scores to measure student performance and growth are discussed. Then we present some common ways of measuring student growth, and introduce the concept of conditional percentile as an indicator of student growth which does not require vertical-linking. We proceed to explain how different conceptualization of growth can be quantified with conditional percentile produced from the growth chart method, and what advantage it has over the conventional methods. After analysis of student performance and growth, we move to school status and change and review some traditional methods of estimating school growth. Next we illustrate how unconditional and conditional growth chart models are constructed. The rest of the chapter will be devoted to the introduction of the key methodological issues involved in the model estimation, hypothesis testing, and inference about goodness-of-fit.

## *2.1 Student Performance and Scale Scores*

### *2.1.1 Student Performance Indicators*

Student performance refers to an individual student's academic standing, usually measured by a certain test at one particular time point. It is typically presented in one or more of the following forms: raw scores such as number correct or percentage correct scores; test-specific scale scores; performance levels; and norm referenced scores such as

percentiles, normal curve equivalents, and grade equivalents (Ferrara and DeMauro, 2006).

Raw scores are produced for practically all educational achievement tests. They are hard to interpret and raw scores from parallel tests administered at different occasions are not easily comparable. Scale scores are derived from raw scores to aid interpretation and comparison of test performances. They may be obtained normatively based on the distribution of a preselected reference group, or they may be transformed from student ability estimates computed through Item Response Theory (IRT). When a score scale developed for a specific test is believed to contain some information about content mastery, standard setting committees can set cut scores on the scale to delineate performance levels (e.g., basic, proficient, and advanced). Thus, student performance may be reported in terms of performance levels and are usually accompanied by performance level descriptions to support criterion referenced interpretations of student achievement. Percentiles, normal curve equivalents, and grade equivalents, on the other hand, support norm referenced interpretations of student performance. They describe how well a student has done on the test in relation to other students who took the same test or tests that are psychometrically parallel to this one.

*2.1.2 Scale Scores*

Scale scores deserve some more detailed discussion since they are the major mode in which student performance is reported in present accountability systems, and since performance levels are based on scale scores. In addition, student growth estimated by the method proposed in this thesis is also based on scale scores. For a good understanding

of student performance it is critical that test scores can be meaningfully interpreted and compared from year to year and sometimes from test to test. As mentioned above, scale scores are produced to satisfy this need.

Interpretability and comparability are actually results from two different psychometric methods, namely scaling and linking. The former usually refers to the construction of a scale for a single test or a test battery so that scores contain some normative or content-related information (Kolen, 2006). The latter mainly means transformation between the scores from one test and those from another (Holland and Dorans, 2006). The following paragraphs present a brief introduction to the theories of scaling and linking.

*2.1.3 Introduction to the Theory of Scaling*

Kolen (2006) summarizes the process of making test scores interpretable as the process of incorporating normative or content information into the score scales. Incorporating normative information into a score scale requires the designation of a norm group, which sets statistical characteristics of the scale score distribution (mean, standard deviation, etc.). The resulting scale scores show the relative standing of individual students with respect to this norm group. In this case, the scores and their meanings are strongly influenced by the choice of the norm group.

The incorporation of content information into score scales is commonly achieved through item mapping and scale anchoring, which associates items of the test with different score points on the scale. This way scores acquire a criterion-referenced meaning, i.e. the chosen score points correspond to some reasonably high probability of

answering the specific items correctly. Based on this mapping, score intervals on the scale are tied to various levels of student knowledge and skills. Attaching criterion-referenced meaning to scale scores is highly desirable from the perspective of test users such as students and teachers, and remains an important goal for psychometricians. Whether this objective has ever been satisfactorily achieved, however, is a much debated issue (Kolen, 2006). Forsyth (1991) has argued that fully incorporating content information in score scales through the current techniques of item mapping and scale anchoring may be unachievable unless the content domains are very well defined, which is not an easy task considering the complexity of human learning. Despite the challenge, criterion-referenced measurement of student performance has proliferated in recent years due to the standard-driven accountability policies.

Whether score scales are developed normatively or non-normatively, Kolen (2006) points out that it is crucial to incorporate score precision information into the scales. This means that the scales should be refined to include enough score units to preserve the precision of measurement in the raw scores, but not so many that small score differences resulting from measurement error are magnified and treated as if they are significant. The refinement of the scales, of course, depend heavily on the magnitude of standard error of measurement of the raw scores; however, it does not mean that the patterns of conditional standard error of measurement of the raw scores are completely preserved in the scales. In fact, they can become markedly different depending on the methods of scaling (Kolen et. al., 1992). As a general rule, conditional standard errors of measurement are reported at various score levels on the scale in standardized large-scale assessments. We will

return to the issues of standard errors of measurement in the next chapter when discussing the conditional growth chart method.

*2.1.4 Introduction to the Concept of Linking*

Scale scores from different tests or different forms of a test are made comparable to each other through linking. Linking is a vast topic and its theories and techniques are developing at a rapid rate. Over the years, psychometricians have proposed various categorization schemes for linking strategies. Mislevy (1992) and Linn (1993) summarize a hierarchy of different linking methods according to their data requirements and their statistical rigor. According to them, linking methods are classified into four categories—equating, calibration, projection (prediction), and moderation.

Equating is the strongest form of linking and is also the most demanding in terms of its assumptions. Two tests that can be equated must be designed to measure the same specific set of knowledge and skills, i.e. they must have the same content specifications, and they must measure the knowledge or skills at the same level of reliability, in other words, they must have the same statistical specifications. Kolen and Brennan (2004) adopt a quite similar definition of equating. They reserve the term to refer to the process where scores from alternate forms of the same test are related to each other, and they stress that "equating adjusts for differences in difficulty, not for differences in content" (Kolen and Brennan, 2004, p.3). As a result of such rigorous requirements, equated scores can be used interchangeably on different test forms.

When two tests do not satisfy the above assumptions and differ in content specification and/or statistical specification, the procedures to relate their scores is

generally called linking by Kolen and Brennan (2004). In Mislevy (1992) and Linn's (1993) taxonomy, this category is further broken down into calibration, projection, and moderation. The details in the differences between these linking methods are not addressed in this thesis.

*2.1.5 Vertical Scaling*

Vertical scaling, also referred to as vertical linking by Mislevy (1992) and Linn (1993), is the basis of many methods for measuring changes in student achievement over time (Kolen and Brennan, 2004; Smith and Yen, 2006; Doran and Jiang, 2006; Schmidt, Houang, and McKnight, 2005). It is a procedure that allows scores of students at different grade levels to be compared. Kolen and Brennan (2004) categorize it as scaling instead of linking, because comparability is not achieved by matching tests directly to each other but by relating scores of each test to a common scale. It makes the assumption that different tests, even though written for different grade levels, measure the same construct, which is usually referred to as unidimensionality. About this assumption Linn (1993) comments that "the calibration requirement that two tests measure the same thing are generally only crudely approximated with tests designed to measure achievement at different developmental levels".

Vertical scaling can only be achieved through certain data collection designs. Either tests administered to adjacent grades must contain overlapping items, or examinees in each grade must be randomly assigned to take the tests designed for their grade and their adjacent grades. Based on an adequate data collection design, various statistical methods can be employed to establish the vertical scale and estimate scale scores (Kolen and

Brennan, 2004).

The typical reason for developing a vertical scale is to compare scores from different grade levels and to use gain scores to measure student improvement directly. It is important to recognize, however, that there are serious limitations in terms of score interpretability and comparability for vertically scaled tests. Because tests of different grades are designed to address different content with different difficulty levels, each test may only exhibit good psychometric properties in a certain scale score region even after they are vertically scaled. Thus, the comparability of scale scores from tests of different grades is usually limited to certain ranges as well (Kolen, 2001). Scale scores corresponding to observed scores that fall outside of the range in a certain grade may contain more than an acceptable amount of measurement error.

An important assumption underlying most hierarchical linear models which are used to project student academic growth and many value-added models which are used to estimate school and teacher effectiveness is that test scores are vertically scaled and have a consistent interpretation over time. Recent studies have found that the construction procedures and psychometric properties of the vertical scale can significantly impact the results of the growth and value-added models in ways that are not completely predictable and not fully understood. For example, Briggs et al. (2008) find that growth modeling results are quite sensitive to the way an underlying vertical scale is established. Based on the same state assessment data, the authors create vertical scales and estimate scale scores using different Item Response Theory (IRT) models and different calibration and estimation methods, all of which are theoretically defensible. Employing a properly

specified growth model leads to strikingly different educational accountability conclusions depending on the vertical scaling procedures used. Doran and Cohen (2005) also show that the vertical scaling process introduces an additional component of error variance, and as a consequence, value-added models may estimate school and teacher effects with much less precision than statisticians used to believe. Martineau (2006) also demonstrates that violations of the vertical scale assumption of unidimensionality can lead to dramatic distortions in value-added estimates.

The conditional growth chart method introduced in this thesis requires longitudinal student assessment data. Moreover, it requires that test scores of the same grade from different years must have consistent meanings, i.e. scores are horizontally scaled, but it does not require vertical scaling. The reason for this will be explained in chapter 3 where the methodology is laid out in detail. Such flexibility is no doubt one of the major advantages of the growth chart method. However, when utilizing tests scores that are *already* vertically scaled, the growth chart method may suffer in the same ways described by the above mentioned authors. We recognize the importance of exploring the impact of scaling on the outcome of the growth chart method, but the topic is not the focus of this thesis.

*2.2 Student Growth*

The above discussion leads us naturally from assessing performance status to evaluating change of status. Measuring individual change is among the most important topics in educational measurement. Educators are ultimately concerned with individual learning, and "the very notion of learning implies growth and change" (Willett, 1988).

Individual academic growth is the principal intended outcome of building effective schools with competent teachers and sound curriculum. To conduct any evaluation of schools, teachers, or educational programs and policies, one has to start with measuring student growth in some way.

*2.2.1 Ways of Measuring Student Growth*

With many achievement testing programs using ordinal achievement levels as a major component of their score reports, perhaps the most intuitive way of indicating student growth is to describe their change in performance levels over time. The current safe harbor provisions under NCLB are aggregated indicators of school growth that are based on individual students' change of achievement levels. This way of describing growth can be easily understood by all stakeholders given the convenient definitions of the labels such as “non-proficient”, “proficient”, and “advanced”. Performance levels, however, offer a very coarse description of student academic status, and any measurement of growth based on these levels involves a loss of information. Students often make substantial progress while remaining in the same performance level. Such growth would be inexcusably lost if we evaluate growth based only on performance levels. On the other hand, very small changes across cut scores of the performance levels would be captured and magnified. To measure change more faithfully requires a more refined scale.

Another straightforward choice in assessing student growth is to use the difference or gain scores. They can be derived from the same tests that are administered at different times or from vertically-linked tests that address different grade levels. In the former

case, bias, reliability, and other relevant coefficients are estimated for difference scores as evidences for its inherent deficiency (see for example Lord, 1956 and Willett, 1988). Besides the psychometric properties, this type of difference score has a critical drawback: Using the same test at different time points makes it impossible for the test to target the current level of student knowledge and skills. Much is taught and learned during an academic year, and with the same test being used in the beginning and end, student learning would likely not be measured with appropriate precision. Over longer time periods, use of the difference score to measure student growth would be even more problematic.

With vertical scaling, difference scores can be produced from tests that are specifically designed for students' current learning levels. But this method of measuring growth is also very problematic as we have discussed in the previous section. Of greatest concern is that difference scores from vertical scales are not interpretable or comparable. Braun's (1988) analysis of the difficulties with measuring change still holds today—gains of scale scores are quantities that cannot be confidently explained or easily compared with each other. A 20 point gain in vertically-linked math scores in an academic year may represent a typical amount of growth for students who start with mid-level achievement, but it could also mean a breakthrough for students starting from very low levels, or an impossible amount of progress for students who already scored very high.

Doran (2004) reviews the normal educational growth model which is based on Normal Curve Equivalent (NCE) scores. In this model, it is considered “adequate” for a student to maintain or exceed the same position in the distribution over time. In other

words, a gain score of 0 computed from the NCE scores is considered as expected growth. Doran (2004) points out three major problems with this method. First, it requires different amounts of growth to maintain one's position in the distribution depending on where one starts, therefore growth measured in this way is not comparable. Second, the gain score in NCE units is obtained using only two data points which does not provide enough information about growth trend over time. And third, growth measured in NCE gain scores provide no information as to whether students are growing toward an acceptable standard of academic performance.

*2.2.2 Definitions of Adequate Growth, Normal Growth, and Expected Growth*

In this method cited by Doran (2004), the terms of "adequate growth", "normal growth", and "expected growth" are used interchangeably. To facilitate discussions in the following paragraphs, we define these terms separately, and explain how these concepts of growth are quantified in this thesis.

Adequate growth, as Doran (2004) argues, is a concept that implies adequacy with respect to some externally defined standard. That is, adequate growth, like adequate achievement (i.e. proficiency), refers to an underlying criterion. Following current growth-to-standard approaches, it could be defined as the rate of growth necessary for a student to reach proficiency in the designated time.

By contrast, normal growth relies more upon norm-referencing than an external criterion. Whether a student's growth is judged normal or typical is dependent on how much other students have grown, especially those students sharing similar backgrounds. We suggest that, in quantifying normal growth, the conditional distribution of current

scores given past scores be used instead of the unconditional distributions of scores from the whole population, because the former employs a more relevant reference group for purposes of comparison. If, among those who started from the same place, a student falls within a reasonable interval around the median in the conditional distribution of current scores, one could consider this growth "normal".

Expected growth is an expectation that takes into account both normal growth and growth history. It is defined for specific individuals according to the magnitude of growth of peers and the historical growth patterns of the student herself. As Linn (2003) argues, "current levels of performance and past gains provide a context for judging future gains".

*2.2.3 Definition of Conditional Percentile or Growth Percentile*

In this thesis, we propose using the conditional percentiles to quantify student growth. A conditional percentile is the percentile or ranking of a student in the conditional distribution described above. It shows the percentile of a student's current score relative to the group of students who have the same past scores. It is also a probability statement about how likely it is for a student to score at or below a specified level (or how unlikely it is for her to score at or above that level) given her past score(s).

Let $Y_1, \ldots, Y_n$ be independently and identically distributed (iid) random variables that denote the current test scores of students $1, \ldots, n$, and let $X_1, \ldots, X_n$ be iid random vectors that denote these students' past scores, which may include last year's score only or several years' scores into the past. Let $F_{Y|X=x}$ be the cumulative distribution function (cdf) of the conditional distribution of current score given past score(s), then $F_{Y|X=x}(y)$ is the conditional percentile corresponding to current score $y$

and past score(s) *x*. This quantity can usually be estimated in two ways. One approach is to derive the estimate of $F_{Y|X=x}(y)$ based on some distributional assumptions about $F_{Y|X=x}$. This approach will be discussed in more details later in this chapter. The other way is to estimate the conditional percentile through the empirical cdf of current score given past score(s), $\tilde{F}_{Y|X=x}$. Let $\tilde{P}$ stands for empirical probability. Then the conditional percentile of student *i* with current score $y$ and past score(s) $x$ can be estimated as:

$$\hat{Conditional\ Percentile}(y|x) \equiv \tilde{F}_{Y|X=x}(y)\cdot 100 = \tilde{P}(Y_i \le y | X_i = x)\cdot 100 \qquad (2.1)$$

Whereas unconditional percentiles normatively quantify achievement, conditional percentiles normatively quantify growth. This is because, in making the above conditional probability statement, students in the population are classified into different groups according to their past scores. Suppose, in one of these groups where everyone has the same level of past achievement, all group members make exactly the same amount of progress (or no progress at all), then their current achievement would not form a distribution, but can only be plotted as a single point. In this case, the conditional percentile as defined in equation (2.1) cannot take on any value other than 0 and 1. The conditional distribution of current achievement given past achievement is a proper distribution if and only if students in the same group grow at different rates. For this reason the conditional percentile is a measure of relative growth, and can also be called student growth percentile.

It was mentioned earlier that the estimation of conditional percentiles does not require test scores to be vertically scaled. Another advantage of using conditional

percentiles is that they are immediately interpretable and comparable. The interpretability mainly comes from the specifically defined reference groups based on the values of the conditioning variables. As discussed earlier, gain scores on vertically-linked scales are often not interpretable or even misinterpreted because the same amount of growth measured by scale scores usually means different amount of progress in learning for students who started high and those started low on the scale. On the other hand, a growth percentiles carries an intuitive message. Being a conditional probability statement, it simply tells how unusual the growth is among students who share the same past achievement. A conditional percentile of 80 means that only 20 percent of the students in that group surpass this one in their growth.

The concepts of normal growth, expected growth, and adequate growth that are defined earlier can be conveniently and properly quantified in terms of conditional percentiles. We have briefly described how to define normal growth using conditional percentiles, as well as the rationale for making normative diagnosis. The definitions of expected growth and adequate growth both involve projecting student achievement into the future. The rest of this section reviews how projection is usually done for educational assessment, how it is done employing conditional percentiles, and how the latter projection is different from the former. The discussions about projection reveal the way in which expected and adequate growth are quantified and explain the basis for making criterion-referenced diagnosis.

*2.2.4 Projecting Student Achievement Using Conditional Percentiles*

Let random variables $X$ , $Y$ , and $Z$ represent scores on standardized tests

targeting three consecutive grades respectively. For simplification of notation, we use these letters to refer to their corresponding grades as well (e.g. grade X). Suppose test scores are available for two cohorts of students. One starts grade X at year 0 , and the earlier one enters grade Y at year 0. Table 2.1 summarizes the hypothetical data in terms of grades and years.

Table 2.1 Grades and Years of Two Hypothetical Cohorts $(X < Y < Z)$

| | **Year 0** | **Year 1** | **Year 2** |
|---|---|---|---|
| **Grade X** | Cohort b | | |
| **Grade Y** | Cohort a | Cohort b | |
| **Grade Z** | | Cohort a | Cohort b |

Holland and Dorans (2006) differentiate the concepts of "prediction" and "projection". They note that prediction is a method that links results from two different tests to each other when scores of both tests are observed for the same sample of students. Whereas projection is made from results of one test to another for a certain sample of students when only the former is available for the sample. In the context of Table 2.1, if data are available for both years 1 and 2, and we are interested in finding out the relationship between Y and Z for cohort b, we can *predict* Z from Y by regression. This process does not require borrowing information from other cohorts. If test scores are only available for years 0 and 1 for the two cohorts, and we are still interested in Z for cohort b, then we must *project* Z from Y. In the projection process, since Z is unknown for cohort b, we must somewhat rely upon our knowledge about the relationship between Y and Z for cohort a. In short, In order to make projection between Y and Z for cohort b, prediction must be carried out first between Y and Z for cohort a. Then certain

assumptions need to be made about the relationship between the populations which each of these two cohorts represents, and projections are made based on these assumptions.

Mislevy (1992) and Linn (1993) both identify "prediction" defined in the above paragraph (which they name "projection" in their own taxonomy) as a very weak form of linking. It is weak in the sense that it requires much less from the tests than other types of linking such as equating and calibration. Tests do not need to measure the same construct. As long as their results are correlated, prediction can be performed. The precision of the prediction depends on the strength of the relationship between the tests. The empirical relationship estimated through this process is quite sensitive to context, group, and time.

Prediction is usually carried out based on linear regression. For example, Pashley and Phillips (1993) model the relationship between results of the International Assessment of Educational Progress and the National Assessment of Educational Progress (NAEP) by administering both assessments to the same sample of students and then regressing the NAEP scores on IAEP values. Williams et. al. (1998) use a weighted least square regression as their basis of prediction of the NAEP scores from a state test. Holland and Hoskens' (2003) method of predicting the true scores of a test from the observed scores of another not necessarily parallel test is also based on OLS linear regression, except that the standard errors of the predicted scores do not come out of the regression analysis and involves the reliability of the predicted test. In the context of table 1, these predictions focus on estimating the conditional mean of Z given Y for cohort a when both Z and Y are observed on the same sample, i.e. $E(Z|Y=y,a)$ , and the imprecision of the prediction is usually measured by the conditional prediction error

variance, which is $Var(Z|Y=y,a)$ .

When the task shifts from prediction to projection, more assumptions have to be made to make up for the missing values of the projected test. In the case of projecting Z from Y for cohort b when year 2 data are not available, we have to assume that the conditional distribution of Z given Y is the same in both cohorts a and b, so that information about the relationship between the two tests in cohort a could be utilized to make projection in cohort b. This assumption can be characterized as

$$Pr(Z \leqslant z|Y=y,a) = Pr(Z \leqslant z|Y=y,b) \quad , \quad \forall\, y \qquad (2.2)$$

If we denote the conditional cdf's on the two sides of equation (2.2) as $F_{Z|y,a}$ and $F_{Z|y,b}$ respectively, then equation (2.2) can also be written as $F_{Z|y,a} = F_{Z|y,b}$ . Assumption (2.2) underlies the projections done in both Pashley and Phillips (1993) and Williams et. al. (1998). Holland and Dorans (2006) call it the "population invariance assumption", as it requires that the same conditional distribution holds for both populations or subpopulations. In our example, because a and b denote two subpopulations (cohorts) that could be indexed by time, we may also call it the time-invariant distribution assumption. It is not only needed for projections, for, as Holland and Dorans (2006) point out, population invariance assumption or assumptions analogous to it "pervade all aspects of scaling and equating where there are missing data" in the sense that in the above example the data for Z in cohort b are missing.

In this thesis, test score projections are performed with the same assumptions and rationales as those of Pashley and Phillips (1993) and Williams et. al. (1998), i.e. we estimate the conditional distribution $F_{Z|y,a}$ with available data from cohort *a* first, then

assume that equation (2.2) holds and project scores of students in cohort *b* according to their available scores and the conditional distribution $F_{Z|y,b}$ . The difference between the projection model proposed here and the conventional one is that the former employs conditional percentiles while the latter mostly centers on conditional means. This difference can be elaborated as follows.

First, models based on conditional percentiles differs from models based on conditional means in the prediction stage. The latter uses linear regression to estimate the mean and variance of $F_{Z|y,a}$ , and then impose strong normality assumption on the distribution (see for example Pashley and Phillips, 1993). In the model based on conditional percentiles, a series of conditional percentiles of Z given Y for cohort *a* are estimated. These conditional percentiles define the distribution $F_{Z|y,a}$ in an empirical way and makes fewer assumptions about its true shape.

Second, the model based on conditional percentiles has greater flexibility than the conventional linear regression model in the projection stage. After assuming that equation (2.2) holds, the best projection of Z in cohort *b* in the linear regression model is $E(Z|Y=y,b)$ ( Holland and Dorans, 2006). In the model based on conditional percentiles, however, there are various choices to make the projection. Suppose that a particular student in cohort *b* receives scores x and y in years 0 and 1 respectively, and has conditional percentile of y given x $P_{y|x,b}$ for year 1, and we are interested in projecting her score in year 2. If we define her expected growth according to how much her peers grow, then it may be reasonable to assume that this student will grow at the median speed compared to those who currently are at the same level as hers, then her

projected score $\hat{z}$ in year 2 will simply be

$$\hat{z}_1 = F^{-1}_{Z|y,b}(0.5) \qquad (2.3)$$

It is also possible to define expected growth according to the student's own past record. In this case we assume that this student will keep growing at her earlier normative speed, i.e. $P_{z|y,b} = P_{y|x,b}$, and her estimated score in year 2 is given by

$$\hat{z}_2 = F^{-1}_{Z|y,b}(P_{y|x,b}) \qquad (2.4)$$

To take both the peer growth and the student's own growth history into account, it is also possible to define expected growth as the greater value of (2.3) and (2.4). The projected score is the result of expected growth. Standard errors can be calculated for the projected score quantifying the degree of imprecision of the projection.

To assess adequate growth, it is necessary to make projections a few years into the future. If the same expected growth every year eventually leads the student to reach proficiency, then the growth is deemed adequate based on this particular criterion. We can directly compute adequate growth in terms of conditional percentiles given that the student must reach proficiency in *m* years.

Compared with conditional mean, which is estimated in most growth projection models, conditional percentile contains more diagnostic information. With the estimate of a conditional mean, the usual statement that can be made is "students with past score(s) $x$ are on average projected to be proficient (or not proficient) in m years". With projection calculated through conditional percentiles, however, one could say, "students with past score(s) $x$ are projected to be not proficient in $m$ years if they grow at conditional percentile $P_{y|x}$. However, if they grow at the rate $P'_{y|x}$ which is higher

than $P_{y|x}$ , then they are projected to be proficient". From the latter projection statement, teachers, parents, and students themselves can easily tell how difficult (or how probable) the task is and how much effort is involved in accomplishing it. Adequate growth, when quantified in terms of conditional percentiles, becomes a probability statement. The higher the required growth percentile, the smaller the probability that the student will eventually reach proficiency. If, according to the projection result, a student has to grow at the 95$^{th}$ conditional percentile every year to be able to reach proficiency, her chance of reaching the goal in the designated time is probably quite slim, or in other words, she will have to make extraordinary amount of effort to get there. Thus, the method of conditional percentiles not only defines adequate growth specifically for each student, but it also helps the student to understand what it takes to reach a specific goal. This information is more meaningful and constructive than a simple linear projection used in most available growth models.

*2.3 School Performance and Growth*

*2.3.1 Clarification of Some Concepts*

In any discussion about school performance and progress, some terms such as "status", "change", "growth", "effectiveness", and "efficiency" will be repeatedly used. Different authors often have different definitions for these terms. To avoid confusion, we explain in the following paragraphs the exact meanings that these terms adopt in this thesis. Before the discussion proceeds, we acknowledge that standardized test scores are by no means the only outcome of interest in evaluating schools, and arguably may not even be the most important outcome. Other valued school outcomes include drop-out

rates, graduation rates, attendance rates, and parent/community satisfaction etc. For this thesis however, we focus exclusively on how to estimate student academic growth which is measured by standardized test scores and how to aggregate the estimated individual growth to school level. Therefore alternative measures of school outcomes are not considered here.

Status of a school refers to the academic performance of the school measured at a single time point. In contrast, school change involves measurement at two or more time points, obviously, as does school growth. However, change and growth are not exactly the same concept. Growth is usually reserved for analyses where individual students are followed over time and measured repeatedly. School change treats the school as the basic unit and disregards the growth of individual student. In other words, school growth is the *aggregation of change* in individual performance which requires longitudinal data, while school change represents *change of aggregated values* of individual achievement in a school which uses cross-sectional measurements. For example, the average of individual gain scores in a school from one year to the next is a measure of school growth (Weeks and Karkee, 2008), while the change of average scores in the school from one year to the next is merely a measure of school status change (Hanushek and Raymond, 2002). Status change does not necessarily reflect real improvement or decline because sample differences are not accounted for. For example, the fact that fourth graders tested in 2006 in a school on average score 10 points higher than fourth graders in 2005 in the same school is no evidence that instructional practices in grade 4 of the school is improving, because the 2006 cohort might have started at a higher level than the 2005 cohort.

Another important term to be clarified is “effectiveness”. Many educational statisticians define school effectiveness as the contribution or causal effect of a school to overall student growth. Raudenbush (2004), for example, uses the term “school effectiveness” interchangeably with “school effect”, and Braun (2005) points out that “the word 'effectiveness' denotes a causal interpretation”. Further, Raudenbush and Willms (1995) differentiate two types of causal effects of a school on student academic growth. According to them, one kind of school effect, or “Type A” effect is evaluated for school choice purposes and contains all the factors of a school that affect student growth, including those not in control of the school faculty and staff, such as the neighborhood effect and peer effect. The “Type B” school effect is the effect of school practice on student growth. In other words, this is the effect that is controllable by school personnels and can be used for accountability purposes.

In this thesis, we use the term “school effectiveness” to denote the “Type A” effect, and let “school efficiency” represent the “Type B” effect of Raudenbush and Willms (1995). Both effects are usually estimated by adjusted versions of overall student gains. The difference is that when estimating school effectiveness, all non-school factors relevant to student growth must be controlled, while in estimating school efficiency, the relevant school factors not in control of the school such as available resource level and neighborhood effect must also be held constant in addition to the non-school factors.

*2.3.2 Measuring School Status and Status Change*

Various methods have been proposed to evaluate school performance and improvement based on test scores alone. School performance can be measured with the

average of student test scores in each grade of a school. Under NCLB's accountability model, the most commonly used school performance indicator is percent proficient—the percentage of students in the school who score at or above proficiency level. Each year, percent proficient of each of the several designated demographic group of a school is compared against a pre-defined target—the annual measurable objective (AMO) to determine whether schools are making adequate progress. Using status measure to evaluate progress, this accountability design is based on a confusion between status and growth.

The dominant approach to measuring a school's status change in a state accountability system is to take the difference between percent proficient (or percent non-proficient) of each subgroup in a school from two consecutive years. This difference of percent proficient between two years is one of the basic measures in the "Safe Harbor" provision. In the previous section, we have discussed the disadvantages of measuring individual growth based on performance levels, all of which apply to the difference of percent proficient as a measure of school change. Another important measure of school performance change is the difference between average scores in a given grade at different time points (e.g. the change of average score of fourth graders one year to fourth graders the next year).

*2.3.3 Measuring School Growth—The Gain Score Model*

Starting in 2005, many states submitted plans of incorporating school growth measures into their AYP models, and some of them have already been approved by the federal government as pilot programs. The gain score model and variations of it are the

most frequently proposed models for the pilot program (Weeks and Karkee, 2008). The individual gain score, as discussed in the previous section, is simply the difference between scores from the same individual at two different time points. The gain score model aggregates individual gain scores to the school level to measure school growth.

Gain score models are usually easy to employ and generate easily interpretable results, but they depend heavily on a sound vertical scale. Weeks and Karkee (2008) point out that the gain score model does not account for regression to the mean. Gain scores themselves also has the problem of containing larger amount of measurement errors as mentioned earlier. Despite these properties, gain score remain a very important concept in educational studies since many other growth models (typically mixed-effects models) often use gain scores as outcome variables (see Ladd and Walsh, 2002; Tekwe et al., 2004).

*2.3.4 Measuring School Growth—The Transition Models*

Another type of model generally referred to as transition models have also been proposed by some states for school accountability purposes. Transition models do not use scale score data, instead, they use longitudinal student performance data presented in achievement levels. These models aim to estimate the probability that an average student in a specific school moves from a given achievement level in one year to another achievement level in the next year. As an example, Betebenner (2007) presents a transition probability model based on the assumption that student growth is a homogeneous first order Markov process. The first order Markov process assumes that given last year's performance, a student's current performance is not related to his/her

earlier academic history, i.e. one's achievement at time *t* is dependent only on his/her level at time *t*-1. That the process is a homogeneous one means that school transition patterns does not change over time. Given *k* performance levels, the model produces a transition matrix $P=\{p_{ij}\}_{i,j\in[1,k]}$ where each $p_{ij}$ represents the probability that students in a specific school move from level *i* to level *j* during the designated time period. Higher probabilities of moving from lower levels to higher levels signify higher rates of growth in the school. Since this probability is not observed for individual students but can be approximated by percentages of a school's students in different achievement categories, this technique is mostly used with school-level data.

The transition matrix has the advantage of being more interpretable and "actionable" for teachers, principals, and other stakeholders (Betebenner, 2007). The method can also adjust for measurement errors of the test by including a misclassification matrix which summarizes the probabilities that students are misclassified in an achievement level (Betebenner et al., 2006). Nevertheless, it has the same shortcoming of measuring student growth with performance level change—progress made within the range of a single level is not counted. The results of the model are dependent on the density of students around the cut-scores, the ranges of different performance levels, and the location of the cut-scores which may be decided subjectively. Strictly speaking, transition analyses usually track cohorts within a school not individual students, and hence should probably be called a "quasi-growth" model.

*2.3.5 Measuring School Growth—The Mixed-Effects Models*

A very popular approach to assessing school growth has relied on mixed-effects

models. This type of model includes regression models which distinguish within-unit variation from between-unit variation so that the researcher can characterize each source separately (Pinheiro & Bates, 2000). The capability of these models to differentiate different sources of variation has great advantages in evaluating student growth due to the nested structure of longitudinal student achievement data. Measurements nested within the same students and students nested within the same schools (and classrooms) almost certainly have correlated regression residuals which, when ignored, will lead to biased estimates. The mixed-effects models minimize these biases by taking into account the effect of each hierarchical level on the variance-covariance matrix of the residuals. They are also able to handle the missing data problem by borrowing strength from similarly nested subjects in estimating growth trajectories (Raudenbush and Bryk, 2002).

Mixed-effects models estimate individual growth trajectories. Different methods are used to aggregate individual results to the school level. Tekwe et al. (2004) adopt a grade point average (GPA) approach to evaluate schools. Since mixed-effects models usually include a random school effect which is a best linear unbiased predictor of the "value added" associated with schools, each of these estimated school-effect coefficients can be divided by its standard error to produce a standardized measure of school growth in a particular subject. Different grades can be assigned to different values of the standardized score, and a school GPA can be calculated across subjects. Another type of aggregation is based on projections. The regression lines for each student in the mixed-effects models are projected out 2-3 years. If a student's projected score exceeds the proficiency cut-score in the projected grade then she is marked as being on track. The

percent of students who are on track to be proficient can be used as a measure of school growth (Weeks and Karkee, 2008).

*2.3.6 Using the Mixed-Effects Models to Assess School Effectiveness and Efficiency?*

Theoretically, if all non-school factors relevant to student academic growth could be controlled for, mixed-effects models can be used to estimate school effectiveness. The English education system, for example, has officially incorporated "contextual value added" performance measures as indicators of school effectiveness to help parents in school choice (Wilson and Piebalga, 2008). Hierarchical linear model is employed in the program to evaluate secondary schools with test scores taken at age 16 as outcome and those taken at age 11 as input. Other individual level predictors included in the model are gender, ethnicity, special educational needs status, eligibility for free school meals as proxy for low income, English as second language, student mobility, age, and measure of deprivation of the student's home neighborhood. School average intake scores and its standard deviation are included in the model as school level predictors to take account of peer effect. The model produces a predicted outcome score for each student. Subtracting the observed scores from the predicted scores, each student has a "contextual value added" (CVA) measure. The school average CVA measure, multiplied by a weight which adjusts for the small school volatility effect (Kane and Staiger, 2002), is used as a measure of school effectiveness. Despite the thoughtful design, Wilson and Piebalga (2008) caution that the model is still not able to isolate the school effect from all non-school factors. For example, student achievement prior to 11 years old that is not captured by the test score at age 11 may still play a role in the outcome scores but is not

accounted for in the model.

In the school accountability system of US, there have also been waves of proposals and intensive studies about using the mixed-effects models to estimate school and teacher effectiveness. As we mentioned earlier, many statisticians argue that estimates from these models should not be taken as direct evidence of school or teacher effects. Rubin et al. (2004) shows that even with randomization, "estimating relevant causal effects of teachers and schools is extremely difficult to conceptualize" due to the complications of peer effects and data missing not at random. When only observational data are available, model-based analysis usually cannot support causal inference without making unwarranted assumptions. In a similar vein, Braun (2005) cautions against making causal statements based on observational data. He argues that while value-added models are known for estimating teacher (or school) effectiveness, they cannot eliminate all competing hypotheses as alternative explanations of the differences in student achievements, such as differences in parental support, motivation, and study habits etc. Raudenbush (2004) also discusses this problem in detail and point out that the "school effect" estimated by the mixed-effects models includes contributions due to factors which cannot be controlled by schools, such as the neighborhood where the school is located, and contributions due to factors that are in the school's control, such as its instructional practice. Without randomization, it is hard to envision that any statistical model can completely isolate the latter from the former. In other words, even if we can estimate school effectiveness, we still do not know how much of it is due to the efficiency of the school practice and competency of its personnels. Therefore Rubin et al. (2004) advocate

a position of taking the estimates of the mixed-effects models as descriptive measures of school growth rather than school effectiveness.

*2.3.7 The Approach of This Thesis to Assessing School Growth*

As mentioned in the introduction, this thesis does not address the problem of making causal inferences about school effectiveness. We are interested in accurately assessing student and school growth. The main question we address is: How fast are students in this school growing academically?

Recall the discussions in the previous section about adequate growth, normal growth, and expected growth, one could go a step further and ask the following questions under the main one:

1. Are most students in this school growing at a rate that is adequate for eventually reaching proficiency?

2. Are most students in this school growing faster or slower compared to their peers in other schools who started from the same levels?

With the reference growth chart method proposed in this thesis, we describe student growth using estimated conditional percentiles and aggregate them to the school level. We experiment with different aggregation methods such as taking the median of students' estimated conditional percentiles and calculating the percentage of students in a school with exceptionally high and low growth percentiles etc. Our purpose is not to rank schools based on their growth from top to bottom, as most schools are hardly distinguishable from each other (Kane and Staiger, 2002; Wilson and Piebalga, 2008). The goal is to identify schools with problematically low rates of growth and unusually

high rates of growth, and to identify specific students within the schools with unusual growth patterns so that the schools and students can be studied and assisted.

*2.4 The Growth Chart Method*

In the introduction, we have briefly explained unconditional and conditional growth charts. In this section, these concepts are clarified further. The construction and application of the unconditional and conditional growth charts are also illustrated.

*2.4.1 Unconditional Growth Chart*

Figure 2.1 presents an unconditional growth chart for length and weight of boys from birth to 36 months old in the United States in the year of 2000. The chart is based on cross-sectional measurements of a nationally representative sample, and it depicts smoothed curves for the 5th, 10th, 25th, 50th, 75th, 90th, and 95$^{th}$ percentiles associated with the length and weight distributions for boys by age. From the percentile curves in the chart, we can roughly reconstruct the distributions of length and weight at each age point. Quick examination of the curves, for example, shows that the distributions of both length and weight are more dispersed at 36 months old than at birth. An individual can be located on the chart according to her age, and with the curves serving as references, it is easy to tell which percentile group the individual falls in, for both height and weight.

The traditional way of creating unconditional growth charts is empirical. In the case of boys' weight, for example, if census data are available the real percentiles can be calculated directly. If, instead of census data, a nationally representative sample is available, percentiles can be estimated empirically at each age point by simply calculating the ratio of boys at or below a given weight level to the total number of boys

in the sample at that age. Then a smooth polynomial curve can be fitted over these estimated points to construct the percentile curves in the growth chart (Hamill, et al., 1979).

The problem with the empirical estimation of percentiles is that standard errors of the estimated percentiles increase steeply towards the tails of the distribution; therefore, the extreme percentiles are often estimated inaccurately (Cole, 1988). One way to get around this problem is to fit a theoretical distribution to the sample, estimate the distribution parameters, and then calculate the expected percentiles from the estimated distribution. Equation (2.5) represents the conventional approach of assuming normality of the population distribution and avoiding the estimation of percentiles by estimating the mean and standard deviation (Wei et al., 2006).

$$\hat{Q}(\tau|t) = \hat{\mu}(t) + \hat{\sigma}(t)\Phi^{-1}(\tau) \qquad (2.5)$$

$\hat{Q}(\tau|t)$ is the estimated value that corresponds to the $\tau$ th percentile of the population at time t, $\hat{\mu}(t)$ and $\hat{\sigma}(t)$ are the estimated population mean and standard deviation at time t, and $\Phi^{-1}$ is the inverse of the standard normal cumulative distribution function, $\Phi^{-1}(\tau)$ therefore denotes the z-score that corresponds to the $\tau$ th percentile. Equation (2.5) represents the procedure of converting z-scores to percentile values in the normal distribution.

For many types of measurement data the assumption of normality does not hold. Various transformation techniques have been proposed to correct for non-normality, though it is still doubtful whether any one of them could yield normality over the range of z-scores of interest. Figure 2.1, for example, is based on the LMS method proposed by

Cole (1988) which transforms skewed data to normality. Nonparametric quantile regression provides distribution-free estimation of unconditional percentiles (Wei, 2004).

*2.4.2 Conditional Growth Chart*

Equation (2.5) shows that, although the growth chart is called "unconditional", its percentiles are estimated conditioning on time or age. If, besides time, other conditioning variables are included such as the previous measurements of the individuals, then the method is called *conditional* growth chart. The conditional growth chart is usually based upon longitudinal data sets. The unconditional growth chart can only be used to determine whether a subject's measurement at a particular time point seems "abnormal" or not, which is not necessarily a useful indicator of true underlying disorders. With regard to both health and education, children who stay well within the normal range of a certain statistic sometimes follow quite distinct developmental paths. For example, a child whose condition steadily deteriorates may have acceptable measurements for many years before she finally drops out of the normal range. On the other hand, those who are below the standard may be catching up. In general, it takes time for a well-functioning child to descend to a problematic level, even if she has evidently displayed such tendency. Similarly, a child with unsatisfactory measurement needs time to reach the normal range no matter how fast she is growing. As educational/medical researchers, policy makers, and practitioners, we do not want to wait until the signal is clear to decide if this growth story has been a failure or a success. The conditional percentiles produced from the conditional growth chart method helps to recognize unusual growth trends before it is too late.

Since conditional percentiles are estimated given at least two covariates, the conditional growth chart model does not have a standard two-dimensional chart form like figure 1. Cole (1994) presents such a model using a longitudinal French children's height data based on the familiar ordinary least square regression:

$$H_{i,t} = b_t H_{i,t-1} + c_t + error \qquad (2.6)$$

where $H_{i,t}$ is height measured for child *i* at age $t$, $H_{i,t-1}$ is height measured for the same child one year earlier, $b_t$ is the regression coefficient which varies from age to age, $c_t$ is an age-dependent intercept, and the error term is assumed to be distributed $N(0, \sigma_t^2)$ for all subjects. If the assumption of error normality holds, then $H_{i,t}|H_{i,t-1}$ also has a normal distribution: $N(b_t H_{i,t-1} + c_t, \sigma_t^2)$. Thus the conditional percentile of $H_{i,t}$ given $H_{i,t-1}$ could be estimated by substituting the mean and standard deviation of this distribution into equation (2.5). If the assumption of error normality is problematic, in other words, if height distribution is skewed, Cole (1994) demonstrates transformation techniques converting skewed distribution to normality, and equation (2.5) can be applied.

The conditional growth chart produced from model (2.6) consists of what Cole (1994) calls the "median conditional velocity centile curves". The conditional velocity is defined by re-arranging (2.6): $H_{i,t} - b_t H_{i,t-1} = c_t + error$. In other words, it is the average growth occurred in a year shared by everyone in a particular age group plus error, therefore the conditional velocity is not dependent on $H_{i,t-1}$, and is distributed as $N(c_t, \sigma_t^2)$. The z-score of the conditional velocity distribution is:

$$Z = \frac{H_{i,t} - b_t H_{i,t-1} - c_t}{\sigma_t}$$

For a child starting at height $H_{i,t-1}$ and growing along the $\tau$ th conditional velocity percentile, height at age t is predicted by:

$$\hat{H}_{i,t} = \hat{b}_t H_{i,t-1} + \hat{c}_t + \hat{\sigma}_t Z_\tau \qquad (2.7)$$

where $Z_\tau$ is the z-score corresponding to percentile $\tau$ , and $\hat{b}_t$ , $\hat{c}_t$ , and $\hat{\sigma}_t$ are estimates of the parameters in (2.6).

For a given starting height $H_s$ and fixing the percentile to be the median, heights in subsequent years can all be estimated by (2.7). A curve smoothing $H_s$ and these fitted heights of subsequent years is called the "median conditional velocity centile curve". The conditional growth chart consisting of such curves looks very similar to the unconditional growth chart of length presented in figure 1, with height plotted against age, except that each curve in the conditional growth chart represents a growth trajectory starting from a given $H_0$ value and follows the median conditional velocity in every year. An individual's measurements along the years can be compared against the curves according to her height at year 0, and it would be easy to tell whether she has been growing at a rate higher or lower than the median velocity. This kind of conditional growth chart only contains a small part of the information provided by model (2.6). Many different types of charts can be produced from the same model to visualize different aspects of the data (Cole, 1994).

Model (2.6) is based on linearity assumption and a strong distributional assumption. In recent developments of the growth chart method, various authors such as Gannoun et

al. (2002) and Carey et al. (2004) apply quantile regression to the estimation of conditional percentiles. This technique does not require normality of the data, and will be the basis of the methodology presented in chapter 3. The next section gives a brief introduction to the assumptions, as well as the estimation and inference procedures of quantile regression.

*2.5 Introduction to Quantile Regression*

Before Koenker and Bassett (1978) formally proposed quantile regression, Bhattacharya (1963) called it "an analog of regression analysis". Indeed, quantile regression and OLS regression essentially aim at the same goals—studying the relationship between dependent and independent variables through the conditional distribution of the former given the latter. The major difference is that OLS regression estimates the conditional mean of the dependent variable given the predictors while quantile regression concerns the conditional quantiles of the dependent variable given the predictors. The former implicitly assumes that the covariates exert a pure location shift effect on the outcome variable. When the predictors affect parameters of the conditional distribution of the dependent variable other than the mean, OLS regression will be inadequate. For example, when regressing students' current year test scores on their previous year's scores it is conceivable that students of different academic levels last year may have different distributions of scores this year—the low-achievers may have wider distributions of scores this year while the high-achievers, due to the ceiling effect, may have less variability in their distributions this year. In this case, the regressor not only affects the mean, but also the variance of the outcome, which is not captured by the OLS

model. Quantile regression, on the other hand, summarize the conditional distributions of the outcome with various conditional quantiles and thus allows a full characterization of the relationship between the outcome and the predictors.

To illustrate this difference, we present the summary plots of an OLS regression and a QR in Figure 2.2. Both regressions have only one independent variable. The data used in the two regressions are simulated with heteroscedastic error. The black line in figure 2.2 (a) represents the conditional mean of *Y* given the respective values of *X.* The slope of the line is the estimated coefficient of *X* in the OLS regression. The lines in figure 2.2 (b) represent the conditional quantiles of *Y.* From the lowest line to the highest line in the counterclockwise direction, the five lines correspond to the $10^{th}$, $25^{th}$, $50^{th}$, $75^{th}$, and $90^{th}$ quantiles of *Y* respectively given the values of *X.* The slope of each conditional quantile line is the estimated coefficient of *X* in the QR model specific to that quantile. The figures show that when the predictor is related not only to the mean but also to the shape of the conditional distribution of the outcome variable, the OLS regression can be a poor fit to the data, and the QR approach provides a better fit while increasing the number of parameters. In the following paragraphs, we introduce the ways of solving for the sample median and other sample quantiles, which provides a starting point of understanding how quantile regression works.

*2.5.1 The Estimation of Sample Quantiles*

Let *Y* be a random variable, and let $y_1, \ldots, y_n$ be a random sample from the distribution of *Y*. The sample median of $y_i$ is the solution to:

$$\min_{\xi \in R} \sum_{i=1}^{n} |y_i - \xi|$$

In other words, it is the value of $\xi$ that minimizes the sum of absolute difference between $y_i$ and $\xi$.

In order to get the $100\tau$ th sample quantile $(\tau \in (0,1))$, the absolute deviance in the above equation need to be weighted by $\tau$ or $1-\tau$ depending on the sign of the deviance (Wei, 2004). It can be shown that the $100\tau$ th sample quantile is given by

$$\min_{\xi \in R} \sum_{i=1}^{n} \rho_\tau(y_i - \xi) \quad \text{where} \quad \rho_\tau(u_i) = u_i(\tau - I(u_i < 0)) = \begin{cases} \tau u_i & u_i \geqslant 0 \\ (\tau - 1) u_i & u_i < 0 \end{cases} \qquad (2.8)$$

*2.5.2 The Estimation of Quantile Regression*

Now suppose we are interested in the relationship between *Y* and an independent variable *X*. Both OLS regression and quantile regression start from the linear model: $y_i = x_i^T \beta + e_i$, for $i = 1, \ldots, n$, where $\beta$ is the coefficient of the independent variable and $e_i$ is the error term. OLS regression assumes that $E(e_i | x_i) = 0$, which means that the conditional mean of $y_i$ has a linear relationship with $x_i$ : $E(y_i | x_i) = x_i^T \beta$. Besides, the errors are also assumed to be independent of each other and have an identical normal distribution. Quantile regression, on the other hand, requires that the $100\tau$ th quantile of $e_i$ given $x_i$ is 0, written as $Q_\tau(e_i | x_i) = 0$, which implies that the conditional $100\tau$ th quantile of $y_i$ is linear in $x_i$ : $Q_\tau(y_i | x_i) = x_i^T \beta$. No additional distributional assumptions are made for QR.

The estimation of the conditional quantiles is a simple extension of (2.8)—substituting $x_i^T \beta$ for $\xi$. Thus the coefficient $\beta$ in quantile regression are estimated by minimizing the following loss function with respect to $\beta$ :

$$R(\beta)=\sum_{i=1}^{n}\rho_{\tau}(y_i-x_i^T\beta) \qquad (2.9)$$

By definition of $\rho_{\tau}$ in (2.8), the loss function (2.9) is piecewise differentiable, i.e. it is only differentiable where $u_i=y_i-x_i^T\beta\neq 0$ . This means that the loss function (2.9) cannot be minimized by simply differentiating the loss function with respect to $\beta$ and setting the derivative equal to 0 as in the estimation of OLS regression. As a result, the estimators of quantile regression parameters do not have closed form expressions. Various ingenious methods were proposed for solving this kind of minimization problem in the regression setting where only one predictor is included, but it was not until the adoption of linear programming that the estimation of quantile regressions with multiple predictors become generally feasible. Koenker (2005) describes the simplex algorithm of linear programming which efficiently searches for the solution to the minimization of $R(\beta)$ . The solution is unique under some mild conditions about the continuity and density of *X* and *Y* (Koenker and Bassett, 1978). The computational details are not addressed in this thesis.

*2.5.3 The properties of Quantile Regression Estimators*

The properties of regression estimators depend largely on how they are estimated. The OLS coefficients are estimated by minimizing squared deviations while the quantile regression coefficients are obtained by minimizing weighted absolute deviations. It has been a long-lasting debate as to which minimization produces more superior estimates in terms of finite-sample and large-sample properties (Koenker and Bassett, 1978).

The Gauss-Markov theorem summarizes the well-known finite-sample properties

of OLS regression estimates—in a linear model in which the errors are independently and identically distributed (*i.i.d.*), and independent variables are not correlated with each other, the least-squares estimates have minimum variance among all unbiased estimates that are linear combination of the *Y*'s. (The OLS estimates of parameter $\beta$ can be seen as a linear combination of the dependent variable *Y,* since $\hat{\beta}=(X^T X)^{-1} X^T Y$ ). Even though this property does not rely on normality of data, OLS regression does need the normality assumption for standard error estimation in finite samples. And with normality, the least squares estimator is also most efficient (i.e. has least variance) among all possible unbiased estimators according to the Rao-Cramer inequality. Without the normality assumption, the OLS estimates still have nice large-sample properties provided that the Gauss-Markov theorem conditions hold, and that the data matrix of the independent variables *X* (also called the design matrix) satisfies the following requirement:

$$\lim_{n\to\infty}(1/n)X^T X=Q \quad \text{where} \quad Q \quad \text{is a positive definite matrix} \qquad (2.10)$$

In equation (2.10), *n* denotes sample size. If these conditions are satisfied, the OLS estimators are consistent, i.e. they converge in probability to the true values of the parameters. It can also be shown that the estimators have an asymptotic normal distribution and that the convergence rate is sufficiently fast (see, for example, Greene, 2000).

There is no well-established finite-sample theory for quantile regression similar to the Gauss-Markov theorem, though the large-sample properties of QR estimators have been studied by many statisticians (see, for example, Koenker and Machado, 1999; and

Koenker and Xiao, 2002). We take comfort in the fact that, as Koenker (2005) points out, even the OLS regression has to rely upon asymptotic approximations as soon as the idealized Gaussian conditions are violated. The sufficient and necessary conditions for the consistency of QR estimators generally include two parts. First, the cumulative distribution function of the conditional distribution of *Y* is absolutely continuous and there is adequate amount of mass at or near the specific quantiles. In other words, the density in the neighborhood of the quantiles should not be zero and should not explode, either. Second, condition (2.10) holds. When these requirements are satisfied, QR estimators are asymptotically normal and converge to their true values at the same rate as the OLS estimators. Note that the residuals are still assumed to be independently distributed but not required to be identically distributed for these asymptotic results (Koenker, 2005).

We need to be aware, however, that QR estimators for different quantiles may not have the same properties. Koenker (2005, p. 120) points out that while estimators of linear QR are consistent across different quantiles under mild conditions, the rates at which they converge to their true values depend crucially on the behavior of the conditional distribution of the outcome given the predictors near the quantile being estimated. With insufficient data in the neighborhood of a given quantile of the conditional distribution, the estimators for that quantile would exhibit slower rate of convergence. The asymptotic variances of QR estimators also vary across quantiles. They depend, again, on the density in the neighborhood of the quantiles, and they are proportional to the quantity $\tau(1-\tau)$ which is maximized at the median $\tau=0.5$ . This

means that holding the densities constant, the QR estimators would be more precise for the tails of the conditional distributions, but this effect is often dominated by the density effect which increase the standard errors of the estimates in regions of low density.

The Gauss-Markov theorem states that the least-squares estimates have minimum variance among all unbiased estimates that are linear combination of the *Y*'s. Outside of this league (for example, when compared with some slightly biased estimates), however, and especially when the distribution of the data is not Gaussian, the least squares estimates may not appear so favorable. Koenker and Bassett (1978) show that, in terms of efficiency, the QR estimators has good properties for a wide variety of distributions and generally out-performs the OLS estimator in non-Gaussian cases by a large margin.

Besides the efficiency properties, there are other advantages of quantile regression estimators over OLS estimators. Least squares estimators are especially sensitive to distributions with longer tails (i.e. the presence of outliers), while absolute-deviation estimators are resistant to outlier contamination. Koenker (2005) points out that as long as the signs of the residuals are not changed, the quantile regression coefficient estimates remain the same when we perturb the dependent variable observations $y_i$ . This can be explained by means of equation (2.9). Since $y_i - x_i^T \beta$ is contained in an indicator function, changing the values of $y_i$ does not change the loss function as long as the sign of the residual stays the same. Outliers in the independent variable $x_i$ 's can still alter the quantile estimates though and need to be watched. The quantile regression estimators also deals with heteroscedasticity with ease, which is a property not shared by the OLS estimators. As figure 2.2 (b) shows, there are different coefficient estimates for different

$\tau$ values and the the slopes are allowed to be nonparallel to each other, the spread of the conditional quantiles are expected to be wider when the $y_i$ 's have larger dispersion, thus the quantile estimates naturally capture the changing variance of the residuals.

*2.5.4 Hypothesis Testing For Quantile Regression*

Hypothesis testing in quantile regression can be done using several common procedures including the likelihood ratio test, the Wald test, and the rank score test (Koenker, 2005). All these tests involve the estimation of the asymptotic covariance matrix for $\hat{\beta}^{(\tau)}$ , which is the regression coefficient at a given 100$\tau$th quantile. In order to estimate this covariance matrix, it is necessary to evaluate the probability density of the error term at the 100$\tau$th quantile of the error distribution. Statisticians have established relatively non-controversial ways of estimating error density in the case that errors are *i.i.d.* (Koenker and Machado, 1999; Koenker, 2005). When errors are non-*i.i.d.*, the problem of estimating coefficient standard errors becomes more complex. Not only does the 100$\tau$th quantile density have to be evaluated for the errors of each subject $e_i^{(\tau)}$ , but these estimated densities also have to be weighted. The number of parameters to be estimated increase dramatically in this process and other distributional assumptions are invariably made. The bootstrap method provides a way of estimating coefficient standard errors regardless of the error density, and makes no assumption about the distribution of the response variable or the error terms. Different studies show that hypothesis tests based on bootstrap and asymptotic standard errors usually yield the same results (Koenker, 2005; Hao and Naiman, 2007).

The bootstrap method is a special type of Monte-Carlo simulation which

approximates the sampling distribution of a parameter estimate by drawing large number of samples from a known distribution and calculating the parameter estimate for each sample (Efron, 1979). With the bootstrap approach to estimating coefficient standard errors, there are different choices of resampling plans. A typical one is to sample with replacement from the observed residuals of the model, which again requires the errors to be *i.i.d.* for the bootstrap method to produce acceptable approximation of the true distribution of parameters (Efron, 1982). Koenker (2005) presents a plan which resamples directly from observed predictors and outcomes instead of the residuals, and then estimates new coefficients with the sampled predictor and outcome values. Since the independent and dependent variable values are randomly sampled with replacement from the observations, some subjects may be sampled multiple times while other subjects may not be included. Thus each of the resamples will randomly depart from the original dataset and so will the corresponding estimates of the parameters. A large number of cycles like this would produce a sampling distribution of the coefficients from which we could obtain standard errors, confidence intervals and perform hypothesis testing. This method proves to be a valid inference tool even if the errors are independent but not identically distributed (Koenker, 2005).

*2.5.5 Goodness-of-Fit of Quantile Regression*

A test of goodness-of-fit is closely related to hypothesis testing. If a hypothesis test result shows that the covariates jointly have no effect on the response, then obviously the model is a bad fit. Yet a separate indicator still has to be developed to signify how well the overall model explains the data and how close the fitted values are to the actual

observed outcomes. Koenker and Machado (1999) propose a convenient statistic for quantile regression that can be used just like the R square in OLS regressions. Going back to the loss function of quantile regression (2.9), let $\hat{V}(\tau)$ denote the minimized sum of weighted absolute residuals under the full model:

$$\hat{V}(\tau) = min \sum_{i=1}^{n} \rho_{\tau}(y_i - x_i^T \beta)$$

and let $\tilde{V}(\tau)$ denote the minimized sum of weighted absolute residuals when the model only includes an intercept parameter:

$$\tilde{V}(\tau) = min \sum_{i=1}^{n} \rho_{\tau}(y_i - \beta_0)$$

The goodness-of-fit criterion is defined as:

$$R^1(\tau) = 1 - \frac{\hat{V}(\tau)}{\tilde{V}(\tau)} \qquad (2.11)$$

$R^1(\tau)$ is the natural analog of $R^2$ in OLS in the sense that if we replace $\hat{V}(\tau)$ and $\tilde{V}(\tau)$ with corresponding minimized residual sum of squares, equation (2.11) would produce $R^2$ instead of $R^1(\tau)$. Like $R^2$, $R^1(\tau)$ lies between 0 and 1. Lower values of this statistic suggests lack of fit. The value of $R^1(\tau)$ is easy to compute and easily interpretable (just like that of $R^2$), but since it is $\tau$ specific, it only shows how closely the model fits the data around the $\tau$ th quantile. Various graphs of goodness-of-fit have been designed which demonstrate model fit across quantiles. One of them will be applied in the data analysis chapter.

# Chapter 3 Methodology

This chapter introduces the methodologies that are used for data analysis in this thesis. We begin with a description of the data consisting of four years of longitudinal test scores from a state. The tests are also briefly described. Next, the quantile regression models are presented, which are used to estimate conditional percentiles for individual students. The model assumptions and specifications are also explained. After this, we proceed to investigate the impact of measurement errors on the estimation of conditional percentiles. A few methods of adjusting for measurement errors are introduced, and we focus on the application of the simulation-extrapolation (SIMEX) method on QR models. This chapter is concluded by a brief exploration of the Bayesian QR model which accommodates random effects.

## *3.1 Data*

### *3.1.1 A Brief Introduction to the State Assessment Program*

The data used in this thesis represent four consecutive years of the scale scores of a state assessment program in reading. Due to confidentiality reasons, the state is not

identified by name and the technical reports of the state assessment are not cited below. The tests are developed to measure the "content standards" determined by the state Department of Education, and are administered in the spring every year. Student results are reported statewide in terms of scale scores and performance levels, and these results are used in the annual school accountability reports of the state. The performance level cut scores were adopted by the state Board of Education, based on the recommendations of standard setting committees composed of qualified educators. The reading and writing assessments are administered in grades 3-10 every year.

According to the annual technical report of the testing program from 2003 to 2006, the test designers use item response theory for test scaling. Specifically, the three-parameter logistic model (Lord and Novick, 1968) is used for the analysis of selected-response items, and the two-parameter partial credit model (Yen, 1993) is used for the analysis of constructed-response items. The item parameters are estimated with these models based on a sample data set of the state's student population. Student scale scores are calculated using the item parameters. Model fit statistics are reported for each grade in each year. Very few items exhibit lack-of-fit, and those items that have poor statistical performance are removed from scale score calculations.

Even though the models we propose in this thesis do not require vertical linking from test scores, the scale scores of the tests in our dataset are vertically linked across different grades. The vertical scale for reading was established in 2001 using the Stocking and Lord (1983) procedures. Scale scores of the same grade in different years are also horizontally equated.

The reading scores range from 150 to 1000 across the eight grades. For each year and each grade, conditional standard errors of measurement at different scale scores are reported. The scale scores are assumed to be on an interval scale for the analyses done in this thesis.

*3.1.2 Data Description*

The data set includes the results of the statewide administrations of the reading test in 2003-06 for students in grades 7-10. Test results of the same students in different years are linked through the same ID number. For the four years altogether, there are 54,625 students in the data . Table 3.1 presents the percentages of male students, each ethnic group, student proficient and partially proficient, and repeaters for grades 7-10 respectively. Besides scale scores, gender, and ethnicity, the data set contains some additional information on students' date of birth, school ID, and school district ID.

Table 3.1 shows that the cohort loses about 10,000 students which is almost 1/5 of its original size by the end of the $10^{th}$ grade. A small percentage of the loss is due to retention, but for most of the missing students, the reason is unclear. We also notice that the percentages of Hispanic and Black students in the cohort become smaller in grade 10, while the percentages of White and Asian students increase noticeably. The timing of these demographic changes suggests that there may be a jump of minority drop-out rates in the $10^{th}$ grade. There is also a jump in retention rate just before the $10^{th}$ grade—1097 students had to repeat the $9^{th}$ grade in 2006, more than doubling the number of repeaters in the previous grades. The percentages of students reaching proficiency and partial proficiency, however, steadily increase over the years. The state's AYP targets use the cut

scores between non-proficiency and partial proficiency as the most crucial cut scores. We see in Table 3.1 that about 92 percent of the 10th graders reached partial proficiency, a quite encouraging achievement if we ignore the somewhat dramatic increase of missing students.

We also present the distributions of the scores in each grade in Figure 3.1. All four distributions are left-skewed, indicating higher density in higher scores. Another issue to be noted is that there is a small jump of density in the far left side of each distribution. These are the students who got the Lowest Obtainable Scale Scores (LOSS). The LOSS are often assigned for administrative reasons instead of based on the real assessments. Therefore these scores contain large amount of errors and their distributions look quite unnatural. We decide to keep these scores in the data set despite the large errors for two reasons. First, in a practical accountability system, students with the lowest scores as well as their parents and teachers will expect a report on their growth percentiles just as those with higher scores do. It will be demoralizing for them if their scores are deleted from the system and their academic growth ignored. Second, quantile regression is generally robust to outliers which means that the inclusion of these scores will not lead to big estimation biases. We do recommend, however, that while the growth percentiles of these students are estimated and reported, the specific administrative reasons for their LOSS are also explained to all stakeholders.

### *3.2 A Local Linear Quantile Regression Model*

#### *3.2.1 Model and Notations*

The basic model to be considered in this thesis is a simple linear QR model. For a

cohort of *n* students, denote by $Y_{i,t}$ the score of the $i$ th student in grade $t$ . The quantile regression function is specified essentially the same as equation (2.6):

$$Y_{i,t}=\alpha_t^{(\tau)}+\beta_t^{(\tau)}Y_{i,t-1}+e_{i,t}^{(\tau)} \qquad i=1,\ldots,n \qquad t=8,9,10 \qquad \tau\in(0,1)$$

(3.1) where $\alpha_t^{(\tau)}$ and $\beta_t^{(\tau)}$ are grade-specific and tau-specific intercept and slope coefficients, and $e_{i,t}^{(\tau)}$ is assumed to be independently distributed. Theoretically, for each individual student, there are infinite numbers of τ's which lead to infinite numbers of model parameter values $\alpha_t^{(\tau)}$ and $\beta_t^{(\tau)}$ and model residuals $e_{i,t}^{(\tau)}$ . The τ-value that corresponds to the smallest model residual among them is the student's growth percentile.

Equation (3.1) denotes a grade-specific model, also known as a local model. Because the model is a local one, and because prior scores are used as conditioning variables, scale scores used to estimate the model are generally not required to be vertically linked.

The subscript *t* is intended to denote time. In this context, time can be represented in two dimensions—grade and year. We now briefly explain why we choose to let *t* represent grade instead of year. Since the model is used to address a particular grade of a certain cohort, whether the subscript *t* is used to denote grade or year would not matter for most subjects in the sample—for students who are never retained, each grade corresponds to a unique year and vice versa. For those who have been retained, however, the exact meaning of *t* does matter. Suppose student *i* is repeating grade 8 this year, and suppose student *k* is also at grade 8 this year but was not retained. If *t* is used to denote year, $Y_{i,t-1}$ would be student *i*'s 8th grade score from last year, while $Y_{k,t-1}$ is student *k*'s 7th grade score from last year. The growth percentile estimated for student *i* based on

this model would mean "the percentile of *i* among the present 8th graders whose 7th grade scores from last year are the same as student *i*'s 8th grade score from last year". Even if the scores are vertically linked, this quantity is still quite hard to interpret. On the other hand, if *t* is used to denote grade, which is what we choose for model (3.1), $Y_{i,t-1}$ would be student *i*'s 7th grade score from two years ago, and the meaning of $Y_{k,t-1}$ remains the same. The growth percentile estimated for student *i* would mean "the percentile of *i* among the present 8th graders whose 7th grade scores from last year are the same as student *i*'s 7th grade score from two years ago". This quantity is easier to understand and more useful in the diagnostic process than the former.

If we let $Q^{(\tau)}$ to denote the τth quantile, equation (3.1) can also be expressed as

$$Q^{(\tau)}(Y_{i,t}|Y_{i,t-1}) = \alpha_t^{(\tau)} + \beta_t^{(\tau)} Y_{i,t-1} \qquad i = 1,\ldots,n \quad t = 8,9,10 \qquad (3.2)$$

The expected conditional $100\tau$ th percentile of the current score distribution given past scores is obtained by simply plugging the coefficient estimates into the right-hand side of (3.2). Compared with the method for obtaining conditional percentiles based on OLS regression, the QRM approach is much easier and saves the fine-tuning procedures in the cases of non-normality. For a sufficiently large set of values of $\tau$, the estimated values of $\hat{Q}^{(\tau)}(Y_{i,t}|Y_{i,t-1})$ summarize the complete distribution of student achievement in grade $t$ given their scores in the previous grade.

To account for sampling variability, it is important to construct confidence intervals for the estimated conditional percentiles. The bootstrap method introduced in section 2.5.4 is suitable for this purpose. Besides constructing confidence intervals and testing the hypotheses that model parameters are equal to zero, the bootstrap method can also be

used to test whether the coefficients of the predictor(s) vary significantly across different quantiles (Hao and Naiman, 2007). This test provides important information about the difference in growth patterns of students at different levels of achievement. For goodness-of-fit assessment, the method of equation (2.11) can be used, and we will also employ the worm plots proposed by Buuren (2007).

Model (3.1) has some advantages over the OLS model for the purpose of educational diagnosis as illustrated in section 2.5.3. To sum up briefly, both the estimates of the conditional percentiles in (3.1) and the inferences based on these estimates are distribution free. Quantile regression estimators may be more efficient than OLS estimators when the error terms are non-normal, they are robust to outliers, and they naturally accommodate heteroscedasticity. These are convenient properties for studying educational data that usually have outliers and non-constant variance. The fact that QRM can characterize the entire conditional distribution of the outcome is important because we aim to make a diagnosis for every student, especially those in the lower tail of the distribution, and not just the average student. Before applying the QRM results, however, densities of score distributions need to be checked to ensure that QR estimators have adequate properties at the specific quantiles.

*3.2.2 Choice of Covariates*

Model (3.1) is a lag-1 model, but it is certainly possible to include more past scores as predictors to form lag-2 or lag-3 models etc. The lag-2 models are specified as

$$Y_{i,t}=\alpha_t^{(\tau)}+\beta_{t,1}^{(\tau)}Y_{i,t-1}+\beta_{t,2}^{(\tau)}Y_{i,t-2}+e_{i,t}^{(\tau)} \qquad i=1,\ldots,n \qquad t=9,10 \qquad \tau\in(0,1)$$

(3.3) where the parameters are interpreted similarly as in (3.1). It is a problem of

balance to choose the number of past scores to be included in the QR model. On the one hand, including more past scores may significantly improve model fit and increase the power of diagnosis, since student growth percentiles are estimated within more homogeneous groups. An F test in the analysis of deviance can be used to determine whether it is the case for our data. On the other hand, we also have to consider the potentially serious disadvantage of multicollinearity. If the regressors are highly correlated with each other, which the past scores very likely are, the over-fitted model could produce inconsistent estimates (Weisberg, 2005). In this thesis, we primarily focus on the lag-1 models since they are the simplest ones for construction of the conditional growth chart. They also have larger sample sizes than the lag-2 or lag-3 models, because they only require two years of scores. Lag-2 models are also estimated in chapter 4, and results are used for school comparison in the next chapter.

One potential problem with the lag-1 models is that the student growth percentiles estimated based on these models may not be completely comparable. Let $\hat{P}_{g8|g7}$ denote the estimated growth percentiles of grade 8 conditioning on scores of grade 7, and let $\hat{P}_{g9|g8}$ and $\hat{P}_{g10|g9}$ have similar interpretations. Suppose $\hat{P}_{g8|g7} < \hat{P}_{g9|g8}$ for a specific student, we are not completely sure what it means, since the two quantities are conditioned on different groups of students. In order to estimate comparable growth percentiles we specify the "common condition" models:

$$Y_{i,t} = \alpha_t^{(\tau)} + \beta_t^{(\tau)} Y_{i,min(t-1)} + e_{i,t}^{(\tau)} \qquad i = 1,\ldots,n \qquad t = 8,9,10 \qquad \tau \in (0,1)$$

(3.4) where the parameters are interpreted similarly as in (3.1). In model (3.4), scores of grade 8, 9, and 10 are all regressed on grade 7 scores, thus producing growth

percentiles $\hat{P}_{g8|g7}$ , $\hat{P}_{g9|g7}$ and $\hat{P}_{g10|g7}$ . For any given student, the three quantities are conditioned on the same group of students and are therefore completely comparable. If, for example, $\hat{P}_{g8|g7} < \hat{P}_{g9|g7}$ for a specific student, we know the student made more progress in grade 9 than in grade 8 within the group which started from the same position in grade 7. Models of equation (3.4) are also estimated in chapter 4, but since these models usually suffer from lack-of-fit compared with the lag-1 or lag-2 models, their results are not used extensively in the analysis of student growth.

Past academic achievement is by no means the only important predictor of a student's current achievement. Theoretically many other predictors can and should be included in (3.1) to improve model fit. Among the observed variables in our dataset, for example, ethnicity might significantly correlate with current scores even when past scores are held constant. In practice, whether to include additional covariates in (3.1) depends on the objectives of the model.

For diagnosis of individual students, we believe that prior test scores are the only relevant conditioning variables as explained earlier. Including other covariates such as ethnicity would be equivalent to setting different academic standards for different demographic groups.

When the model is not used to compute conditional percentiles, including other relevant covariates may be meaningful. For example, adding ethnicity as a fixed effect intercept and testing its significance is a relatively sophisticated way of detecting and quantifying an achievement gap. With a simple comparison of mean scores in different ethnic groups at grade 8, for example, the achievement gap can be detected, but it is the

estimation of a cumulative gap, i.e. it indicates the achievement gap between ethnic groups *by* grade 8. There is no information about when this gap started to form and when it widened or narrowed. On the other hand, using the grade 8 model defined in (3.1), because scores from the previous grade are held constant, the estimated fixed effect for ethnicity is the gap of academic growth that occurred *in* grade 8. Moreover, if the slope coefficient is also allowed to vary with ethnicity, we can compare not only the average growth in grade 8 between different ethnic groups but also the distributions of growth in grade 8 between ethnic groups. This is very useful information for educational administration, research, and policymaking.

Another aspect of model specification that needs to be discussed is linearity. A non-linear QR model usually fits the data better than a linear one, especially considering the existence of outliers discussed in figure 3.1. Wei et al. (2006) model the nonlinear relationship between age and height by using a cubic B-spline basis as predictors instead of age itself, and achieved excellent model fit. In fact, the QR models can be made arbitrarily close to the data if we are willing to use a more liberal B-spline basis. In doing so, however, we may be modeling sampling and measurement errors as well as the true trend of data. It is not undesirable if our sole objective is to produce accurate descriptive statistics, but if we also want to make inferences based on the models and use the results to make projections for later cohorts, it is better to have models that are less sensitive to sampling noise. This is not to say that the linear models are always more generalizable than the nonlinear ones. We choose to focus on linear models in this thesis because the data do not demonstrate systematic departure from linearity (see figure 4.1), and also

because linear models are easier to interpret and provide a starting point to understand the growth chart method.

### *3.3 Adjusting for Measurement Error*

Measurement error is associated with all test scores. The models proposed in this chapter use test scores as both predictor and response variables. It was mentioned in section 3.1.1 that standard errors of measurement (SEM) are estimated in all grade levels in this state assessment program. In fact, they are reported at a grid of scale scores with a spacing of 25 points. Since the SEM vary at different scale scores, they are called the conditional SEM. Figure 3.2 plots the reported conditional SEM against scale scores in grades 7, 8, 9, and 10. The plots show that, in each grade, the amount of measurement error within some ranges of scale scores are non-negligible. How do the measurement errors impact the results of the QR models? And how can we minimize this effect?

In this section we begin by explaining the effect of measurement error on the estimation of simple OLS models, and then extends these results to QR models. We proceed to examine some methods that correct for the estimation bias caused by measurement errors. We finally focus on the simulation-extrapolation (SIMEX) method that corrects for measurement error induced bias for both the OLS and QR models.

#### *3.3.1 Types of Measurement Errors*

In both classical test theory and item response theory, a student's true score is defined as the expected value of the measurement she has obtained, the expectation being taken over the hypothetical set of parallel or “nominally parallel” measurements (Lord and Novick, 1968, p. 173). The measurement error is defined as the discrepancy between

the observed score and the true score. Two important properties of the measurement error follow from this definition. First, the errors are additive, meaning the observed score is the sum of the true score and error. Additive and non-additive errors have very different effects on analysis results and need different treatments. Second, the expected value of the measurement error is zero. Carroll et al. (2006) call it unbiased measurement error. Besides these properties, Lord and Novick (1968, p. 493, p. 509) also concluded that a sufficient condition for the measurement errors on a test to be distributed independently of true score on the same test is that the observed scores are not artificially bounded.

*3.3.2 Measurement Error in the Dependent Variable*

Additive unbiased measurement error in the dependent variable is usually considered to be a minor problem in regression analysis, since the error simply gets absorbed into the regression residuals as long as it is not correlated with the covariates (Ladd and Walsh, 2002; Abrevaya and Hausman, 2004). Carroll et al. (2006), after studying this type of error in detail, claims that additive, unbiased, and homoscedastic response measurement errors only increase variability of the estimated coefficients in linear and nonlinear regressions and thus renders statistical tests less powerful. No estimation bias is introduced by this type of measurement error. Moreover, they show that even if the response measurement error is heteroscedastic, which is likely to be the case in model (3.1), it only changes the form of the residual variance function and would not lead to additional bias if the model is specified to accommodate heteroscedastic residuals. These conclusions are drawn from OLS and generalized linear models, but are also readily extended to quantile regression models.

*3.3.3 Measurement Error in the Independent Variables*

When measurement error affects the predictor variable(s), the problem becomes more complex. Consider a simple OLS regression:

$$Y = \beta_0 + \beta_1 X + e \qquad (3.5)$$

where $X$ is the only variable measured with error, and $X = \dot{X} + v$ where $X$ is the observed variable, $\dot{X}$ represents the true value that is not observed, and $v$ is the measurement error. As mentioned earlier, Lord and Novick (1968) argue that measurement errors and true scores are assumed to be and usually are independent of each other, i.e. $Cov(\dot{X}, v) = 0$ . Therefore $Cov(X, v) = Cov(\dot{X} + v, v) = \sigma_v^2$ where $\sigma_v^2$ is the variance of the measurement error. In short, the observed scores are correlated with the measurement errors. Now consider the model that we are really interested in, i.e. the model that contains the true variable instead of the observed one:

$$Y = \dot{\beta}_0 + \dot{\beta}_1 \dot{X} + \dot{e} = \dot{\beta}_0 + \dot{\beta}_1 X + (\dot{e} - \dot{\beta}_1 v) \qquad (3.6)$$

where $\dot{\beta}_0$ and $\dot{\beta}_1$ are the intercept and slope coefficients in the measurement error free model, and $(\dot{e} - \dot{\beta}_1 v)$ is the composite residual of model (3.6). Since $X$ and $v$ are correlated with each other, it follows naturally that $X$ and $(\dot{e} - \dot{\beta}_1 v)$ are correlated with each other, too. In OLS models, the situation that one or all of the covariates are correlated with the model residuals is usually called endogeneity, and it causes bias and inconsistency in the estimates of all the model parameters, not just the coefficient of the endogenous variable. In other words, using error-contaminated observed scores instead of the true test scores as predictors leads to endogeneity in the OLS model, in which case the parameter estimates are biased.

The effect of covariate measurement error in simple OLS regression is a well-studied topic. Inconsistency of parameter estimation in these models can be directly calculated. Estimation bias in such models is called attenuation, which refers to the fact that slope estimates are biased in the direction of zero. In multiple regression, covariate measurement error leads to the same problem of endogeneity and cause estimates to be biased, but unlike simple OLS, the coefficients estimates in multiple regression are biased in unknown directions. Depending on the number of predictors measured with error and the extent of multicollinearity present in the predictors, the effects of measurement error may vary. Real relationships may be hidden, observed data may exhibit false relationships that are not present in error-free data, and even the signs of estimated coefficients may be different from those of the true coefficients (Fuller, 1987).

*3.3.4 Adjusting for Covariate Measurement Error—Instrumental Variable*

Various methods have been proposed to correct for measurement error induced bias in OLS models. One of the most established methods is the instrumental variable (IV) approach (Carroll et al., 2006; Wooldridge, 2002). In the context of model (3.3), an observable instrumental variable for the error contaminated $X$ would be a variable $Z$ that fulfills the following requirements. First, $Z$ must be uncorrelated with the model error $e$. Second, $Z$ must also be uncorrelated with the measurement error $v$. And third, $Z$ must be correlated with the true variable $\dot{X}$.

The IV technique is often implemented by the two-stage least-squares method. In the first stage, we regress the endogenous $X$ on the instrument $Z$ and other exogenous independent variables if there are any. Let the fitted values from this

regression be denoted as $\hat{X}$ . In the second stage, model (3.5) is estimated, except that the values of $X$ are replaced by $\hat{X}$ . The slope coefficients thus estimated will be consistent for the true coefficients. A proof of this consistency can be found in Wooldridge (2002). In this process, only a minor correction needs to be made on the sum of squared residuals in the second stage model for the estimation of the standard errors of the parameter estimates.

Ladd and Walsh (2002) provide a good example of adopting the IV method to adjust for measurement error contained in test scores. In their value-added model, students' fifth grade scores are predicted from their fourth grade scores, and the third grade scores from the same subjects are used as an instrumental variable for the error-contaminated fourth grade scores. This choice of IV, according to the authors, fulfills the conditions mentioned above, since the third grade scores are clearly correlated with the fourth grade scores, but are presumably independent of the measurement errors contained in the fourth grade scores. It is not demonstrated in the paper, however, whether 3rd grade scores are uncorrelated with the errors of the original model. It could be argued that 3rd grade scores do not provide any information about 5th grade score that 4th grade scores have not already provided, but this claim remains to be tested.

In our dataset, the only possible choice of instrumental variable is scores from earlier grades as in Ladd and Walsh (2002). The problem with this approach is that students with only two waves of scores will be dropped from the models. Considering that our goal is to make diagnosis for every individual student for whom we have more than one measurement, this is not an ideal method. Moreover, there is evidence in our

data that scores from non-consecutive grades are still correlated with each other, which means that scores from earlier grades may not fulfill the requirements for instrumental variable. Later in this section, we introduce a method to adjust for covariate measurement error that works for both OLS and QR models and does not create any more missing value problems than we already have.

*3.3.5 Measurement Error in Quantile Regression*

Compared with ordinary regression and generalized linear models, the effects of measurement error on QRM are much more poorly understood. Since the estimates of QRM parameters do not have closed form expressions, it is hard to derive inconsistency results due to measurement error in an analytic form. Chesher (2001) provides some insight into this problem by working out a second order Taylor expansion of the error contaminated conditional quantiles around the error-free conditional quantiles, and thus approximating the bias of parameter estimation. He shows that when the quantile regression model contains only one covariate, the slope estimate tends to be flattened by the covariate measurement error. In other words, the attenuation effect is also present in single-predictor quantile regression. In the case that quantile regression function is nonlinear, covariate measurement error would dampen the curvature of the error contaminated regression function relative to the error free function. Covariate measurement error also tends to interact with the preselected quantiles $\tau, (\tau \in [0,1])$ so that the error contaminated conditional quantile functions may not be parallel to each other even if the error free quantile functions are parallel.

To our knowledge, studies of the methods of correcting for measurement-error-

induced bias in QRM are scant. Chesher (2001) develops a method which greatly reduces estimation inconsistency for error contaminated linear quantile regression models based on his results of the approximation of the measurement-error-induced bias. But the method only works for regressions with a single covariate and parallel slopes across quantiles (i.e. the residuals of the model are homoscedastic). Schennach (2008) introduces the instrumental variable method for dealing with covariate measurement error in QRM. The author proves consistency for the IV estimators which follow the same rationale as for IV for OLS models.

*3.3.6 Adjusting for Covariate Measurement Error—The SIMEX Method*

Cook and Stefanski (1994) proposed a simulation-based method of estimating and reducing bias due to measurement error. The method, which is called simulation extrapolation (SIMEX), has since gained much popularity among statisticians and biostatisticians for its simplicity and generality (see, for example, Carroll et al., 2006; Carroll et al., 1999; Kuchenhoff, et al., 2006). It is ideally suited to models with additive measurement error in the covariates when the measurement error variance is known or can be reasonably well estimated.

The basic idea of the method is to add simulated additional measurement error with increasing variance to the original data in a resampling-like stage, identify a trend of measurement error-induced bias versus the variance of the added measurement error, and extrapolate the trend back to the point with no measurement error. This algorithm does not help us to derive an analytic form of the measurement error-induced bias, which is perhaps one of its drawbacks. But what makes the method powerful is its wide

applicability. The technique lends itself easily to measurement error models of the simplest form, and "because the method is completely general, it is also useful in applications when the particular model under consideration is novel and conventional approaches to estimation with the model have not been thoroughly studied and developed" (Cook and Stefanski, 1994). Considering that conditional standard errors of measurement are well estimated and routinely reported for standardized tests used in educational accountability systems, the SIMEX method seems to be a good fit for our data and our methodology.

*3.3.6.1 Basic rationale and algorithm of the SIMEX method*

To illustrate the SIMEX algorithm in detail we again consider model (3.5) where the observed covariate *X* is error-contaminated, and the measurement error variance is $\sigma_v^2$. Let $\theta$ denote the vector of the coefficients in the error-prone model, and let $\theta_{true}$ denote the vector of the true parameters in the error-free model, i.e. $\theta=(\beta_0, \beta_1)$, and $\theta_{true}=(\dot{\beta}_0, \dot{\beta}_1)$. The estimate of $\theta$ is denoted $\hat{\theta}_{naive}$, and the final result of the SIMEX algorithm, which is an estimate of $\theta_{true}$, is denoted $\hat{\theta}_{simex}$.

The procedure starts from the simulation stage where additional data sets are created with the predictor containing increasingly larger measurement error $(1+\lambda)\sigma_v^2$, where λ usually ranges between 0 and 2. Within each simulated data set, a $\hat{\theta}_{naive}$ is produced ignoring the existence of measurement errors. Theoretically, the bigger the measurement errors are, the farther away $\hat{\theta}_{naive}$ is from $\theta_{true}$. In fact, $\hat{\theta}_{naive}$ should be a function of λ. If we could extrapolate the value of this function to the point where

$\lambda = -1$ , this value would correspond to the measurement error variance $(1+\lambda)\sigma_v^2 = 0$ , and thus we obtain $\hat{\theta}_{simex}$ which is our estimate of the true parameters.

The implementation of the SIMEX algorithm is straightforward. First, choose a set of λ values. In our case, we choose $\lambda = 1/4, 2/4, \ldots, 8/4$ . For each value of $\lambda$ , create B datasets each of which contains a new error-prone predictor: $X(\lambda) = X + \sqrt{\lambda}\, U$ , where U is a randomly-generated variable from the distribution $N(0, \sigma_v^2)$ , or $N(0, \hat{\sigma}_v^2)$ when the variance of the measurement error is not known but estimated. For expositional ease, we only consider the case when $\sigma_v^2$ is known. It can be shown that the variance of $X(\lambda)$ conditional on the true values $\dot{X}$ would be $(1+\lambda)\sigma_v^2$ . The new variable $X(\lambda)$ is called a remeasurement of *X*, and the newly-created datasets with inflated measurement errors are called remeasured data.

With each of the B remeasured data sets, a $\hat{\theta}_{naive}(\lambda, b) \quad (b = 1, \ldots, B)$ is estimated from model (3.5) ignoring the measurement error, and for each value of $\lambda$ , a $\hat{\theta}_{naive}(\lambda)$ is obtained which is the sample mean of the B $\hat{\theta}_{naive}(\lambda, b)$ values. The purpose of using the sample mean for extrapolation instead of the individual naïve estimates is to reduce the sampling variation in the simulation process which may compound the measurement error variance and mask the bias trend.

At this point we have a series of values for $\hat{\theta}_{naive}(\lambda)$ which is a function of λ. Figure 3.3 (a) and (b) presents the dependence of $\hat{\theta}_{naive}(\lambda)$ on λ for a simple OLS regression and a simple quantile regression respectively. The average estimate at each λ value is generated with $B = 200$ . The naive intercept and slope estimates in both OLS

and QR models appear to have a strong relationship with λ.

Theoretically, an asymptotic true function $\theta(\lambda)$ of $\lambda$ exists which can be directly extrapolated back to the true parameters (Stefanski and Cook, 1995). In reality, this true function usually remains unknown and has to be approximated. There are many possible choices for the extrapolation function. The usual candidates include the linear extrapolant, the quadratic extrapolant, and the rational extrapolant (Carroll et al., 2006). The linear extrapolant function is defined as:

$$\hat{\theta}_{lin}(\lambda)=\gamma_1+\gamma_2\lambda \qquad (3.7)$$

The simple quadratic extrapolant function is:

$$\hat{\theta}_{Q}(\lambda)=\gamma_1+\gamma_2\lambda+\gamma_3\lambda^2 \qquad (3.8)$$

And the rational extrapolant function is:

$$\hat{\theta}_{RL}(\lambda)=\gamma_1+\frac{\gamma_2}{\gamma_3+\lambda} \qquad (3.9)$$

where $\gamma_1$ , $\gamma_2$ and $\gamma_3$ are the function parameters to be estimated. Functions (3.7) to (3.9) can be estimated as the usual linear and non-linear least-square models with the values of $\hat{\theta}(\lambda)$ and $\lambda$ being dependent and independent variables respectively. And the conventional tools of model diagnostics can be applied to assess model fit. Results from applications and simulations as well as theoretical analysis suggest that the SIMEX estimator derived from (3.9) usually contains less bias than those derived from (3.7) and (3.8), but the linear and quadratic extrapolants, being more conservative in correction for bias, may have less variability (Carroll et al., 2006).

Figure 3.4 plots $\hat{\theta}_{naive}(\lambda)$ against λ for the slope coefficient in a QR model at

$\tau = 0.25, 0.5, 0.75$ . As in figure 3.3, the average estimate at each λ value is generated with $B = 200$ . The "naive est" dot in each plot that lies at λ = 0 is the original naïve estimate of the slope before applying simulation and extrapolation. The straight line in each plot represents the linear extrapolation, and the "simex lin" dot that lies at λ = -1 is the SIMEX estimate of the slope based on linear extrapolation. Similarly, the curve in each plot represents the quadratic extrapolation, and the "simex quad" dot that lies at λ = -1 is the SIMEX estimate of the slope based on quadratic extrapolation. The true slope in all three plots is equal to 1. Figure 3.4 shows that the SIMEX estimate based on either linear or quadratic extrapolation is much closer to the true value than the original naïve estimate, and the quadratic extrapolation produces a less biased estimate than the linear extrapolation in each of the three cases.

*3.3.6.2 The SIMEX method in some special cases—heteroscedastic measurement error and multiple error-prone predictors*

The algorithm described above can easily accommodate heteroscedastic measurement error. Figure 3.2 shows that the standard errors of measurement in our data are not constant over different test score values, which means that instead of using $\sigma_v^2$ to denote the measurement error variance, we should denote the measurement error variance corresponding to each $x_i$ by $\sigma_{v_i}^2$ . The estimate of this quantity, $\hat{\sigma}_{v_i}^2$ , is reported for the test at each grade level at a grid of $x_i$ values as mentioned earlier. To accommodate this heterogeneity, a variable $u_i$ is randomly-generated from the distribution $N(0, \hat{\sigma}_{v_i}^2)$ corresponding to each $x_i$ , and thus the remeasurement of *X,*

$X(\lambda) = X + \sqrt{\lambda}\,U$ , is created where $U$ is a vector of $u_i (i = 1, \ldots, n)$ . The rest of the procedures stay the same as in the case of homogeneous measurement errors.

In figure 3.5, we simulate a data set with heteroscedastic measurement error in the predictor. The measurement error variance corresponding to each predictor value is equal to its absolute true value $(\sigma_{v_i}^2 = |x.true_i|)$ . The left graph in figure 3.5 plots $Y$ against the error-free predictor with quantile regression lines at 3 different quantile values, and the right graph plots $Y$ against the predictor containing heteroscedastic measurement error with quantile regression lines. The data points in the graph containing measurement error appear to be much more dispersed, especially at the lower and higher ends of $X$, and the quantile regression lines are all attenuated toward zero. Figure 3.6 shows the SIMEX correction for slope estimates using the same data as figure 3.5. The three plots correspond to quantiles $\tau = 0.25, 0.5, 0.75$ . In each plot, the straight line represents the linear extrapolation and the red curve represents the quadratic extrapolation. The "simex lin" dot and the "simex quad" dot mark the SIMEX estimates of the slope based on the respective extrapolation method, while the "naive est" dot marks the naïve estimate. The figure shows that both SIMEX estimates are closer to the true value than the naïve estimate, and the quadratic extrapolation leads to the least biased estimate.

The SIMEX method is also applicable when more than one regressors are error-prone. In this case, the computer-generated errors U would have multivariate normal distribution with zero mean and the same variance-covariance matrix as that of the measurement errors contained in the regressors. The number of remeasured datasets needs to be greatly increased to match the increase of dimensions of the simulation. The

extrapolation procedures would remain the same.

*3.3.6.3 Estimating the variance of the SIMEX estimators*

Asymptotic results are hard to derive for the SIMEX estimators. Simulation studies have been presented to show that they are nearly asymptotically unbiased and efficient for logistic regressions and different types of nonparametric regressions (Cook and Stefanski, 1994; Carroll et al., 1999).

Resampling techniques such as bootstrapping (i.e. drawing randomly with replacement from the current sample) or Jackknifing (i.e. using subsets of available data) can be implemented to estimate the standard errors of the SIMEX estimator at the price of some additional computational burden (Stefanski and Cook, 1995; Carroll et al., 2006, p. 110). For the study in this thesis, however, bootstrapping is impractical. In order to achieve sufficient accuracy in estimating student growth percentiles, we estimate the quantile regressions at 99 different quantiles. Due to the large sample sizes of our data and to the computationally intensive simulation stage, applying the SIMEX method to the QR models at each grade level takes a few weeks. Bootstrapping the SIMEX method would then entail months of computation time. As a result, we choose to rely on the theoretical results for estimating the variances of the SIMEX estimates.

Two quantities are important in the variance estimation. The first is $\hat{v}^2(\lambda)$, which denotes the mean of variances of $\hat{\theta}_{naive}(\lambda, b) \quad (b = 1, \ldots, B)$ averaged over a large $B$. The second quantity is $s_{\Delta}^2(\lambda)$, which denotes the sample variance of $\{\hat{\theta}_{naive}(\lambda, b)\}_{b=1}^{B}$. Stefanski and Cook (1995) show that, under a mild smoothness condition and when $B$ is very large, extrapolation of the difference, $\hat{v}^2(\lambda) - s_{\Delta}^2(\lambda)$ to $\lambda =$

-1, provide an unbiased estimator of the variance of the SIMEX estimate. While this variance estimation is developed in the context of homoscedastic measurement errors, Carroll et. al. (1996) argue that it can also be used as an approximation of the SIMEX estimator variance when measurement errors are heteroscedastic, and it is precisely what will be done in the next chapter.

*3.4 Random Effect*

As mentioned in section 2.5.3, residuals are assumed to be independently distributed in QR models although they are not required to be identically distributed. The independence assumption is almost certainly violated with our dataset due to its nested structure, a well-known characteristic of educational assessment data. Individual students are nested within classrooms and schools, and academic achievement and growth of students in the same classrooms and schools are likely to be correlated due to the effects of teachers, schools, neighborhoods, and peers. In order to take this clustering into account, a random effect of classrooms or schools can be added to model (3.2).

For student *i* at grade *t* and school *j,* the $100\tau$ th conditional quantile of her score given past score(s) is estimated by:

$$Q^{(\tau)}(Y_{i,t,j}|Y_{i,t-1}) = \alpha_t^{(\tau)} + \beta_t^{(\tau)} Y_{i,t-1} + u_{t,j} \quad , \quad i = 1,\ldots,n \qquad t = 8,9,10 \qquad j = 1,\ldots,m \tag{3.10}$$

where $u_{t,j}$ is the random effect for school *j*. The distribution of this random term is generally assumed to be $N(0,\sigma_{u,t}^2)$ . It can also be assumed to have other distributional forms if evidences are available. The reason why $u_{t,j}$ does not have a superscript τ is because it is considered to be a random draw from a distribution assumed to be known, which is not τ-specific. All other terms are defined the same as in

(3.2).

Koenker (2005) regards fixed and random effects QR model as the "twilight zone of quantile regression", not because the idea is new. What makes model (3.10) challenging is its estimation. The fundamental dilemma is that the estimation of regular QRM is not based on distributional assumptions but the estimation of random effects has to be achieved through such assumptions.

Progress in the likelihood approach to quantile regression was made in the field of Bayesian QR. Yu and Moyeed (2001) showed that, irrespective of the actual distribution of the data, the Bayesian estimation of QR parameters can proceed based on the likelihood function of the asymmetric Laplace distribution (ALD) (also see Koenker and Bassett, 1978). A random variable $U$ is said to follow the ALD if its probability density is given by

$$f_\tau(u)=\frac{\tau(1-\tau)}{\sigma}\exp(-\rho_\tau(\frac{u-\mu}{\sigma})) \qquad (3.11)$$

where $u$ is a specific value of $U$, $0<\tau<1$, $\mu$ is the location parameter of the ALD function, $\sigma$ is the scale parameter, and $\rho_\tau$ is defined in equation (2.8). It can be shown that the minimization of the loss function (2.8) with respect to $\xi$ is exactly equivalent to the maximization of (3.11) with respect to $\mu$. Therefore the estimation of the $\tau$ th quantile of the outcome in QR is equivalent to the estimation of the location parameter $\mu$ of a $\tau$-specific ALD density function.

Based on this conclusion, regardless of the error distribution of the QR model, the likelihood function of the data for a given $\tau$ can be written as

$$P_\tau(y_{i,t,j}|\alpha_t^{(\tau)}, \beta_t^{(\tau)}, \sigma_t, u_{t,j}, y_{i,t-1}) = \frac{\tau(1-\tau)}{\sigma_t}\exp(-\rho_\tau(\frac{y_{i,t,j}-\alpha_t^{(\tau)}-\beta_t^{(\tau)}y_{i,t-1}-u_{t,j}}{\sigma_t}))$$

where $\sigma_t$ is the standard deviation of the residual, and the other parameters are explained in (3.6). The posterior distribution of the model parameters given the data is proportional to the above likelihood function:

$$P_\tau(\alpha_t^{(\tau)}, \beta_t^{(\tau)}, \sigma_t, u_{t,j}|y_{i,t,j}, y_{i,t-1}) \propto \frac{\tau(1-\tau)}{\sigma_t}\exp(-\rho_\tau(\frac{y_{i,t,j}-\alpha_t^{(\tau)}-\beta_t^{(\tau)}y_{i,t-1}-u_{t,j}}{\sigma_t}))$$

(3.12)

A Metropolis-Hastings algorithm can be implemented to sample from this joint distribution and estimate the full posterior distributions for each of the parameters.

Geraci and Bottai (2007) also utilized the ALD and proposed a quasi-Bayesian Monte Carlo EM algorithm to estimate quantile regression with random effects. Assume that the outcome $Y_{ij}$ conditional on all the model parameters $\boldsymbol{\eta}$ plus the random effect $u_j$ has the distribution $f(Y_{ij}|\boldsymbol{\eta}, u_j)$ which is an ALD, and assume that $u_j \sim N(0, \sigma_u^2)$ . It can be shown that the posterior distribution of the random effect $u_j$ conditional on the model parameters and $Y_{ij}$ has the core:

$$f(u_j|Y_{ij}, \boldsymbol{\eta}) = f(Y_{ij}|\boldsymbol{\eta}, u_j) f(u_j|\sigma_u^2) \qquad (3.13)$$

Setting an initial values to η and $\sigma_u^2$ , a Markov Chain Monte Carlo sampling technique can be used to sample from the distribution of (3.13). The likelihood $f(Y_{ij}|\boldsymbol{\eta}, u_j)$ can then be updated substituting $u_j$ with the samples. After that the maximum likelihood estimates of η is obtained which maximilze the updated $f(Y_{ij}|\boldsymbol{\eta}, u_j)$ . The next iteration will start with sampling from $f(u_j|Y_{ij}, \boldsymbol{\eta})$ again with the updated η. The

iterations go on until the parameters reach convergence. The authors' simulation study shows that their estimates have good properties and have much lower mean squared error than the estimates from regular QR models without random or fixed effects when the data is nested.

In order to explore the estimation of the QR models with random effects, we adopt the full Bayesian approach[1]. Non-informative normal priors are chosen for the model intercept and slope, and non-informative uniform priors are chosen for the standard deviations of the model residual and the random effect. Starting values of these parameters are decided by a random draw from their prior distributions. A Metropolis-Hastings algorithm is implemented with 5,000 iterations and 2,500 burn-in. We conduct a simulation study to evaluate the performance of the Bayesian estimators. Data is simulated from the model $y_{ij} = 1 + 3 \cdot x_{ij} + u_j + \epsilon_{ij}$ . We fix the distribution of the random effect $u_j$ at *N(0,1),* and let the distribution of the model residual $\epsilon_{ij}$ vary between a standard normal distribution and a t-distribution with degrees of freedom of 3. The model is estimated at $\tau = 0.25, 0.5, 0.75$ . Table 3.2 presents the results from the simulation study. In table 3.2, $\beta_0$ and $\beta_1$ referring to the intercept and slope respectively, QR refers to the QR model which ignores the nested structure of the data and does not include a random effect, and QRRE refers to the QR model with random effect.

The simulation is conducted with 500 replications in each scenario. Bias is calculated with

1 Geraci and Bottai (2007) produced their results with a C program. A new program is written in R based on their algorithm for this thesis. Unfortunately, the program is extremely inefficient in computation. As a result, we choose the full Bayesian approach as the only practical alternative.

$$\hat{bias}(\hat{\beta}_k) = \frac{1}{500}\sum_{r=1}^{500}(\hat{\beta}_k^{(r)} - \beta_k)$$ for $k = 0,1,$ where $\beta_k$ is the true value[2] for the parameter. Mean squared error (MSE) is calculated with

$$\hat{mse}(\hat{\beta}_k) = S^2(\hat{\beta}_k) + [\hat{bias}(\hat{\beta}_k)]^2 = \frac{1}{500}\sum_{r=1}^{500}(\hat{\beta}_k^{(r)} - \bar{\beta}_k)^2 + [\hat{bias}(\hat{\beta}_k)]^2$$ where

$\bar{\beta}_k = \frac{1}{500}\sum_{r=1}^{500}\hat{\beta}_k^{(r)}$ and $S^2(\hat{\beta}_k)$ is the Monte Carlo variance of the estimates.

Results in table 3.2 show that the nested structure of the data does not seem to produce large amount of bias in the simple QR models in the first place. The bias of the intercept estimates are smaller than 2 percent of the true values, and the bias of the slope estimates are still smaller than that. Table 3.2 also shows that QRRE leads to slight improvement in the MSE of the slope estimates in comparison to QR, and, in a few cases, the QRRE slope estimate has smaller bias than the QR estimate as well. When estimating the intercepts, however, QRRE invariably leads to greater bias and higher MSE than QR. The exact reason for this result is not clear. We hypothesize that using more informative priors for the parameters and increasing the number of replications will lead to better results of the QRRE method. This hypothesis warrant future investigation. For the present analysis, since the simulation study does not show disturbing effect of data clustering on the estimation of QR models, and since the available QRRE method does not have a clear advantage over the simple QR models, we choose not to apply the QRRE method to the analysis of student academic growth.

2 The true value for the slope is always one. The true value for the intercept, however, changes at different quantiles. For example, when $\tau = 0.25$, the true value is equal to 1 plus the $0.25^{th}$ quantile of the total error, $u_j + \epsilon_{ij}$. Similar calculations are extended to other quantiles. This way of choosing the true values is confirmed through correspondence with Matteo Bottai.

*3.5 Summary*

In this chapter we have summarized the characteristics of the state testing program and the cohort of longitudinal test scores obtained from it. We introduced the QR model to estimate student growth percentiles, and also discussed model specifications including the choice of predictors and linearity. The impact of measurement error contained in the test scores on the estimation of the QR models is then analyzed in detail. Several methods of adjusting for measurement error-induced bias are briefly reviewed, and the SIMEX method is proposed to be used in combination with the QR model for data analysis in this thesis. Several graphs with simulated data are presented to explain and demonstrate the performance of the SIMEX estimators with QR models. Lastly, we discuss the nesting structure of the data, its possible consequences on the QR model estimation, and two options of dealing with this problem. After a small-scale simulation study with the Bayesian QR model with random effects, we decide not to apply this method to analyze student academic growth in the next chapter.

## Chapter 4 Results

In this chapter we present the results of the QR models combined with the SIMEX method and discuss their methodological significance and practical implications. In the first section, we present results from the basic QR models, interpret the results and

explore their implications for students' growth patterns. We also analyze the goodness-of-fit of the models and present results from two QR models with different specifications. The second section is devoted to the results of the SIMEX method. The section starts with a simulation study which demonstrates the performance of the SIMEX method applied to QR models relative to the naive QR models ignoring measurement error in the predictor(s). Next we present the results of the SIMEX method applied to the QR models with the longitudinal testing data, and compare the results before and after the SIMEX correction. We also explain the practical meaning of the SIMEX method results. In the third section, we present the conditional growth charts based on the QR and SIMEX models. The growth percentiles estimated for each student using these models are also analyzed. The final section of this chapter shows some examples of using the conditional growth chart method to diagnose student and school growth.

*4.1 The QR Models ignoring measurement error*

*4.1.1 The Lag-1 Models—Results and Interpretation*

Table 4.1 summarizes the results for the lag-1 QR models specified in equation 3.1. The table includes intercepts α, slope coefficients β, and goodness-of-fit statistics $R^1$ for seven different quantiles $(\tau = 0.03, 0.1, 0.25, 0.5, 0.75, 0.9, 0.97)$. Sample sizes in each of the three models and results from tests of equality of slopes are also presented.

In these models, the values of the independent variables are centered about their means for ease of model interpretation. Thus the τ-specific intercept α is the predicted scale score for students who got the average score in the previous year and grew at growth percentile τ in the current grade. The slope coefficient $\beta^{(\tau)}$ is interpreted as the

change in the τth quantile of the response variable corresponding to a unit change in the predictor. If the β's across all quantiles were equal to each other in a particular model, it would mean that the conditional distributions of the outcome change their locations but preserve their shapes as the predictor value changes, which is one of the central assumptions of OLS regression—homoscedasticity. We test the equality of slopes using the test proposed in Koenker and Bassett (1982). Results from the joint tests of equality of slopes, presented in Table 4.1, show that the seven slopes are not equal to each other in any of the three models. In fact, further tests show that each pair of slopes in a given model differ significantly from each other (the results are not presented). It means that changes in the predictor is associated with not only location shifts but also shape alterations in the response distributions.

Several patterns are notable in Table 4.1. First, all the coefficients are significantly above zero, as the coefficient estimates are orders of magnitude greater than their standard errors. It means that past scores are significant predictors of current scores at all seven conditional quantiles. Second, the intercepts are increasing functions of τ, since, among all students who obtained the average score in the previous year, those who grew at higher growth percentiles should naturally have higher fitted outcome values.

The third pattern is that the slopes decrease as τ increases in all three models. To understand the implications of this pattern, suppose the interquartile range of grade 8 scores (i.e. score at the 75$^{th}$ percentile minus score at the 25$^{th}$ percentile) is $r$ given that grade 7 score equals $x$. When the grade 7 score increases to $x + 1$, the estimated 75$^{th}$ percentile in grade 8 increases by 0.81 unit while the estimated 25$^{th}$ percentile increases

by 0.87 unit according to the slope estimates at τ equals 0.75 and 0.25 in the model of grade 8 scores regressed on grade 7 scores. Thus the interquartile range becomes (*r* + *0.81 – 0.87) = (r* – 0.06). This example shows that, when the slope at the .25$^{th}$ quantile is steeper than that at the .75$^{th}$ quantile, the interquartile range of the outcome reduces when predictor value increases. Since this statement holds true for other quantiles as well, and since slopes at lower quantiles are always steeper than those at higher quantiles, we can conclude that the variability of the conditional distributions of the outcome decreases as the predictor value increases. The data and quantile regression lines in Table 4.1 are plotted in Figure 4.1 (a), (b), and (c), and the shape change of the response distributions are apparent—the scatter plots become narrower at the higher ends of the predictors.

In order to analyze the location and shape shifts of the response distributions in more details, we examine QR coefficients of the lag-1 models for a more dense sequence of quantile values. Specifically, coefficients are estimated at 97 quantile values $(\tau = 0.02, 0.03, \ldots, 0.97, 0.98)$. Figure 4.2 provides a graphical view for the intercept and slope estimates as functions of τ. The 95% confidence envelopes around each curve of coefficients are also drawn, which are very thin relative to the ranges of the graphs. The horizontal line in each plot marks the coefficient of OLS regression for the particular model, also with 95% confidence intervals.

The plots in Figure 4.2 show that summarizing the data with OLS regressions alone would lead to the loss of much information regarding shape shifts. The plots repeat the patterns in Table 4.1 that the intercepts are monotonically increasing with τ while the slopes are monotonically decreasing with τ, with their confidence envelopes far above

zero. It confirms our conclusion that students with higher past scores are less varied in their current scores than those with lower past scores. Figure 4.1 shows that most of the outliers occur in the lower ends of the predictor values, which is perhaps one of the major factors that contribute to the difference of variability, but why do outliers mostly occur among low-achieving students? What are the reasons behind the difference of variability in both achievement and growth between low-scoring and high-scoring students?

We provide several possible explanations for the above-mentioned phenomenon. The first one is that the difference of variability is mostly an artifact of test design. As figure 3.2 suggests, measurement errors in the lower end of score distribution are much larger than those in the higher end. This may be the reason why the conditional growth distributions of the low-achieving students are more dispersed than those of the high-achieving students.

To evaluate the plausibility of this hypothesis, we choose a group of low-scoring students in grade 7 with a fixed variance and a group of high-scoring students in grade 7 with similar variance. We follow the variance change of these two groups through grade 10, and also observe the reported standard errors of measurement associated with their scores in each grade. Specifically, the groups chosen are students who scored between 709 and 719 in grade 7 (709 is the $90^{th}$ unconditional percentile in grade 7 in this cohort), and those who scored between 541 and 551 in grade 7 (551 is the $10^{th}$ unconditional percentile in grade 7 in this cohort). Table 4.2 (a) reports the sample sizes, sample standard deviations of scores, and range of standard errors of measurement (SEM) of the two groups in grades 7, 8, 9, and 10. Both groups have some attrition. The attrition rate of

the low-achieving group is much higher than that of the high-achieving group. In section 3.1.2, we noted the high attrition rate in the whole cohort—by the end of grade 10, the cohort loses almost 1/5 of its population size in grade 7, and only about 10 percent of this loss is due to retention. We also noted the demographic change of the cohort—the percentages of White and Asian students increase from grade 7 to grade 10, while the percentages of Black and Hispanic students decrease during the same period. Sample sizes reported in Table 4.2 (a) provide a piece of evidence that the student loss mostly occurred in the low-achieving end. While we do not have sufficient data to study the causes and consequences of the attrition, it remains a very important research topic for anyone interested in understanding the true educational achievement and growth in the state.

The two groups chosen in Table 4.2 (a) have similar sample variances in grade 7, since a range of 11 is imposed on the scores of both groups. After grade 7, the variance of the low-achieving group in any given grade is much greater than that of the high-achieving group, but the SEM ranges of the low-achieving group are also much larger. We check the distributions of SEM and find that in each grade, there is a small number of students in the low-achieving group with unusually large SEM. Many of these students have the lowest obtainable scale scores which, as discussed earlier, are especially error-prone. We exclude the students with large SEM from the low-achieving group so that the upper bound of SEM in this group is no larger than the upper bound of SEM in the high-achieving group. Table 4.2 (b) reports sample sizes, standard deviations, and SEM ranges for the same two groups after the exclusion. The distributions of SEM in the two groups

are now alike judging from histograms (which are not presented). Variances of scores in the low-achieving group excluding outliers are much smaller than those with outliers, and yet they are still consistently larger than variances of the high-achieving group over the years. Tables 4.2 (a) and (b) show that larger measurement errors in the lower end of score distributions explains part but not all of the variance difference between low and high achievers, and measurement error does not explain why most of the outliers occur among low-achievers, either.

A second candidate for the explanation of the variance difference is that it is a selection/accumulation effect. Since academic achievement is accumulative, and since becoming a high-achiever by grade 7 requires long-term consistent effort, it is likely that high-achievers have more stable and similar growth histories, and tend to continue growing stably as well. Low achievement, on the other hand, may be a result of unsteady levels of effort in many cases, and will predict heterogeneous growth paths in the future, too.

The best way to back up this argument would be to follow the same cohort from the 3rd grade to 10th grade. Third graders supposedly have not established their growth patterns, thus both high-achievers and low-achievers in the 3rd grade may follow quite diverse growth paths in the next year. After several years of selection process, high-achievers become a more and more uniform group and their courses of growth should also be less varied. In short, if the selection/accumulation effect exists, the variance difference in conditional growth distributions between high-achievers and low-achievers should become more pronounced along the years. Since we only have four years' worth

of data, this analysis cannot be done. In Figure 4.2, we do see that the downward curve in (f) is steeper than the curves in (b) and (d), which means that shape shift is more substantial in that year than those in the earlier years. Nevertheless, there is no noticeable pattern along the years, and the data we have do not provide strong support for the above-mentioned argument.

For a third possible explanation, we hypothesize that there are some social/behavioral factors that make high-achievers resemble one another and low achievers each grow in his own way. Suppose high academic achievement requires the assembly of adequate family socio-economic status, parental guidance, instruction, personal ability, personal motivation, and peer influence, and the breaking-down of any of these factors may lead the student astray, then naturally the high-achievers form a more homogeneous group than other students. Again, we do not have enough data to test this theory.

Besides the above-mentioned hypotheses, the ceiling effect among the high-achievers also appears to be a possible explanation of the shape shift in the conditional distributions. However, if ceiling effect exists among the high-achievers, then floor effect should also exist among the low-achievers, and it is not clear why the the former should be more influential to lead to smaller variability among the high-achievers. Since this topic is not the central focus of this thesis, we leave these questions for future investigation.

*4.1.2 Goodness-of-fit*

The goodness-of-fit statistics $R^1$ presented in Table 4.1 is defined in equation

2.11. It stands for the magnitude of residuals in an intercept-only model that is reduced by including the predictors, and ranges from 0 to 1 as the familiar R square in OLS regression. According to the results in Table 4.1, adding the previous year's scores as predictor reduce about half of the sum of weighted absolute residuals for each model and at each quantile. The .97th quantile appears to have the worst fit among all seven quantiles in the three models, yet its $R^1$ value is still quite high (around 0.4).

Buuren (2007) introduces a worm plot to diagnose fit in quantile regression. The worm plot is essentially a Q-Q plot. The rationale is that, if the model has good fit, the empirical cumulative probability of the estimated conditional percentiles must be close to the τ's. The difference between worm plot and $R^1$ in assessing goodness-of-fit is that the former assesses model fit across quantiles and at fixed predictor values, while the latter is across predictor values, but τ-specific.

To construct the worm plots we estimate the QR models at 99 quantile values $(\tau = 0.01, 0.02, \ldots, 0.98, 0.99)$. Let $x_i$ denote a specific predictor value, let $\hat{\beta}^{(\tau)}$ denote the estimated model coefficients at a particular τ, and let $\tilde{F}_y(x_i'\hat{\beta}^{(\tau)})$ denote the empirical quantile corresponding to the predicted value $x_i'\hat{\beta}^{(\tau)}$. It is straightforward to obtain the value of $\tilde{F}_y(x_i'\hat{\beta}^{(\tau)})$ —among all the students who have the same predictor value as $x_i$, calculate the proportion of them whose outcomes are below $x_i'\hat{\beta}^{(\tau)}$. The last step is to simply plot the difference between $\tilde{F}_y(x_i'\hat{\beta}^{(\tau)})$ and τ for selected predictor values. If the model has perfect fit, the worm plot should be a straight horizontal line at zero.

Figure 4.3 (a), (b), and (c) are worm plots for the lag-1 QR models. The x-coordinates of the points in the worm plots are the standard normal z-scores corresponding to the τ's. We convert the quantiles to z-scores simply to make the units on the axis larger and more understandable. Similarly, the y-coordinates are the z-scores corresponding to the difference between $\tilde{F}_y(x_i'\hat{\beta}^{(\tau)})$ and τ. Each subplot in each page contains goodness-of-fit results aggregated over a range of predictor values. The subplots are ordered from the lower-left panel to the upper-right panel corresponding to increasing predictor values. Below each 4 by 5 worm plot, there is a 4 by 5 table that presents the predictor range and sample size for each subplot. For example, in figure 4.3 (a), the cell in the 4th row and 1st column of the accompanying table is "320-519 [2422]". It means that the subplot in the 4th row and 1st column of the worm plot depicts model fit for the predictor values ranging from 320 to 519, which covers 2,422 students.

Although the predictors are centered in table 4.1, the predictor values in the tables of the worm plots are non-centered scale scores. Typically, standardized deviations between $\tilde{F}_y(x_i'\hat{\beta}^{(\tau)})$ and τ that are below 0.2 are considered small[3] following the rules of thumbs of regular effect sizes (e.g. Cohen's d). The plots show that, for all three models, except for the extreme predictor values and at the extreme quantiles, there are generally quite decent fit between the models and the data.

*4.1.3 Model results with other specifications*

Table 4.3 summarizes the results for the lag-2 models defined in equation (3.3). Table 4.4 summarizes the results for the "common condition" models defined in equation

3 According to my correspondence with Stef van Buuren

(3.4). The predictors in both models are centered about their means. The tables include estimates of intercepts, slope coefficients, and goodness-of-fit statistics $R^1$ for seven different quantiles.

Table 4.3 shows some similar patterns as those in table 4.1. First, the intercepts are increasing functions of $\tau$, which means that for students who got the mean scores in both grades ($t - 1$) and ($t - 2$), their predicted scores in grade $t$ increase as their growth percentiles increase. Second, $\beta_1$ is a decreasing function of $\tau$ in each model, which means that, when holding scores from grade ($t - 2$) constant, as scores in grade ($t - 1$) increase, the variability of the conditional distributions of grade $t$ scores decreases. The same pattern does not repeat for $\beta_2$, i.e. when holding scores from grade ($t - 1$) constant, scores in grade ($t - 2$) do not have a linear relationship with the variability of the conditional distributions of grade $t$ scores. Moreover, $\beta_2$ is always smaller than $\beta_1$ for any given grade or any given $\tau$ value, which means that grade ($t - 2$) scores are not as closely related to grade $t$ scores as grade ($t - 1$) scores do. Table 4.4 demonstrates all the patterns in table 4.1, and share similar interpretations. Results from the models in tables 4.3 and 4.4 are used in the analysis of student and school growth in later sections.

In terms of model fit, the $R^1$ statistics show that the common condition models in table 4.4 have worse fit than their corresponding models[4] in table 4.1, due to the longer distance in time between the outcomes and the predictors in the former models. The lag-2 models have better fit than both the lag-1 models and the common condition models. The

4 By "corresponding models" we mean models that have the same outcome, e.g. the model of grade 9 scores regressed on grade 7 scores in table 4.4 corresponds to the model of grade 9 scores regressed on grade 8 scores in table 4.1.

differences in $R^1$ between the lag-2 and the lag-1 models are small (around 0.03) at lower quantiles and become greater (around 0.05-0.07) at higher quantiles (i.e. for $\tau >$ 0.75). All the differences, including the smallest ones, are significant according to the F test results in analysis of deviance (which are not presented in the table)[5]. The test is not done for any other pair of models because they are not nested.

The worm plots show that the lag-1 linear models have lack-of-fit in the tails of the predictors. Figure 4.1 also suggests that there may be non-linearity, especially in the lower tails of the predictors. We choose to focus on linear lag-1 models in this thesis, again, because they are easier to interpret and provide a starting point for understanding the growth chart method.

*4.2 Applying the SIMEX Method*

*4.2.1 Simulation Study*

We conducted two simulation studies to evaluate the performance of the SIMEX method with respect to models with one and two error-prone predictors. In each study, we used sample sizes of 400 at three different quantiles, 0.25, 0.5, and 0.75. To generate the data for the first study, we used the model

$$Y = 1 + X.true + \epsilon \quad \text{and} \quad X = X.true + v \qquad (4.2.1)$$

where $X.true$ is the true value of $X$, the observed predictor, $v$ is measurement error and is drawn from a standard normal distribution, and $\epsilon$ is the model residual. We consider two different probability distributions for $\epsilon$ : the standard normal, $N(0,1)$, and the chi-square with 2 degrees of freedom, $\chi^2_2$ , which is highly skewed. With each

5 The test is done between the lag-1 model of grade 9 scores regressed on grade 8 scores and the lag-2 model of grade 9 scores regressed on grade 8 and 7 scores across all $\tau$ values. Similarly, the test is also done between the lag-1 and lag-2 models with grade 10 scores as outcomes.

distribution of $\epsilon$ we also let the variance of *X.true* change from 4 to 9[6]. When *X.true* has a variance of 4, measurement error accounts for 20 percent (1/5) of *X's* total variance. When *X.true* has a variance of 9, 10 percent of the predictor's total variance is due to measurement error.

We simulated 2000 replications from each combination of the model residual and *X.true* distributions. In each replication, a quantile regression of *Y* on *X* is estimated ignoring the measurement error, and SIMEX estimates of intercepts and slopes are also obtained based on linear and quadratic extrapolants respectively. Table 4.5 presents the estimated bias averaged over the simulations for each $\hat{\beta}_k$ , $k=0,1$ ,

$$\hat{bias}(\hat{\beta}_k)=\frac{1}{2000}\sum_{r=1}^{2000}(\hat{\beta}_k^{(r)}-\beta_k) \qquad (4.2.2)$$

where $\beta_k$ is the true value for the parameter. According to equation (4.2.1), the true value for the slope is always one. The true value for the intercept, however, changes at different quantiles. It is equal to 1 plus the 25th, 50th, or 75th percentile of the model residual $\epsilon$ respectively.

We also reported the estimated mean squared error (MSE):

$$\hat{mse}(\hat{\beta}_k)=S^2(\hat{\beta}_k)+[\hat{bias}(\hat{\beta}_k)]^2=\frac{1}{2000}\sum_{r=1}^{2000}(\hat{\beta}_k^{(r)}-\bar{\beta}_k)^2+[\hat{bias}(\hat{\beta}_k)]^2 \qquad (4.2.3)$$

where $\bar{\beta}_k=\frac{1}{2000}\sum_{r=1}^{2000}\hat{\beta}_k^{(r)}$ . With equation (4.2.3) we are only calculating the approximation of MSE, since we use the Monte Carlo variance of the estimates. Still, it is

6 The 400 values of *X.true* are generated from a normal distribution with a fixed mean and a variance of 4 or 9. In this simulation study we let the variance of *X.true* change, because we believe that the extent of bias in the naïve estimation and the performance of the SIMEX estimators depend ultimately on the ratio of variances between the predictor and its measurement error, not the variance of the measurement error alone.

meaningful to compare this MSE for different estimators to judge their relative performance in efficiency and bias. For a more intuitive understanding of the results, Figure 4.4 plots the density of the 2000 naïve and SIMEX estimates for each scenario with normally distributed model residuals. The scenarios for model residuals with chi-square distributions look quite similar to Figure 4.4, and are not presented.

To generate the data for the second study, we used the model $Y = 1 + X_{1.\text{true}} + X_{2.\text{true}} + \epsilon$ , $X_1 = X_{1.\text{true}} + v_1$ , and $X_2 = X_{2.\text{true}} + v_2$ , where $X_{1.\text{true}} \sim N(1,9)$ , and $v_1 \sim N(0,1)$ . We let the distributions of $X_{2.\text{true}}$ vary from *N(1,4)* to *N(1,16),* corresponding to $v_2 \sim N(0,1)$ and $v_2 \sim N(0,2)$ respectively. The model residual $\epsilon$ is drawn from the standard normal distribution. Like in the first study, we simulated 2000 replications from each scenario, and in each replication, a quantile regression of *Y* on $X_1$ and $X_2$ is estimated ignoring the measurement error, and SIMEX estimates of intercepts and slopes are also obtained based on linear and quadratic extrapolants. Table 4.6 presents the estimated average bias and MSE calculated with equations (4.2.2) and (4.2.3).

Table 4.5 shows that, first of all, regressing the dependent variable on error-contaminated predictors does cause considerable amount of bias. When measurement error accounts for 20 percent of the observed predictor's total variance, average bias in the slope is about 20 percent of the slope's true value at any given quantile. When measurement error accounts for 10 percent of the observed predictor's total variance, average slope biases are about 10 percent of their true values across quantiles. The slope biases are all negative, meaning that the slopes are biased toward zero. This is the

attenuation effect discussed in the previous chapter. The intercept biases are negative at .25$^{th}$ quantile in all cases, but monotonically increase at higher quantiles and become relatively large positive numbers at .75$^{th}$ quantile. The difference between intercept at .75$^{th}$ quantile and .25$^{th}$ quantile is the interquartile range of the conditional distribution of the outcome when the predictor is equal to zero. Thus we see that the interquartile range is overestimated when measurement error is present in the predictor and ignored in the model estimation. Also since the slopes at different quantiles are almost parallel to each other, this overestimation of interquartile ranges (and possibly variances) is extended to other predictor values as well. These patterns of biases in the intercepts and slopes do not seem to vary significantly when the model residuals change from a normal distribution to a highly skewed chi-square distribution.

The results in Table 4.5 also show that the SIMEX estimators perform much better than the naïve estimators in terms of bias and MSE across the simulated scenarios except in the intercepts at the 0.25$^{th}$ quantile. The SIMEX estimates of slopes are still attenuated, and the intercept estimates show that the interquartile ranges (and possibly the variances) of the outcome conditional distributions are still overestimated, but the amount of these biases are greatly reduced. For example, in the case where measurement error accounts for 20 percent of the observed predictor variance, the quadratic SIMEX estimator reduce the slope biases from 20 percent to about 3 percent of their true values across quantiles.

We see in Figure 4.4 that the SIMEX estimates have larger variability than the naïve estimates, but they still reduce MSE in most cases due to their effective reduction

of the biases. More specifically, Figure 4.4 shows that the SIMEX estimates based on quadratic extrapolations (SIM.q) have larger variability than the SIMEX estimates based on linear extrapolations (SIM.lin), and the naïve estimates have smaller variability than the SIMEX ones. In terms of bias, both Table 4.5 and Figure 4.4 show that the SIM.q estimators are generally more effective in reducing bias than the SIM.lin estimators, but the former does not necessarily have smaller MSE than the latter. When the variance of measurement error is 10 percent that of the observed predictor, the SIM.lin estimators generally have similar MSE as the SIM.q estimators, and the former perform better than the latter in a few cases. When the measurement error variance is 20 percent that of the observed predictor, however, the MSE of the SIM.lin estimators are usually larger than those of the SIM.q estimators.

Table 4.6 shows similar results. First, the naïve estimates of slopes are attenuated for both error-prone predictors, and the attenuation effect is smaller for the predictor which contains a smaller proportion of measurement error. Biases in the intercept estimates show that the interquartile range of the outcome conditional distribution is overestimated, just like in the single-predictor models. Second, the SIMEX estimators perform better than the naïve estimators in bias and MSE except for the intercept at $.25^{th}$ quantile in the first scenario. The SIM.q estimator is more effective than the SIM.lin estimator in reducing biases in all cases, but neither estimator show a clear advantage over the other in terms of MSE.

### *4.2.2 Applying the SIMEX Method to the State Assessment Data*

We now apply the SIMEX method to the lag-1 QR models with the state assessment

data. We adopt methods discussed in section 3.3.6 to accommodate heteroscedastic measurement error and to compute the standard errors of the SIMEX estimates. The latter task requires that a large number of remeasured data sets are simulated at each value of $\lambda$ (i.e. $B$ is large). In this application, we let $B = 500$, which leads to about 400 hours computing time on a 2.10 GHz processor for each of the 3 lag-1 models at $\tau = 0.01, 0.02, \ldots, 0.99$. The SIMEX estimates of the intercepts and slopes of the models are obtained, again, at $\tau = 0.01, 0.02, \ldots, 0.99$, but their standard errors are only estimated at $\tau = 0.04, 0.05, \ldots, 0.96$, due to the computational instability of variance estimation at extreme quantile values. Quadratic extrapolations are used to obtain both the SIMEX estimates and their standard errors.

Figure 4.5 presents a comparison of the SIMEX estimates to the naïve estimates of QR lines. The seven blue lines represent the SIMEX estimates, and the red lines are the naïve QR estimates plotted in Figure 4.1. Each set of lines corresponds to seven different quantiles from low to high $(\tau = 0.03, 0.1, 0.25, 0.5, 0.75, 0.9, 0.97)$. The vertical and horizontal green lines in each plot mark the cut scores that differentiate "non-proficient" from "partially proficient" in the corresponding grades.

The plots in Figure 4.5 show that the blue lines preserve the pattern of the red lines, i.e. there appears to be smaller variability of the conditional outcome distribution at the higher end of the predictor than at the lower end of the predictor in each model, which is the phenomenon discussed in section 4.1.1, even though the blue lines appear to be less spread out than the red ones across the predictor values.

It is also clear in the plots that all the SIMEX regression lines have steeper slopes

than the lines without correction, which indicates that the SIMEX method has probably corrected for attenuation to a certain extent. In short, the comparison of SIMEX estimates and naïve estimates with the reading assessment data generally resembles the comparison with simulated data—the SIMEX correction leads to lower estimates of intercepts, higher estimates of slopes, and smaller variability of the conditional outcome distribution at any given value of the predictor.

The practical consequence of the SIMEX corrections is quite complex. We now illustrate an aspect of this change by the dashed and dotted lines in Figure 4.5. In the plot of grade 8 scores against grade 7 scores, there are two vertical dashed lines and two vertical dotted lines. The dashed line in the left marks the crossing point of the .97$^{th}$ red QR line and the horizontal green line. Any student who scored below this dashed line in grade 7 needed to grow at a rate higher than the 97$^{th}$ conditional percentile to reach partial proficiency in grade 8. Thus the left dashed line define the group at high risk based on the simple QR model—most of this group stay non-proficient for the two years. The dashed line in the right marks the crossing point of the .03$^{th}$ red QR line and the horizontal green line. Any student who scored above this dashed line in grade 7 needed to grow at a rate lower than the 3$^{rd}$ conditional percentile to drop to non-proficiency in grade 8. Thus the right dashed line define the "secure" group based on the simple QR model—most of this group stay at least partially proficient for the two years. The left dotted line marks the crossing point of the .97$^{th}$ blue QR line and the horizontal green line and define the high-risk group based on the QR model with SIMEX correction. The right dotted line marks the crossing point of the .03$^{th}$ blue QR line and the horizontal green line, and define the

"secure" group based on the QR model with SIMEX correction.

The SIMEX estimates define a larger high-risk group and a larger not-at-risk group compared to those defined by the simple QR model (i.e. the low-achieving are more likely to stay low-achieving, and the high-achieving are more likely to stay high-achievers). The group in between, i.e. the group that had a moderate chance of changing their status through growth (whether it is to grow from non-proficient to partially proficient or the other way around), has a smaller size based on the SIMEX estimates compared with that based on naïve estimates. In short, students' trajectories seem to be more strongly determined by their achievement histories and have relatively less chance of changing classifications according to the SIMEX estimates in comparison to the conclusions based on the naïve estimates.

The SIMEX method leads to slightly different estimates of individual student's growth percentiles, but it does not change students' *relative* positions in the conditional distributions. In other words, a student who appears to grow faster than another student will remain the faster grower after we adjust for measurement errors. What the SIMEX method changes is the difference between the two students' growth rate (i.e. how much faster one is than the other). For example, student A scored 585 in grade 9 (just above the partially proficiency cut score in grade 9) and 607 in grade 10 (just above the partially proficiency cut score in grade 10). Student B scored 585 in grade 9 and 673 (above the proficiency cut score in grade 10). Based on the naïve QR estimates we conclude that student A grew at the $45^{th}$ conditional percentile, and student B grew at the $97^{th}$ conditional percentile in grade 10. However, after applying the SIMEX method we

estimate that student A grew at the 54th conditional percentile, and student B grew at the 98th conditional percentile. After the SIMEX correction, reaching partial proficiency in grade 10 is estimated to be a more difficult task for those starting at 585 in grade 9 than before. The difference between the two students' growth is still big, but not as big as that estimated previously. In general, the SIMEX correction may play a role in how individual students are diagnosed and how challenging and attainable objectives are set. We will explore this further later in the chapter.

### *4.3 Analysis of Growth*

In the previous sections, we have examined the simple QR models and QR models combined with the SIMEX method. In this section, we will study the results of these models as measures of student academic growth. In the first part of this section, a conditional growth chart is presented and explained. In the second part, we focus on the estimated growth percentiles of the students and analyze their patterns.

#### *4.3.1 The Conditional Growth Chart*

Figure 4.6 presents conditional growth charts based on fitted values of the lag-1 models. The left subplot is drawn based on the simple QR results, while the right subplot is based on the SIMEX results. The starting points in both subplots are chosen to be the scale score of 597, which is the 25th unconditional percentile of grade 7 reading scores in 2003. Three different growth paths at the 25th, 50th, and 75th conditional percentiles are drawn from the starting point in each subplot. The blue lines mark the proficiency cut scores in grades 7-10, while the red lines mark the partial proficiency cut scores in these grades.

The gray areas around the growth paths represent confidence bands. For example, in each subplot, the gray band around the lowest growth path at year 2004, or grade 8, is the 95% confidence interval of the projected value[7] for a student who started from the score of 597 in grade 7 and grew at the 25th conditional percentile. The same band at year 2005 has a slightly different interpretation. It marks the range of the 95% confidence intervals of the fitted values for students who scored *within the band* in 2004 and grew at the 25th conditional percentile. The band at year 2006, similarly, represent the range of 95% confidence intervals of the projected values for those who scored within the band in 2005 and grew at the 25th conditional percentile. For a student who started in grade 7 at the score of 597 and followed the 25th conditional growth path every year, the probability of falling within the lowest gray band in grade 10 is obviously smaller than 0.95. The exact probability is hard to determine, but $(0.95)^3$ serves as a very conservative lower bound. In other words, for a student who started in grade 7 at the score of 597 and grew consistently at the 25th conditional percentile for three years, the probability of falling within the lowest gray band in grade 10 is higher than 0.86. The other confidence bands

7 To estimate the variance of the fitted/projected values in QR, we simply bootstrap the lag-1 QR models with 2000 bootstrap samples. To calculate the fitted value variance based on the SIMEX results, the CPU time required for bootstrapping is beyond the limit of our resource, and an approximation based on theoretical results is adopted. Let $\hat{Y}^{(\tau)}$ denote the fitted values at the τth conditional quantile, and let $X$ denote the predictor: since $Var(\hat{Y}^{(\tau)}|X=x_i)=Var(\hat{\beta}_0^{(\tau)}+\hat{\beta}_1^{(\tau)}x_i|X=x_i)=Var(\hat{\beta}_0^{(\tau)}|X=x_i)+x_i^2Var(\hat{\beta}_1^{(\tau)}|X=x_i)+2x_iCov(\hat{\beta}_0^{(\tau)},\hat{\beta}_1^{(\tau)}|X=x_i)$ and since $Var(\hat{\beta}_0^{(\tau)}|X=x_i)$ and $Var(\hat{\beta}_1^{(\tau)}|X=x_i)$ are estimated during the SIMEX computation as explained in section 3.3.6, we only need to calculate $Cov(\hat{\beta}_0^{(\tau)},\hat{\beta}_1^{(\tau)}|X=x_i)$ to get an estimation of the fitted value variance. Stefanski and Cook (1995) point out that the variance-covariance matrix of all parameters can be estimated using the simulation-extrapolation method. In the process of our SIMEX computation, however, the extrapolation of the remeasurements of the covariance between the parameters seem to be quite problematic. We therefore use the $Cov(\hat{\beta}_0^{(\tau)},\hat{\beta}_1^{(\tau)}|X=x_i)$ estimated in the naïve lag-1 QR models as an approximation. Since the covariance is always negative, and since its absolute value in the naïve models is usually slightly smaller than that for the SIMEX results (as shown in our simulation study, but not presented), this approximation probably leads to an overestimation of the variance and the confidence intervals. Also because the covariance is very small in all cases (<0.01), the extent of the overestimation is likely to be small.

around the 50th and the 75th conditional growth paths follow the same interpretations.

The most noticeable difference between the two subplots lies in the width of the confidence bands. It is reasonable that the width of the bands should increase along the years since it accounts for the accumulation of projection standard errors in the series of lag-1 models. This trend is much more obvious in the SIMEX plot than in the naïve QR plot, because all the parameters in the former model have larger standard errors than those in the latter. In a sense, the SIMEX method acknowledges the uncertainties represented by the measurement errors and incorporate them into the uncertainties of its projections, whereas the naïve QR model does not take these types of uncertainties into account. The point projections of the two methods, however, hardly differ from each other in figure 4.6, although it appears that students who follow the 50th conditional percentile growth paths consistently are projected to be barely above proficiency in the left plot and slightly below proficiency in the right plot.

Both subplots show the benefit of constant effort. For a student who started from well below proficiency in grade 7, three years of consistent growth at the 50th or the 75th conditional percentiles carries her very close to or well above proficiency in grade 10. On the other hand, the plots also demonstrate the attainability of the standards. As mentioned in chapter 3, the state in question uses the cut score of partial proficiency as the sole standard in its definition of AYP. Considering that, for a student who started from the 25th unconditional percentile in the state, even three years of consistent low growth at the 25th conditional percentiles ensures her to be comfortably above partial proficiency by grade 10, this standard is not unreasonably challenging for most students in the state.

Of course, numerous different conditional growth charts can be drawn with different starting points and different growth percentiles. Figure 4.6 is just one of them. These charts can be used to evaluate the difficulty of reaching certain objectives for people starting at certain levels. But for diagnosis of individual students' growth, it is better to estimate growth percentiles for each student.

*4.3.2 Student's Growth Percentiles*

Growth percentiles are estimated for every student based on the lag-1 QR models with and without the SIMEX correction. The specific estimation algorithm is explained in chapter 3. With the estimated growth percentiles, we first seek to answer the following question—how typical are the growth paths depicted in Figure 4.6?

Let $\hat{P}_{g8|g7}$ , $\hat{P}_{g9|g8}$ , and $\hat{P}_{g10|g9}$ denote the estimated growth percentiles of grades 8, 9, and 10 conditioning on scores of grade 7, 8, and 9 respectively. If the growth percentiles from different years are independent of each other, it is easy to calculate the joint probability of several years' growth paths. For example, the probability of growing at or below the 75th growth percentile for three years is $0.75^3$ , and the probability of growing at or above the 75th percentile for three years is $(1-0.75)^3$ . Table 4.7 presents the probabilities of growing at or above the 25th, 50th, and 75th conditional percentiles for one year, two years, and three years. The probabilities are calculated theoretically assuming independence, empirically based on naïve QR results, and empirically based on SIMEX results. The empirical results are consistently lower than the theoretical ones when it comes to two or three years of joint probabilities, indicating dependence between growth percentiles from different years. We find from this table that at least two growth

paths depicted in Figure 4.6 are quite atypical. According to the SIMEX results, slightly over a quarter of students grew at or above the 25th conditional percentile for three years continuously; only about 6 percent of students grew at or above the median rate for three years continuously; and growing at or above the 75th conditional percentile consistently for three years indicate some truly extraordinary effort—only one student out of every 200 could do it.

In fact, the growth percentiles from two consecutive years are negatively correlated, which explains why consistent growth is so unusual. The correlation between $\hat{P}_{g8|g7}$ and $\hat{P}_{g9|g8}$ is -0.324 based on naïve QR results, and -0.382 based on SIMEX results. The correlation between $\hat{P}_{g9|g8}$ and $\hat{P}_{g10|g9}$ is -0.267 based on naïve QR results, and -0.318 based on SIMEX results. The correlation between $\hat{P}_{g8|g7}$ and $\hat{P}_{g10|g9}$ is positive but quite small, below 0.06 for both methods.

In order to visualize the relationships between growth percentiles in different years, we plot the conditional density of growth percentiles in Figure 4.7. Specifically, students are divided into four groups based on their growth percentiles in grade 8 ( $\hat{P}_{g8|g7}$ ), below 25, between 25 and 50, between 50 and 75, and at or above 75. Density of growth percentiles in grade 9 ( $\hat{P}_{g9|g8}$ ) is then plotted for each of these groups, and these density plots make up the first row of plots in figure 4.7. Similarly, the second row of plots depict the density of grade 10 growth percentiles ( $\hat{P}_{g10|g9}$ ) for four groups of students whose grade 9 growth percentiles ( $\hat{P}_{g9|g8}$ ) range from [1, 25), [25, 50), [50, 75), and [75, 99] respectively. The third row consists of density plots of $\hat{P}_{g10|g9}$

similarly conditioning on $\hat{P}_{g8|g7}$ .

Figure 4.7 shows that students who grow at below the 25th conditional percentile one year tend to grow at very high conditional percentiles (the mode is above 90) the next year, whereas those who grow at above the 75th conditional percentile one year tend to grow at very low conditional percentiles (the mode is close to 10) the next year. Among students who have moderate growth percentiles (between 25 and 75), the distributions of next year's growth percentiles are almost uniform. The distributions of $\hat{P}_{g10|g9}$ conditioning on $\hat{P}_{g8|g7}$ also look uniform, since growth percentiles from non-consecutive years have very small correlation. In other words, figure 4.7 shows that there is a group of noticeable size who tends to alternate between very low and very high growth percentiles from one year to the next. For convenience's sake, we call this group of students the "radical growers". Their growth pattern appears to be a regression to the mean effect.

Several questions need to be answered to understand the alternation between radical growths. First, how many percent of students belong to the group of "radical growers"? Second, does the phenomenon of fluctuating growth persist if growth percentiles are estimated based on lag-2 models or the "common base" models instead of the lag-1 models? Lastly, do fluctuating growth lead to completely different results compared with stable moderate growth?

We answer the first question with Table 4.8, which presents the percentages of students whose growth percentiles from consecutive years differ by less than 25, between 25 and 50, between 50 and 75, and above 75. The left part of Table 4.8 is about the

difference between growth percentiles in grade 8 and grade 9, and the right part is about the difference between growth percentiles in grade 9 and grade 10. According to the results from the simple QR models, about 1/3 of the students who have complete data from grades 7-9 have highly fluctuating growth percentiles in grades 8 and 9, i.e. the difference between $\hat{P}_{g9|g8}$ and $\hat{P}_{g8|g7}$ is at or above 50. Based on the SIMEX results, the percentage of this group is even higher, above 40 percent. Similar conclusions can be drawn for growth percentiles in grades 9 and 10.

By the second question we try to probe the true story behind the swinging growth percentiles. We again observe highly fluctuating growth percentiles when the growth percentiles are estimated based on lag-2 models instead of lag-1 models, i.e. the difference between $\hat{P}_{g10|g8,g9}$ and $\hat{P}_{g9|g7,g8}$ is still large for a sizable group. However, fluctuations of growth largely disappeared after growth percentiles are estimated based on the "common condition" models. Recall that in this type of models, scores in grades 8-10 are all regressed on grade 7 scores, and the growth percentiles produced from these models are therefore all conditioning on the grade 7 scores $(\hat{P}_{g8|g7}, \hat{P}_{g9|g7}, \hat{P}_{g10|g7})$.

Figure 4.8 plots the conditional density of the growth percentiles estimated from the "common condition" models. Plots in the first row present the density of $\hat{P}_{g9|g7}$ for four groups of students—those whose $\hat{P}_{g8|g7}$ is below 25, between 25 and 50, between 50 and 75, and at or above 75. The second row of plots depict the density of $\hat{P}_{g10|g7}$ conditioning on $\hat{P}_{g9|g7}$ for the same four groups, and the third row consists of density plots of $\hat{P}_{g10|g7}$ conditioning on $\hat{P}_{g8|g7}$. These plots form sharp contrasts to those in

figure 4.7. They show that, among students who scored the same in grade 7, those who grew at lower percentiles in grade 8 tend to stay in lower percentiles in grade 9 and 10. The same pattern is found in students who had higher growth percentiles.

Figure 4.8 is an evidence that the fluctuations of lag-1 growth percentiles are at least partly due to the change of conditioning variables. In Figure 4.9, we specifically choose students whose lag-1 growth percentiles are highly fluctuating. The two groups of students presented in the two plots in Figure 4.9 fulfill the standards $\left|\hat{P}_{g9|g8}-\hat{P}_{g8|g7}\right|\geq 50$ and $\left|\hat{P}_{g10|g9}-\hat{P}_{g9|g8}\right|\geq 50$, respectively. We draw the distribution of difference between $\hat{P}_{g9|g7}$ and $\hat{P}_{g8|g7}$ for the first group in the left plot, and the distribution of difference between $\hat{P}_{g10|g7}$ and $\hat{P}_{g9|g7}$ for the second group in the right plot. If change of conditioning variables is the only major reason for the lag-1 growth percentile fluctuation, both the distributions in Figure 4.9 should have their modes around zero. Instead, figure 4.9 shows that both distributions are bimodal with one mode between 0 and 50 and the other model between 0 and -50. It means that growth fluctuations still subsist, only on a smaller scale, after common conditioning variables are used.

A second conditional growth chart is plotted in Figure 4.10 to help answer the third question—whether the "radical growers" and the stable moderate growers end up with similar achievements. Three different growth paths starting from the same score in grade 7 are depicted in the chart. The middle line represents consistent growth at the 50$^{th}$ conditional percentile in grades 8, 9, and 10. The lower line represents a growth path that alternate between the 25$^{th}$, 75$^{th}$, and 25$^{th}$ conditional percentiles, and the higher line

alternates between the 75th, 25th, and 75th conditional percentiles in grades 8, 9, and 10 respectively. The confidence bands around the growth lines are drawn in the same way as those in figure 4.6. The plot shows that swinging growth does not necessarily lead to different results than stable growth, and that, ultimately, a student's academic achievement after several years seems to depend on the sum of growth percentiles across the years. For example, in the year of 2005, the three hypothetical students following different growth paths had the same sum of growth percentiles for the past two years and therefore obtained very similar scores. In the year of 2006, however, their achievement diverged. The one who had the highest sum of growth percentiles (i.e. the one who follows the 75th-25th-75th growth percentiles) obtained the highest score, and the one who had the lowest sum of growth percentiles got the lowest score.

4.3.3 Summary

In this section, we plot the conditional growth chart and analyzed student growth percentiles. We find that a considerable proportion of students (at least 1/3) grow at drastically different conditional percentiles (i.e. conditional percentiles that differ by more than 50) from one year to the next. This fluctuation of growth demonstrates an regression-to-the-mean effect which is not completely gone when the growth percentiles are estimated with models of different specifications (such as the lag-2 model and the "common condition" model). Moreover, the fluctuating growth could lead to similar outcomes as the consistent moderate growth, provided that the ups and the downs are balanced. We have gone through analyses of growth percentiles in order to develop a diagnosis rule for student academic growth. As we see, low growth percentile in any year

should raise alarm but does not necessarily mean that the student is slipping away from her original level—she may belong to the "radical growers" that tends to leap forward with pauses in between. Two years of continuous low growth percentiles, however, is a sign of real risk. Similarly, one year of high growth percentiles should not be the basis of reward for individual students or schools, but two years of consistent high growth rates indicate truly outstanding merit.

### *4.4 Using the Conditional Growth Chart Method to Diagnose Student and School Growth*

We have discussed the construction and interpretation of the conditional growth charts. The patterns of longitudinal growth percentiles have also been examined. In this section, we use some real examples of students and schools to demonstrate the diagnosis of student and school growth.

#### *4.4.1 Student Examples*

Student 251337 remained below partial proficiency from grade 7 to 10, but made tremendous progress during those years. Figure 4.11 depicts an unconditional and a conditional growth chart to screen this student's growth. The left plot is an unconditional growth chart with 7 different quantile curves at the $0.03^{th}$, $0.1^{th}$, o.$25^{th}$, $0.5^{th}$, $0.75^{th}$, $0.9^{th}$, and $0.97^{th}$ unconditional quantiles for this cohort. The proficiency and partial proficiency cut scores for each year are also marked by two curves. The four black dots in the plot correspond to student 251337's scale scores in the four years. We see in this plot that the student made his way from below the $0.03^{th}$ quantile to somewhere below the $0.1^{th}$ quantile, which does not seem very impressive. And if we evaluate him by his

achievement level alone, he has made no progress during the years since he never made partial proficiency. With the conditional growth chart on the right, however, we are able to reach a completely different conclusion. We plot three different growth paths from where the student started in 2003. The three black lines follow steady growths at the $0.25^{th}$, $0.5^{th}$, and $0.75^{th}$ conditional percentiles each year. The student's growth is above the confidence band of the highest growth curve. Recall that only 0.5 percent of students are able to keep their growth percentiles above 75 for three years continuously, as calculated in Table 4.7. This particular student, who is estimated to have grown at the $84^{th}$, $88^{th}$, and $91^{th}$ conditional percentiles in grades 8, 9, and 10, has made an extraordinary amount of progress, and should be recognized even though he remained non-proficient by the end of grade 10.

Figure 4.12 depicts the growth story for student 561315. He remain proficient from grades 7 to 9, and dropped slightly below proficiency in grade 10. Again, we present the unconditional growth chart on the left and the conditional growth chart on the right. The student got almost identical scores for the four years, therefore his growth path measured by the scale scores look undramatic. Even though he dropped from above the $50^{th}$ unconditional percentile in grade 7 to the $25^{th}$ percentile in grade 10, his parents and teachers may not be able to see, from his scale scores or the unconditional chart alone, how extremely inadequate his progress has been. In the conditional growth chart, we see that the student grew steadily at the $25^{th}$ conditional percentile each year. There is, in fact, about the same proportion of people (0.5 percent) who grew continuously at or below the $25^{th}$ conditional percentile as those who grew steadily at or above the $75^{th}$ conditional

percentile. In short, the student's growth path is very rare and indicates serious problems in his academic development.

*4.4.2 School Comparisons*

A school's academic growth can be evaluated using a suitable summary of students' growth percentiles, such as the median, the mean, or the whole distribution. Figure 4.13 presents the boxplots of students' growth percentiles for several schools. We include 10 high schools in Figure 4.13 (a), with their school numbers at the horizontal axis. The boxplots represent the distributions of the growth percentiles in grade 9 conditioning on grade 8 scores. The short horizontal bar in the middle of each box designates the median growth percentile[8] in the school, and the upper and lower limit of the box are the 75$^{th}$ and 25$^{th}$ percentiles of the growth percentiles in the school. The highest and lowest bars in the plot correspond to the highest and lowest growth percentiles in the schools respectively. The notches around each median represent a rough 95% confidence interval of the median, calculated based on the asymptotic normality of the median (Chambers et al., 1983, p.62). So if the notches of two boxes do not overlap it is strong evidence that the two medians differ. Smaller schools tend to have wider notches. For example, school 209 has 6 students in grade 9 and 26 students in grade 10, therefore in plot (a), the notches of school 209 are wider than the box itself, and in plot (b), the notches are almost as wide as the box. The wide notches are signs that the school sizes are small and results are more dependent on sampling fluctuations.

In Figure 4.13 (b), we plot the growth percentiles of grade 10 conditioning on grade

8 The median growth percentile in a school is found empirically, i.e. we simply aggregate all the conditional growth percentiles in the school in one list, sort them by their values, and find the median. The 75$^{th}$ and the 25$^{th}$ percentiles of the growth percentiles are found the same way.

9 scores for the same 10 schools plus a school 3105. The latter does not have grade 9 and cannot be included in plot (a). Comparing (a) and (b), we notice the differences in the distributions of the growth percentiles in some of the schools. For example, the distribution of growth percentiles in school 24 is positively skewed in (a) and negatively skewed in (b). School 1402 and school 8050 demonstrate noticeable change of skewness in their distributions of growth percentiles as well.

We choose school 10 and 3105 for further comparisons. Figure 4.13 (b) shows that school 10's distribution of growth percentiles has a much higher median than that of school 3105's distribution with non-overlapping notches, and the 25$^{th}$ and 75$^{th}$ percentiles of the former are higher than those of the latter as well. School 10 has 196 students with 10$^{th}$ grade scores, 170 of whom are Hispanic, and 8 are White. State accountability report card shows that this school repeatedly fails the AYP reading and math targets (including the safe harbor targets) during the years of 2003-06, and is put under corrective action. School 3105 has 265 students with 10$^{th}$ grade scores, 245 of whom are White. The school was newly opened in the summer of 2004 and has never failed any AYP target in 2005 and 06. We plot the 9$^{th}$ grade and the 10$^{th}$ grade scores of the students in the two schools in Figure 4.14. The seven black lines in the plot are the SIMEX quantile regression lines estimated with the whole cohort's data—exactly the same as the blue lines in the third plot of figure 4.5. It is clear in the scatter plot that students in school 3105 generally have higher academic achievements than those in school 10 in both grades 9 and 10. There were also 35 students in school 10 who were retained in the 9$^{th}$ grade in 2006. Figure 4.14 plots the 9$^{th}$ grade scores in 2006 against the 9$^{th}$ grade scores in 2005 for 33 of them. The

other 2 repeaters have missing scores. School 3105 has no repeaters in 2006.

Judging from student achievement or school's AYP targets, school 3105 is a far more superior school than school 10, but if the overall student growth in the schools are considered, we will reach different conclusions. Figure 4.15 plots the distributions of student growth percentiles for the two schools. Plot (a) is the density of growth percentiles in grade 10 conditioning on grade 9 scores for the two schools without the repeaters. In plot (b), the repeaters of grade 9 are included, i.e. we calculate their percentiles of grade 9 scores in 2006 conditioning on their grade 9 scores in 2005, and include these conditional percentiles in Figure 4.15 (b). Figure 4.15 (c) plots the growth percentiles of the two schools without repeaters based on the lag-2 models, and plot (d) present the same growth percentiles with the repeaters by calculating their $9^{th}$ grade percentiles in 2006 conditioning on their $8^{th}$ and $9^{th}$ grade scores in 2004 and 05. In other words, the grade 9 repeaters are treated as $10^{th}$ graders in this plot. We are theoretically allowed to do this because test scores from different grades are vertically equated in this particular state. As discussed in the last chapter, the repeaters should be dropped from the present cohort and treated as members of the younger cohort. Here, we include the repeaters in the present cohort to obtain a more conservative evaluation of school 10's growth. Figure 4.15 shows that, whether the repeaters are included or not, and whether the growth percentiles are estimated with the lag-1 or lag-2 models, school 10 has larger proportion of students with higher growth rates and smaller proportion with lower growth rates compared with school 3105.

Of course, test score growth should not be the only standard in evaluating a school.

Other factors such as graduation/drop-out rates are equally important. In fact, we have also found that 37 students who were in school 10 in the year of 2005 disappeared from the data system in 2006. One possible explanation is that they have transferred to schools outside of the state, but given that a considerable proportion of them have a record of low achievement and low growth rates (three quarters of them are below the 17$^{th}$ unconditional percentile in their grade 9 scores, the median of their growth percentile $\hat{P}_{g8|g7}$ is 26, and their median of $\hat{P}_{g9|g8}$ is 33), it is also highly likely that some of them have dropped out from school. The potentially high drop-out rate should be a factor in school evaluation, but it does not mean that the particular school is necessarily at fault. For example, we notice that the missing group generally has low growth percentile in grade 8 before their enrollment in high school 10, which suggest that some of the drop-out students may be already at high risk when they enter the school.

To sum up, Figure 4.15 shows an example that schools that fulfill the AYP requirements are not necessarily demonstrating satisfactory student growth, while schools that repeatedly fail the AYP targets are not necessarily demonstrating poor student growth, either. Note that our models and results are not causal ones, therefore it is not justifiable to conclude that the relatively slow growth of school 3105 students, the relatively fast growth of school 10 students, or the potential high drop-out rate of school 10 students in 2006 are "school effects". We believe that the growth percentiles estimated based on QR models and the SIMEX method proposed in this thesis is a useful tool in diagnosing student growth, is an improvement over the AYP or the Safe Habor provisions in diagnosing school growth, and provides a good starting point for causal

investigations of school effects.

## Chapter 5 Summary and Discussions

### *5.1 The significance of student growth measurement*

The objective of this thesis is to develop a method to measure student academic growth that is sensitive enough to capture small changes, accurate in terms of accounting for test measurement errors, and easily interpretable to all stakeholders. As explained in the first chapter, the significance of student growth measurement in the educational accountability system lies in the following aspects.

First, teachers, parents, and students themselves need growth measurement to diagnose student academic progress in the past and to inform future practices. The improvement of instructional and learning strategies and effort relies heavily on interpretable, sensitive, and accurate individual diagnoses.

Second, student growth measurement is the basis of teacher and school accountability. Individual growth measurement is a necessary condition for the growth measurement at the class or school level, since the latter is simply an aggregation of the

former. Class or school growth measurement, again, is a necessary condition for the estimation of teacher or school effect. If student growth cannot be measured with sensitivity and accuracy, teachers and schools cannot be held accountable on any meaningful basis.

Third, student growth measurement is used in evaluating the achievability of an academic goal for different students and schools. There is no doubt that the same academic expectation should be held for students across different schools, geographic regions, and demographic groups. However, the same expectation entails different amount of effort for different students. Proper evaluations must be conducted to assess different schools and student groups' probabilities of achieving the goals, and adequate assistance must be provided accordingly, otherwise the accountability system can easily turn counterproductive by assigning more punishment than assistance to the disadvantaged schools/student groups. To estimate the probability of reaching proficiency in the $10^{th}$ grade for a $7^{th}$ grader, for example, it is necessary to project the student's academic achievement to 3 years later. The projection process, described in chapter 2, is dependent on the measurement of growth for an older cohort of students who have their full longitudinal records available.

*5.2 Conditional percentile as a measure of student growth*

Measuring student academic growth is not an easy task. Chapter 2 summarizes and discusses some of the usual methods of growth measurement, such as using the change of achievement levels (e.g. from “nonproficient” to “proficient”) to measure growth, and using the difference of scale scores to measure growth. The former method

lacks sensitivity, while the latter is not interpretable and has poor psychometric properties (such as low reliability). Based on a long tradition of pediatric practices (Cole, 1988 & 1994) and recent educational research (Betebenner, 2008), we proposed the concept of "conditional percentile" (or "growth percentile") as a measure of academic growth. The conditional percentile is interpretable for all stakeholders as diagnosis of student's academic progress (or the lack of it), it can be aggregated at the class or school level to form sensitive measurement of class or school growth, and it can be conveniently used to project students' future achievement and to estimate the probability of achieving a certain goal for a given individual. One of the most attractive properties of conditional percentile is that it does not require test scores at different grade levels to be vertically linked.

*5.3 Methodological significance*

Conditional percentiles are estimated using quantile regression in this thesis following the recent breakthrough in the research on the conditional growth chart method (Wei, 2004). Compared with the OLS regression, quantile regression makes less distributional assumptions about the data. The QR model also estimates the conditional percentiles directly, unlike the OLS regression which estimates the conditional mean and variance.

The major methodological contribution of this thesis is the combination of the SIMEX method with the QR model to improve the accuracy in estimating conditional percentiles. Since conditional percentiles are estimated with standardized test scores, and since test scores contain measurement errors, the accuracy of the estimated conditional percentiles is compromised, especially for those with very high or very low test scores.

Previous research has shown the effect of covariate measurement error on the estimation of quantile regression (Chesher, 2001), but few methods have been proposed to correct for measurement error-induced bias in QR. We adopt the simulation-extrapolation method of Cook and Stefanski (1994) to approach this problem. Results of our simulation study show that the SIMEX correction significantly reduce the amount of bias and the magnitude of the mean squared errors in the QR estimators resulted from covariate measurement errors in most scenarios.

*5.4 Major findings*

The fourth chapter is based on the longitudinal data of a specific student cohort (i.e. those who were in grades 7 to 10 in the years 2003-06) from a state assessment program. We analyzed the results of the simple QR models, QR models with SIMEX corrections, and the conditional percentiles produced from these models. There are the following major findings.

First, the distributions of growth in any given year are heteroscedastic. We define the distributions of growth in a given year as the conditional distributions of the test scores in that year given the test scores in the previous year. Heteroscedasticity refers to the fact that the distributions of growth have different variances corresponding to different test scores in the previous year. Specifically, for the students who scored at the lower end in the previous year, the variances of their distributions of growth are larger than those of the students who scored at the higher end. By the tests of slope equality in the QR models, this difference in variances is shown to be statistically significant. We have shown that the heteroscedasticity in the distributions of growth is not completely

explained by the heteroscedasticity in measurement errors. The practical implications of the heteroscedasticity is that low-achievers tend to follow more heterogeneous growth paths than high-achievers do.

Second, the SIMEX method corrects for attenuation in the QR model due to covariate measurement errors. Thus the SIMEX correction leads to steeper slopes and lower intercepts, and the estimated inter-quartile ranges of the distributions of growth also decreases. The practical consequence is that the number of people who have very high estimated conditional percentiles and the number of those with very low conditional percentiles both increase, while the number of students with moderate growth percentiles decrease. In terms of projection for a younger cohort, the SIMEX correction will lead to more students being classified in the "high-risk" group (meaning those who need to grow at extremely high conditional percentiles to reach partial proficiency), more in the "low-risk" group (i.e. those who can reach partial proficiency even with very low conditional percentiles), and less in the "moderate-risk" group. Aside from the change in point estimation, the SIMEX method also leads to wider confidence intervals in the model parameters and predicted values. In other words, the uncertainties of model projections increase after taking the measurement errors into account.

Third, a considerable proportion of students (about one third based on the simple QR results and more than one third based on the SIMEX results) grow at highly fluctuating conditional percentiles from one year to the next. Part of this phenomenon is a regression-to-the-mean effect, and part of it is due to model specification. For instance, after changing from the lag-1 QR models to the "common condition" QR models for the

estimation of the conditional percentiles, the degrees of the fluctuations lessen noticeably.

Finally, we use examples to illustrate the use of the conditional growth chart method in diagnosing student growth. The student and school examples show that the conditional percentiles contain immediately interpretable information about student academic progress, and when aggregated to the school level, they also provide valuable information about school growth. This type of diagnostic information is not easily obtainable from other sources of growth measurement, such as the change of scale scores, achievement levels, and unconditional percentiles.

Besides diagnosis, the conditional growth chart can also be used for projection. Consider a present seventh grader who got the same score as the student in figure 4.11 did in grade 7. If we are willing to assume that the cohort of the former does not differ significantly from the cohort of the latter, the three growth paths in the right plot of figure 4.11 project the present seventh grader's position in grade 10 based on different levels of effort. The projection shows that the student needs to make an exceptional amount of consistent effort in order to reach partial proficiency in three years. For a particular school that has a large number of students with similar scores, the projection suggests that the state needs to provide tremendous amount of instructional support and tutorial assistance to the school in question before expecting the school to reach the state standards.

*5.5 Policy Implications*

To sum up, the conditional percentile as measure of student academic growth effectively serves the purpose of diagnosis and projection for educators and policy

makers. Based on the analysis of this thesis, the following cautions should be taken when adopting this concept in the educational accountability system. First, for each student, it is advisable to estimate several conditional percentiles based on different model specifications (such as the lag-1, lag-2, and the "common condition" models etc.). The different quantities carry different estimation advantages, disadvantages, and slightly different information, and will supplement each other to provide more comprehensive diagnostic pictures. Second, conditional percentiles from a single year can be used for diagnosis but not for accountability. As we have shown, growth fluctuations are prevalent. Slow growth during a single year for a student or a school does not necessarily mean that the student or school has failed the growth expectations. Longitudinal data from several years should be used when judging the growth of a school. Third, the accuracy of conditional percentiles is affected by the measurement errors in the test scores. Besides improving the reliability of the tests, the SIMEX method can be used in combination with the QR models to produce more accurate results.

*5.6 Limitations and directions for future studies*

The study done for this thesis has many limitations. One of the major methodological limitations lies in the fact that we were not able to build a QR model with random effects to account for the nested structure of the data. The simulation study presented in chapter 3 shows that the Bayesian QR model with random effects does not perform well with nested data, the reasons for which are not completely understood. More research also needs to be done to understand the exact consequences of nested data on the QR model estimation.

Another methodological limitation lies in the variance estimation of the fitted values of the QR models with SIMEX corrections. As explained in note 7 in chapter 4, the estimation of the fitted value's variance relies on the variance estimation of the intercept and the slope, and the estimation of the covariance between the intercept and the slope. The variance estimation is explained in chapter 3, but the covariance estimation is problematic. We used the estimated covariance between the intercept and the slope in the naïve QR model to substitute for that in the SIMEX model, which results in an overestimation of the fitted value variance. That is to say, the confidence bands in figures 4.6, 4.10, 4.11, and 4.12 are slightly wider than they should be. More study is needed to attain a more accurate and feasible method of estimating the variance of the SIMEX fitted values.

With respect to the application of the methodology of this thesis in educational accountability, one interesting topic for future research is test score projection using the QR model combined with the SIMEX method. We have mentioned that, when projecting the current cohort's test scores into the future using results from a previous cohort, an important assumption is that the two cohorts have similar distributions of growth. The validity of this assumption needs to be studied with more cohorts of testing data. We also discussed the linearity assumption of the QR model in chapter 3. While nonparametric QR will surely demonstrate better model fit than linear QR when applied to the testing data, it is not clear which model is more generalizable and more suitable for projection purposes, and this question should be answered in future studies.

Finally, it is emphasized repeatedly in this thesis that the QR models in this thesis

do not support causal inferences. The estimated growth percentiles are measures of student and school academic growth but do not provide any information about the factors contributing to that growth. Is it possible to incorporate the ideas of the value-added models in quantile regression and separate school effects from the effects of families and neighborhoods? The answer to this question is contingent on the success of building a QR model with random effects, but extends further than that. It is a topic for long-term investigation in the future.

## Reference


Abrevaya, J. & Hausman, J. A. (2004). Response error in a transformation model with an application to earnings equation estimation. *Econometrics Journal,* 7, 366-388.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29 (1), 37-62.

Bassett, G. & Koenker, R. (1982). Tests of linear hypotheses and L1 estimation. *Econometrica*, 50, 1577-83.

Bhattacharya, P. K. (1963). On an analog of regression analysis. *The Annals of Mathematical Statistics*, 34 (4), 1459-73.

Bhattacharya, P. K., & Gangopadhyay, A. K. (1990). Kernel and nearest-neighbor estimation of a conditional quantile. *The Annals of Statistics,* 18 (3), 1400-1415.

Betebenner, D. W. (2007) Growth as a Description of Process. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. (To appear in

Festschrift dedicated to the life and work of Robert Linn)

Betebenner, D. W. (2008) Toward a Normative Understanding of Student Growth. In K. E. Ryan & L. A. Shepard (Eds.). *The Future of Test-based Educational Accountability (*pp. 155-170). New York: Taylor & Francis.

Betebenner, D. W., Shang, Y, Xiang, Y, Yue, X, & Zhao, Y. (2008). Performance Level Misclassification: Bias and Variability of School Quality Measures. *Journal of Educational Measurement,* 45(2),119-137 .

Braun, H. I. (2005). *Using Student Progress to Evaluate Teachers: A Primer on Value-added Models*. (Tech. Rep.). Princeton, New Jersey: Educational Testing Service.

Braun, H. I. (1988). A New approach to avoiding problems of scale in interpreting trends in mental measurement data. *Journal of Educational Measurement*, 25(3), 171-191.

Briggs, D. C., Weeks, J. P., & Wiley, E. (2008). The impact of vertical scaling decisions on growth projections. (Paper Presented at the 2008 Annual Conference of the National Council for Measurement in Education).

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. New York: Springer-Verlag.

Buuren, S. v. (2007). *Worm plot to diagnose fit in quantile regression.* Statistical Modeling, 7(4), 363-376.

Carey, V. J., Yong, F. H., Frenkel, L. M., & McKinney, R. M. (2003) Growth velocity assessment in paediatric AIDS: smoothing, penalized quantile regression and the definition of growth failure. *Statistics in Medicine*, 23(3), 509-526

Carlson, D. (2002). Focusing state educational accountability systems: Four methods for judging school quality and progress. In W. J. Erbpenbach (Ed.). *Incorporating multiple measures of student performance into state accountability systems: A compendium of resources* (pp. 285-297). Washington DC: Council of Chief State School Officers.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4), 305-331.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective.* Boca Raton, FL: Chapman & Hall/CRC.

Carroll, R. J., Maca, J. D., & Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika*, 86(3), 541-554.

Carroll, R. J., Kuchenhoff, H., Lombard, F., & Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement models. *Journal of the American Statistical Association*, 91, 242-250.

Center for Education Policy. (2007). *Answering the question that matters most: Has student achievement increased since No Child Left Behind?* Washington, DC: Author.

Chesher, A. (2001). Parameter approximations for quantile regressions with measurement error. Working paper CWP02/01, Department of Economics, University College London.

Cole, T. J. (1988). Fitting Smoothed Centile Curves to Reference Data. *Journal of the Royal Statistics Society, Series A*, 151, 385-418.

Cole, T. J. (1994). Growth Charts for Both Cross-Sectional and Longitudinal Data. *Statistics in Medicine*, 13, 2477-2492.

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89, 1314-1328.

Cullen, J. B. & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system. WP 12286, National Bureau of Economic Research (June).

Devanarayan, V. (1996). *Simulation extrapolation method for heteroscedastic measurement error models with replicate measurements.* Unpublished Ph.D. Thesis, North Carolina State University, Raleigh, NC.

Doran, H. C. (2004, June). Value-added models and adequate yearly progress: Combining growth and adequacy in a standards-based environment. (Paper Presented at the 2004 Annual CCSSO Large-Scale Assessment Conference).

Doran, H. C., & Cohen, J. (2005). The confounding effects of linking bias on gains estimated from value-added models. In R. W. Lissitz (Ed.), *Value Added Models in Education: Theory and Applications* (pp. 80-111). Maple Grove, MN: JAM Press.

Doran, H. C., & Jiang, T. (2006). The impact of linking error in longitudinal analysis: An empirical demonstration. In R. W. Lissitz (Ed.), *Longitudinal and Value Added Models of Student Performance* (pp. 210-230). Maple Grove, MN: JAM Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Ferrara, S., & DeMauro, G. E. (2006). Standardized Assessment of Individual Achievement in K-12. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 579-621). Westport, CT: American Council on Education and Praeger Publishers.

Figlio, D. N. (2005). Testing, crime and punishment. WP 11194, National Bureau of Economic Research (March).

Figlio, D. N. & Getzler, L. S. (2002). Accountability, ability and disability: Gaming the system. WP 9307, National Bureau of Economic Research (October).

Figlio, D. N. & Rouse, C. E. (2005). Do accountability and voucher threats improve low-performing schools? WP 11597, National Bureau of Economic Research (August).

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues & Practice*, 10(3), 3-9, 16

Fuller, B., Wright, J., Gesicki, K, & Kang E. (2007). Gauging growth: How to judge No Child Left Behind? *Educational Researcher*, 36, 268-278.

Fuller, W. A. (1987). *Measurement Error Models*. New York: John Wiley & Sons.

Gannoun, A., Girard, S., Guinot, C., & Saracco, J. (2002). Reference curves based on non-parametric quantile regression. *Statistics in Medicine*, 21, 3119-3135.

Geraci, M. & Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, 8, 140-154.

Greene, W. H. (2000). *Econometric Analysis.* Prentice Hall, Inc.

Gutenbrunner, C., & Jureckova, J. (1992). Regression rank scores and regression quantiles. *Annals of Statistics,* 20, 305-330.

Gutenbrunner, C., Jureckova, J., Koenker, R., & Portnoy, S. (1993). Tests of linear hypotheses based on regression rank scores. *Journal of Nonparametric Statistics*, 2, 307-331.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data in accountability testing. Yearbook of the National Society for the Study of Education* (Vol. 104, Part I, pp. 1-34). Boston, MA: Blackwell Publishing.

Hamill, P. V. V., Dridzd, T. A., Johnson, C. L., Reed, R. B., Roche, A. F. &Moore, W. M. (1979). Physical growth: National Center for Health Statistics percentiles. *American Journal of Clinical Nutrition*, 32, 607-629.

Hanushek, E. A., & Raymond, M. E. (2002). Improving educational quality: How best to

evaluate our schools? Retrived on Aug. 22nd, 2008 from http://edpro.stanford.edu/hanushek/admin/pages/files/uploads/accountability.BostonFed.final%20publication.pdf

Hanushek, E. A., & Raymond, M. E. (2003). Lessons about the design of state accountability systems. In P. E. Peterson and M. R. West (Eds.) *No Child Left Behind? The Politics and Practice of Accountability*. Washington, DC: Brookings: 127-151.

Hao, L. & Naiman, D. Q. (2007). *Quantile Regression*. Thousand Oaks, CA: Sage Publications.

Hardle, W., Muller, M., Sperlich, S., & Werwatz, A. (2004). Nonparametric and Semiparametric Models. Berlin, Germany: Springer.

Hausman, J. (1978). Specification tests in Econometrics. *Econometrica*, 46, 1251-1271.

Hausman, J. & Taylor, W. (1981). Panel data and unobservable individual effects. *Econometrica*, 49, 1377-1398.

He, X., & Ng, P. (1999). Quantile splines with several covariates. *Journal of Statistical Planning and Inference*, 75, 343-352.

He, X., Ng, P., & Portnoy, S. (1998). Bivariate quantile smoothing splines. *Journal of Royal Statistical Society B*, 60, 537-550.

Holland, P. W. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81, 945-960.

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, 68, 123-149.

Holland, P. W., & Dorans, N. J. (2006) Linking and Equating. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 187-220). Westport, CT: American Council on Education and Praeger Publishers.

Hsiao, C. (2003). *Analysis of Panel Data*. Cambridge, UK: Cambridge University Press.

Jacob, B. A. (2002). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. WP 8968, National Bureau of Economic Research (June).

Jacob, B. A. (2007). Test-based accountability and student achievement: An investiation of differential performance on NAEP and state assessments. WP 12817, National Bureau of Economic Research (January).

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores: Implications for test-based accountability systems. In Diane Ravitch (Ed.) *Brookings Papers on Education Policy, 2002* Washington DC: Brookings Institution.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33-50.

Koenker, R., & Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association.* 94, 1296-1310.

Koenker, R., Ng, P., & Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*. 81(4), 673-680.

Koenker, R. & Xiao, Z. (2002). Inference on the quantile regression process. *Econometrica*, 70, 1583-1612.

Koenker, R. (2005). *Quantile Regression.* Cambridge, UK: Cambridge University Press.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 155-186). Westport, CT: American Council on Education and Praeger Publishers.

Kolen, M. J. (2001). Linking assessments effectively: Purpose and design. *Educational Measurement: Issues and Practice,* 20(1), 5-19.

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

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, 29(4), 285-307.

Kuchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, 62, 85-96.

Ladd, H. F. & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education Review*, 21, 1-17

Linn, R. L. (2006). *Educational* a*ccountability systems* (Tech. Rep.). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, CRESST.

Linn, R. L. (2005a). *Test-based educational* a*ccountability in the era of No Child Left Behind* (Tech. Rep.). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, CRESST.

Linn, R. L. (2005b). *Issues in the design of* a*ccountability systems.* (Tech. Rep.). Los

Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, CRESST.

Linn, R. L. (2003). *Accountability: Responsibility and reasonable expectations* (Tech. Rep.). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, CRESST.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6(1), 83-102.

Linn, R. L., & Haug, C. (2002) Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis*, 24 (1), 29-36.

Lissitz, R. W., Doran, H., Schafer, W.D., & Willhoft, J. (2006). Growth Modeling, Value Added Modeling and Linking: An Introduction. In R. W. Lissitz (Ed.), *Longitudinal and Value Added Models of Student Performance* (pp. 1-46). Maple Grove, MN: JAM Press.

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) (Retrieved Jan. 14, 2008 from the World Wide Web: http://pareonline.net/getvn.asp?v=8&n=10)

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, 32(2), 125-150.

Lord, F. M. (1956). The measurement of growth. *Educational and psychological measurement*, 16, 421-437.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley Publishing Co.

Martineau, J. A. (2006). Distorting value added: the use of longitudinal, vertically scaled student achievement data for growth-based value-added accountability. *Journal of Educational and Behavioral Statistics*, 31(1), 35-62.

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), 67-102.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value Added Models for Teacher Accountability* (Report for the Carnegie Corporation). Santa Monica: Rand Corporation.

McDonnell, L. M. (2005). Assessment and accountability from the policymaker's

perspective. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data in accountability testing. Yearbook of the National Society for the Study of Education* (Vol. 104, Part I, pp. 35-54). Boston, MA: Blackwell Publishing.

Meng, X. L. & Van Dyk, D. (1998). Fast EM-type implementations for mixed effects models. *Journal of the Royal Statistical Society, Series B*, 60, 559-68.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and practice.* Princeton, NJ: Educational Testing Service Policy Information Center.

O'Day, J. A. (2004). Comlexity, accountability, and school improvement. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning Accountability Systems for Education* (pp. 15-46). New York: Teachers College Press.

Pashley, P. J., & Phillips, G. W. (1993). *Toward world-class standards: A research study linking international and national assessments*. Princeton, NJ: Educational Testing Service.

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

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29(1), 121-129.

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

Rogosa, D. (2002). Irrelevance of reliability coefficients to accountability systems: Statistical disconnect in Kane-Staiger "Volatility in School Test Scores". Retrived on Aug. 2nd, 2008 from http://www-stat.stanford.edu/~rag/api/kscresst.pdf

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.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688-701.

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.

Schemo, D. J. (2007). Failing schools strain to meet U.S. Standard. *New York Times*, Oct,

16.

Schennach, S. M. (forthcoming). Quantile regression with mismeasured covariates. *Econometric Theory.*

Schmidt, W. H., Houang, R. T., and McKnight, C. C. (2005). Value-added research: Right idea but wrong solution? In R. W. Lissitz (Ed.), *Value-Added Models in Education: Theory and applications* (pp. 145-164). Maple Grove, MN: JAM Press.

Smith, R. L., & Yen, W. M. (2006). Models for Evaluating Grade-to-grade Growth. In R. W. Lissitz (Ed.), *Longitudinal and Value Added Models of Student Performance* (pp. 82-94). Maple Grove, MN: JAM Press.

Stefanski, L. A., & Cook, J. (1995). Simulation extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 90, 1247-1256.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement,* 7, 201-210.

Stiefel, L., Schwartz, A. E., Rubenstein, R., & Zabel, J. (2005). Measuring school efficiency: What have we learned? In L. Stiefel, A. E. Schwartz, R. Rubenstein, & J. Zabel (Eds.), *Measuring School Performance and Efficiency: Implications for Practice and Research* (pp. 1-17). Larchmont, NY: Eye On Education, Inc.

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M., Roth, J., Ariet, M., Fisher, T., & 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.

Weeks, J. P., & Karkee, T. (2008). From status to growth: The impact on school accountability ratings. Paper presented at the annual meeting of the National Council on Measurement in Education, March 26, 2008, New York, NY.

Wei, Y. (2004). Longitudinal Growth Charts Based on Semi-parametric Quantile Regression. Ph.D. dissertation, University of Illinois at Urbana-Champaign.

Wei, Y. & He, X. (2006). Conditional Growth Charts. *The Annals of Statistics*, 34(5), 2069-2097.

Wei, Y., Pere, A., Koenker, R., & He, X. (2006). Quantile Regression Methods for Reference Growth Charts. *Statistics in Medicine*, 25, 1369-1382.

Weisberg, S. (2005). *Applied Linear Regression*. Hoboken, N. J.: John Wiley & Sons, Inc.

Wilcoxon, F. (1945). Individual Comparisons by ranking methods. *Biometrika*, 1, 80-83.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, 15, 345-422.

Williams, V. S. L., Rosa, K. R., McLeod, L. D., Thissen, D., & Sanford, E. E. (1998). Projecting to the NAEP scale: Results from the North Carolina End-of-Grade Testing Program. *Journal of Educational Measurement,* 35, 277-296.

Wilson, D. & Piebalga, A. (2008). Accurate performance measure but meaningless ranking exercise? An analysis of the English school league tables. Retrieved on Aug. 2nd, 2008 from http://www.bris.ac.uk/Depts/CMPO/workingpapers/wp176.pdf

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press.

Yen, W. M. (1993). Scaling performance assessment: Strategies for managing local item idependence. *Journal of Educational Measurement,* 30, 187-213.

Yu, K., & Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54, 437-447.

## Chapter 2

### Figure 2.1 Unconditional Growth Charts of Height and Weight

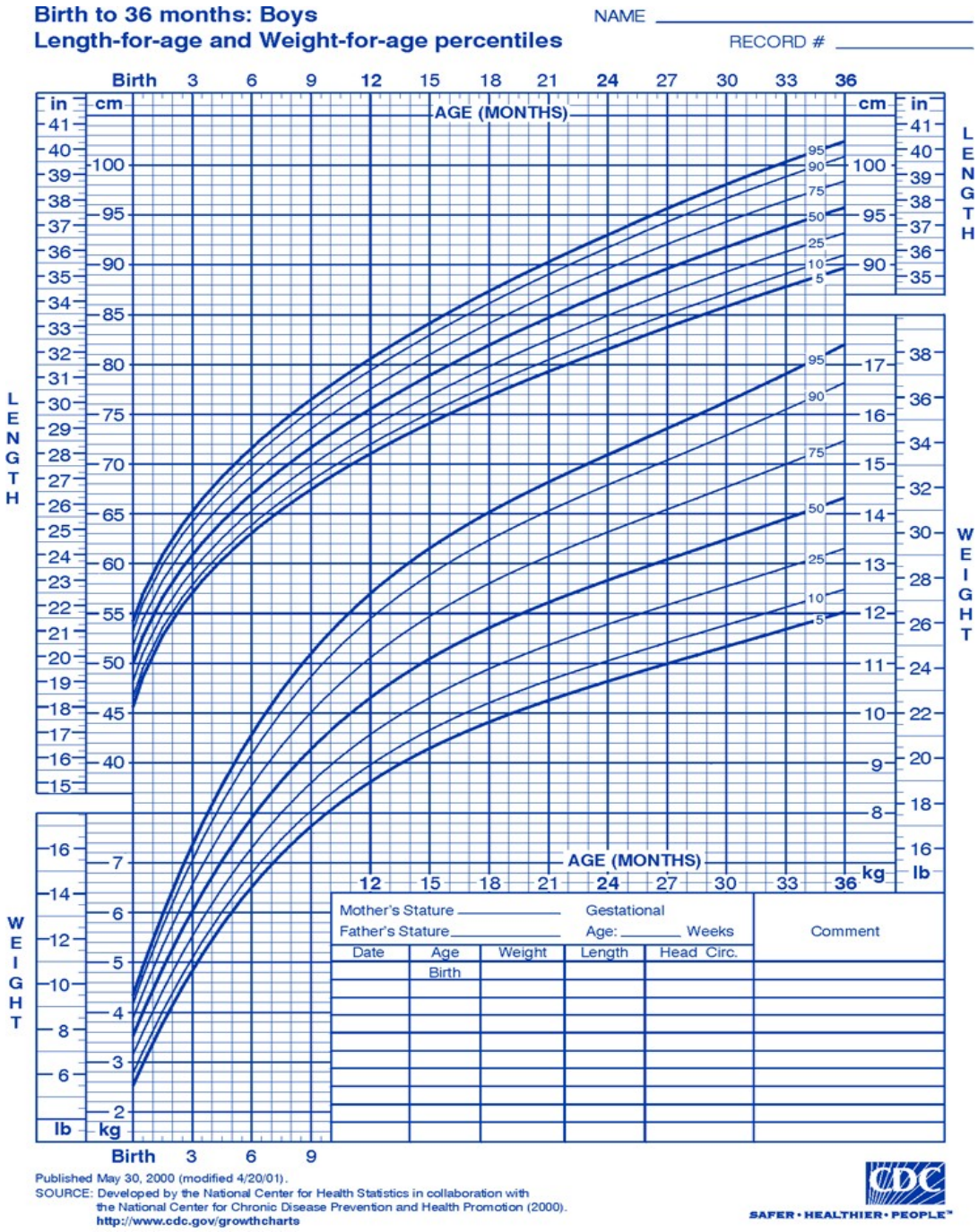
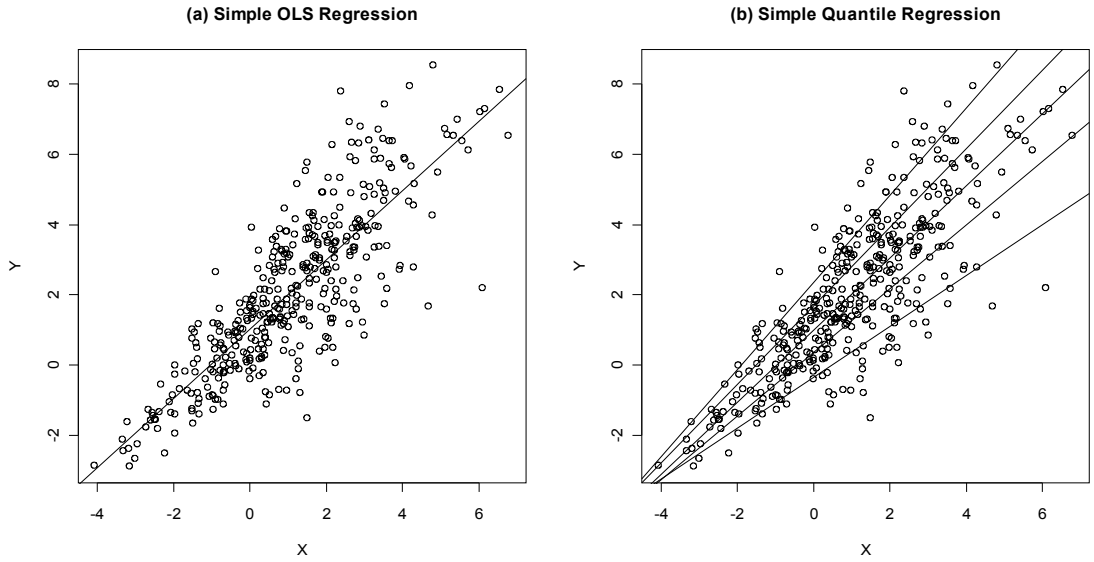


Figure 2.2 Comparison of Simple OLS regression and QR




Data simulated in Figure 2.2 (a) and (b):

$$X \sim N(1,4) \quad , \quad e \sim N(0,1) \quad , \quad Y = 1 + X + (1 + 0.2\mathrm{X})e$$

Four hundred data points are randomly generated for *X* and *e* according to the distributional specification.

## Chapter 3

Table 3.1 Descriptive Statistics of the State Assessment Data

| | 2003 (G 7) | 2004 | 2005 | 2006 |
|---|---|---|---|---|
| Number of Students with Non-missing Scores | 53210 | 53225 | 48788 | 43352 |
| Percent Male | 51.11 | 51.01 | 50.77 | 50.00 |
| Percent Native American | 1.07 | 1.10 | 1.06 | 0.94 |
| Percent Asian | 2.79 | 2.82 | 2.88 | 3.02 |
| Percent Hispanic | 22.70 | 22.95 | 22.51 | 20.49 |
| Percent Black | 5.85 | 5.83 | 5.61 | 5.31 |
| Percent White | 67.58 | 67.30 | 67.95 | 70.25 |
| Percent Proficient | 63.55 | 65.62 | 70.04 | 72.26 |
| Percent Partial Proficient | 86.18 | 88.17 | 93.07 | 92.27 |
| Number of 1st time repeaters | | 402 | 250 | 1097 |
| Number of 2nd time repeaters | | | 6 | 9 |

Table 3.2 Results from the simulation study to evaluate the performance of the full Bayesian estimators of QR models with random effects.

| Model | | $u \sim N(0,1)$ $\epsilon \sim N(0,1)$ | | | | $u \sim N(0,1)$ $\epsilon \sim t(3)$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\beta}_0$ | | $\hat{\beta}_1$ | | $\hat{\beta}_0$ | | $\hat{\beta}_1$ | |
| | | $\hat{bias}$ | $\hat{mse}$ | $\hat{bias}$ | $\hat{mse}$ | $\hat{bias}$ | $\hat{mse}$ | $\hat{bias}$ | $\hat{mse}$ |
| τ = 0.25 | QR | -0.005 | 0.041 | 0.001 | 0.0006 | -0.009 | 0.062 | 0.0004 | 0.0010 |
| | QRRE | 0.263 | 0.601 | 0.0007 | 0.0003 | 0.336 | 2.702 | 0.001 | 0.0006 |
| τ = 0.5 | QR | 0.021 | 0.042 | -0.0030 | 0.0005 | 0.016 | 0.045 | -0.001 | 0.0006 |
| | QRRE | -0.036 | 0.353 | -0.0020 | 0.0003 | 0.068 | 0.910 | -0.001 | 0.0004 |
| τ = 0.75 | QR | 0.0003 | 0.046 | -0.0002 | 0.0006 | -0.002 | 0.057 | 0.002 | 0.0009 |
| | QRRE | -0.290 | 0.512 | 0.0004 | 0.0005 | -0.272 | 0.830 | 0.003 | 0.0006 |

Figure 3.1 Distribution of reading scores in grades 7-10 during the years of 2003-06

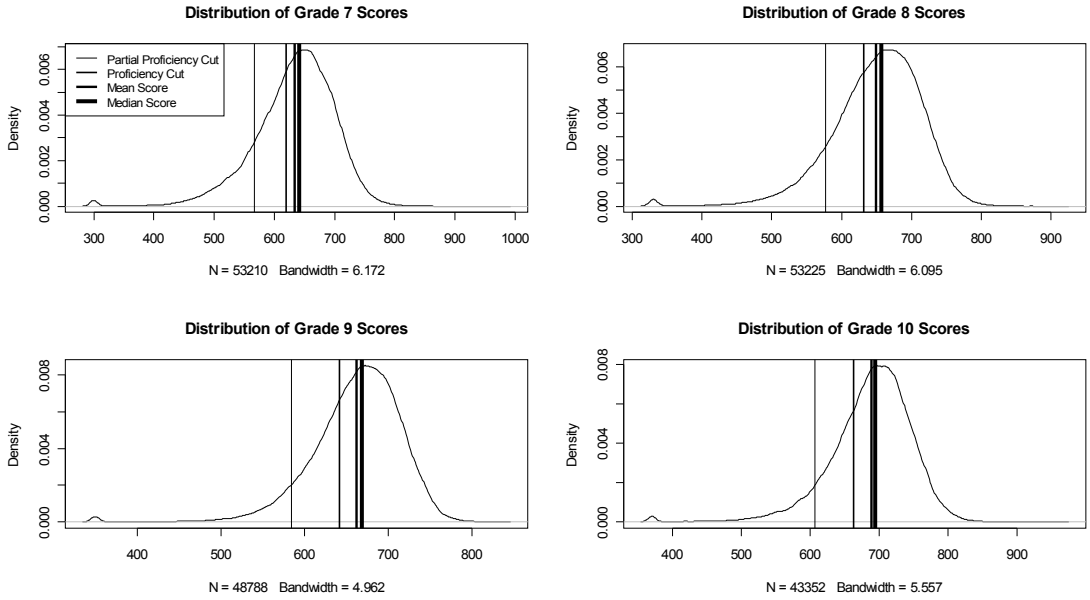



Figure 3.2 Conditional standard errors of measurement plotted against the scale scores for each grade.

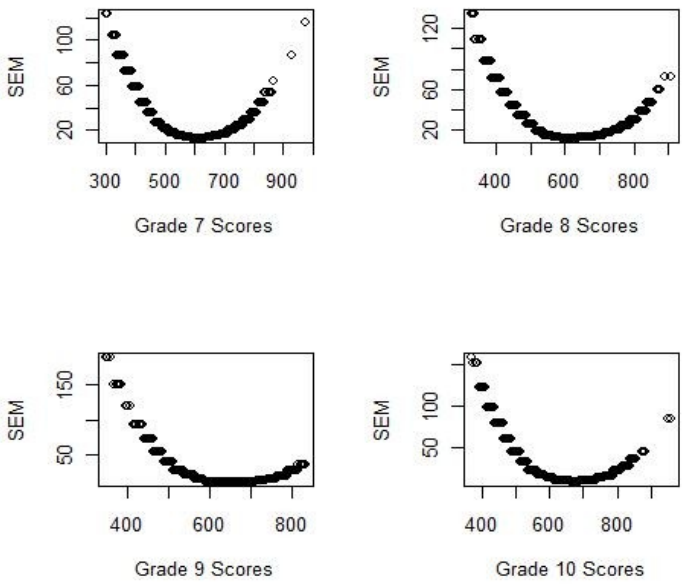


Figure 3.3 (a) The dependence of $\hat{\theta}(\lambda)$ on $\lambda$ for a simple OLS regression

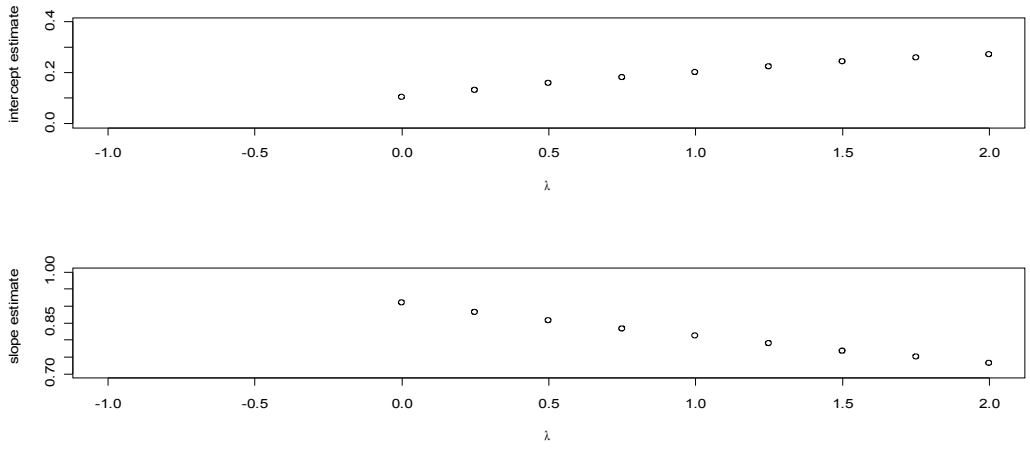
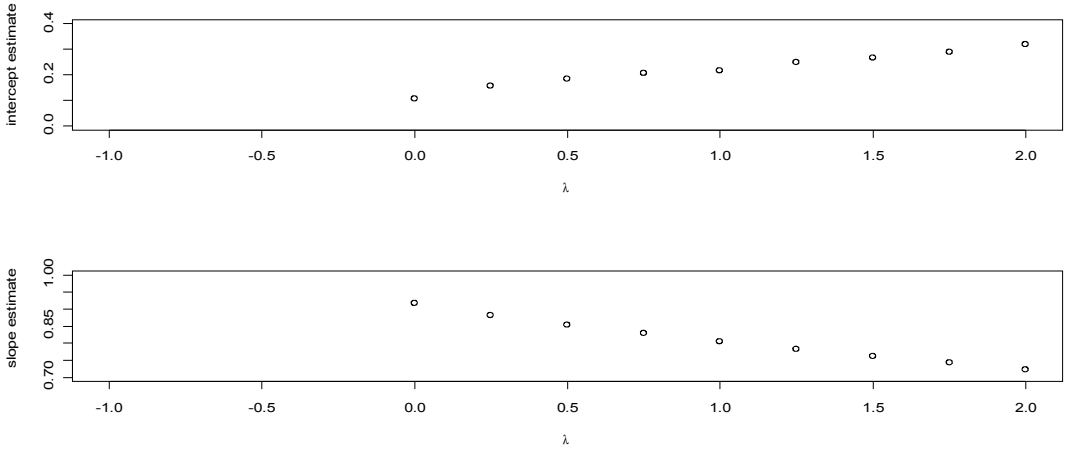



Figure 3.3 (b) The dependence of $\hat{\theta}(\lambda)$ on $\lambda$ for a simple QR model at $\tau = 0.5$



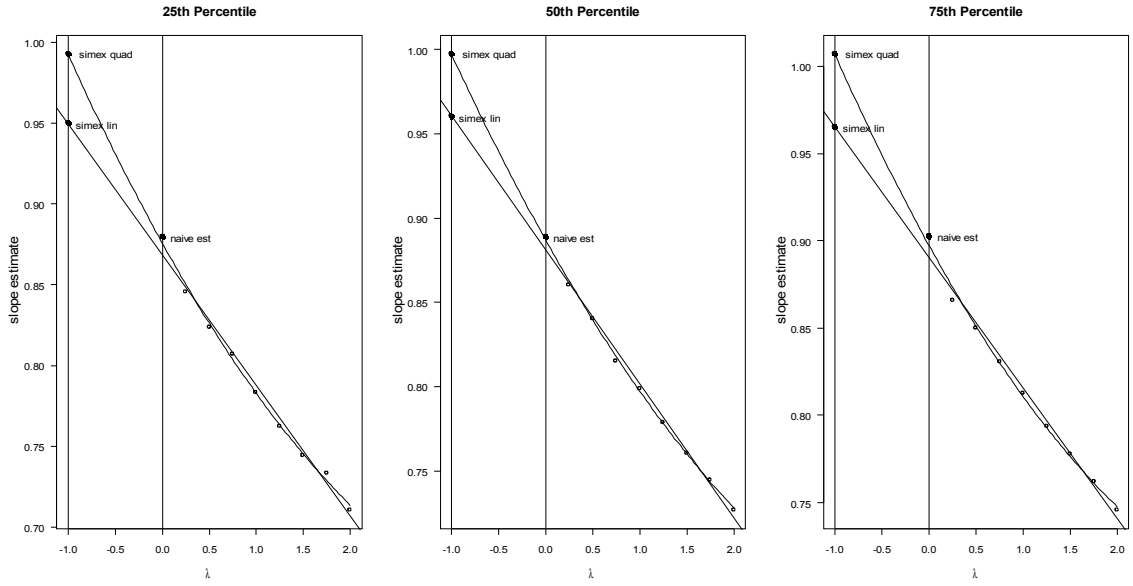

Simulated Data:

$$Y = \beta_0 + \beta_1 \dot{X} + e \qquad X = \dot{X} + v \qquad \dot{X} \sim N(1,9) \qquad e \sim N(0,1) \qquad v \sim N(0,1)$$

$$\beta_0 = 0 \qquad \beta_1 = 1$$

Four hundred data points are randomly generated for *X, v,* and *e* according to the distributional specification.

Figure 3.4 The SIMEX estimates of quantile regression slopes at three different quantiles based on linear and quadratic extrapolations compared with the naïve estimate. The true slope is 1 for all quantiles. (The average estimate at each λ value is generated with $B = 200$ . The "naive est" dot in each plot that lies at λ = 0 is the original naïve estimate of the slope before applying simulation and extrapolation. The "simex lin" dot that lies at λ = -1 is the SIMEX estimate of the slope based on linear extrapolation, and the "simex quad" dot that lies at λ = -1 is the SIMEX estimate of the slope based on quadratic extrapolation.)




Simulated Data:

$$Y = \beta_0 + \beta_1 \dot{X} + e \qquad X = \dot{X} + v \qquad \dot{X} \sim N(1,9) \qquad e \sim N(0,1) \qquad v \sim N(0,1)$$

$$\beta_0 = 0 \qquad \beta_1 = 1$$

Four hundred data points are randomly generated for *X, v,* and *e* according to the distributional specification.

Figure 3.5 Quantile regression slopes with error-free predictor and with error-prone predictor which contains heterogeneous measurement errors.

Simulated Data:

$y_i = \beta_0 + \beta_1 x.true_i + e_i$ , $i = (1, \ldots, 400)$ , $x.true_i \sim N(1,9)$ , $e_i \sim N(0,1)$ ,
$\beta_0 = 0$ , $\beta_1 = 1$ , $x.observed_i = x.true_i + v_i$ , $v_i \sim N(0, |x.true_i|)$

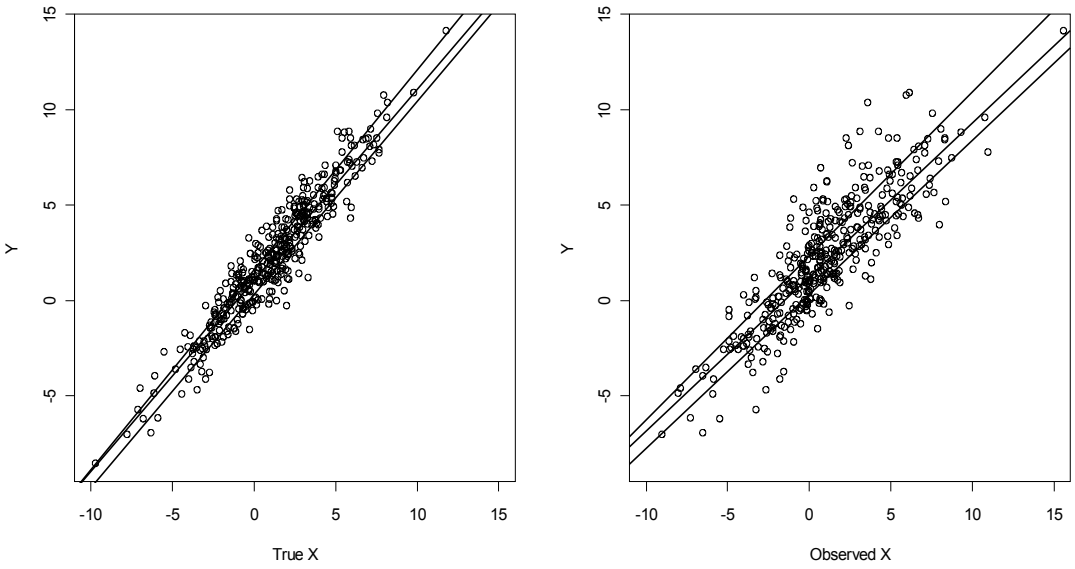
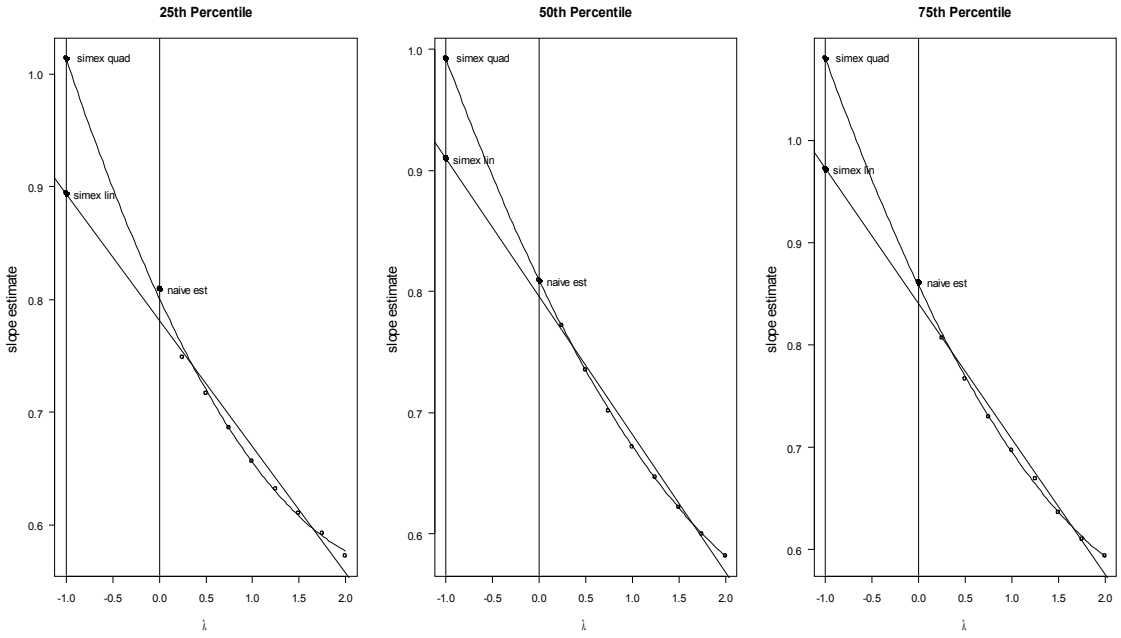


Figure 3.6 The SIMEX estimates of quantile regression slopes at three different quantiles with heteroscedastic measurement error in the predictor. The data used is generated in Figure 3.5. The true slope is 1 for all quantiles. (The average estimate at each λ value is generated with $B = 200$ . The "naive est" dot in each plot that lies at λ = 0 is the original naïve estimate of the slope before applying simulation and extrapolation. The "simex lin" dot that lies at λ = -1 is the SIMEX estimate of the slope based on linear extrapolation, and the "simex quad" dot that lies at λ = -1 is the SIMEX estimate of the slope based on quadratic extrapolation.)



## Chapter 4

Table 4.1 Results of the Quantile Regressions of reading scores in grades 8, 9, and 10 regressed on scores one year earlier (standard errors in brackets)

| Model | | τ | | | | | | | Sample Sizes N & Tests of Equality of Slopes F |
|---|---|---|---|---|---|---|---|---|---|
| | | 0.03 | 0.1 | 0.25 | 0.5 | 0.75 | 0.9 | 0.97 | |
| | α | 591.58 | 615.59 | 633.91 | 651.54 | 668.56 | 684.78 | 703.21 | N = 52471 |
| Grade | | [.61] | [.27] | [.17] | [.14] | [.15] | [.21] | [.41] | |
| 8~7 | β | 1 | 0.91 | 0.87 | 0.84 | 0.81 | 0.78 | 0.72 | F = 169.8 *** |
| | | [0.01] | [0.004] | [0.002] | [0.002] | [0.002] | [0.003] | [0.01] | DF = 6 |
| | $R^1$ | 0.52 | 0.54 | 0.55 | 0.54 | 0.51 | 0.47 | 0.41 | |
| | α | 612.45 | 632.66 | 647.67 | 661.79 | 675.54 | 688.64 | 704.07 | N = 48312 |
| Grade | | [.56] | [.22] | [.14] | [.12] | [.13] | [.18] | [.38] | |
| 9~8 | β | 0.85 | 0.78 | 0.74 | 0.71 | 0.68 | 0.64 | 0.57 | F = 262.28 *** |
| | | [0.01] | [0.003] | [0.002] | [0.002] | [0.002] | [0.002] | [0.01] | DF = 6 |
| | $R^1$ | 0.48 | 0.52 | 0.54 | 0.53 | 0.51 | 0.47 | 0.4 | |
| | α | 621.94 | 650.07 | 668.29 | 685.41 | 701.76 | 717.39 | 736.48 | N = 42796 |
| Grade | | [.99] | [.30] | [.20] | [.15] | [.18] | [.24] | [.52] | |
| 10~9 | β | 1.28 | 1.09 | 1.02 | 0.97 | 0.94 | 0.89 | 0.8 | F = 151.7 *** |
| | | [0.02] | [0.01] | [0.004] | [0.003] | [0.003] | [0.01] | [0.01] | DF = 6 |
| | $R^1$ | 0.43 | 0.48 | 0.5 | 0.49 | 0.46 | 0.43 | 0.37 | |

*** P < 0.001

Table 4.2 (a) Sample sizes, sample standard deviations, and range of standard errors of measurement (SEM) of two groups in each grade. The first group scored between 541 and 551 in grade 7, and the second group scored between 709 and 719 in grade 7.

| | Grade 7 score ϵ [541,551] | | | Grade 7 score ϵ [709,719] | | |
|---|---|---|---|---|---|---|
| | N | sd | SEM | N | sd | SEM |
| grade 7 | 1046 | 3.17 | 15 | 1882 | 3.18 | [17,20] |
| grade 8 | 1015 | 42.26 | [13,134] | 1871 | 25.44 | [13,47] |
| grade 9 | 878 | 36.6 | [11,189] | 1784 | 21.94 | [11,26] |
| grade 10 | 684 | 50.46 | [12,158] | 1721 | 29.05 | [18,37] |

(b) Results from the same groups in (a) after deletion of outliers

| | Grade 7 score ϵ [541,551] | | | Grade 7 score ϵ [709,719] | | |
|---|---|---|---|---|---|---|
| | N | sd | SEM | N | sd | SEM |
| grade 7 | 1046 | 3.17 | 15 | 1882 | 3.18 | [17,20] |
| grade 8 | 1002 | 33.45 | [12,44] | 1871 | 25.44 | [13,47] |
| grade 9 | 852 | 25.33 | [11,20] | 1784 | 21.94 | [11,26] |
| grade 10 | 659 | 35.14 | [10,34] | 1721 | 29.05 | [18,37] |

Table 4.3 Results of the lag-2 QR models of grade 9 and 10 reading scores regressed on two previous scores respectively (standard errors in brackets)

| | | τ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | | 0.03 | 0.1 | 0.25 | 0.5 | 0.75 | 0.9 | 0.97 |
| | α | 616.24 | 635.32 | 649.06 | 662.15 | 674.49 | 686.24 | 699.01 |
| Grade | | [.51] | [.19] | [.12] | [.11] | [.12] | [.15] | [.29] |
| 9~8+7 | $\beta_1$ | 0.53 | 0.48 | 0.45 | 0.42 | 0.39 | 0.36 | 0.33 |
| | | [.01] | [.01] | [.004] | [.003] | [.004] | [.004] | [.01] |
| (N=47748) | $\beta_2$ | 0.37 | 0.33 | 0.31 | 0.31 | 0.31 | 0.31 | 0.30 |
| | | [.01] | [.01] | [.004] | [.003] | [.003] | [.004] | [.01] |
| | $R^1$ | 0.51 | 0.55 | 0.57 | 0.57 | 0.55 | 0.52 | 0.47 |
| | α | 624.63 | 651.21 | 668.77 | 684.89 | 700.37 | 714.65 | 731.02 |
| Grade | | [.84] | [.29] | [.18] | [.14] | [.16] | [.21] | [.40] |
| 10~9+8 | $\beta_1$ | 0.84 | 0.71 | 0.65 | 0.61 | 0.58 | 0.53 | 0.46 |
| | | [.03] | [.01] | [.01] | [.01] | [.01] | [.01] | [.02] |
| (N=42481) | $\beta_2$ | 0.39 | 0.34 | 0.33 | 0.32 | 0.32 | 0.35 | 0.37 |
| | | [.02] | [.01] | [.01] | [.004] | [.01] | [.01] | [.01] |
| | $R^1$ | 0.45 | 0.50 | 0.52 | 0.52 | 0.50 | 0.47 | 0.43 |

Table 4.4 Results of the Quantile Regressions of reading scores in grades 8, 9, and 10 regressed on grade 7 scores (standard errors in brackets)

| Model | | τ | | | | | | | Sample Sizes & Tests of Equality of Slopes |
|---|---|---|---|---|---|---|---|---|---|
| | | 0.03 | 0.1 | 0.25 | 0.5 | 0.75 | 0.9 | 0.97 | |
| | α | 609.67 | 631.50 | 647.51 | 662.71 | 677.31 | 690.50 | 705.11 | N = 48072 |
| Grade | | [0.59] | [0.23] | [0.16] | [0.12] | [0.14] | [0.17] | [0.32] | |
| 9~7 | β | 0.81 | 0.74 | 0.70 | 0.67 | 0.64 | 0.61 | 0.56 | F = 411.06 *** |
| | | [0.005] | [0.004] | [0.003] | [0.002] | [0.002] | [0.003] | [0.003] | DF = 6 |
| | $R^1$ | 0.45 | 0.49 | 0.51 | 0.50 | 0.48 | 0.45 | 0.40 | |
| | α | 612.76 | 645.51 | 666.90 | 686.16 | 704.66 | 721.33 | 740.26 | N = 42744 |
| Grade | | [1.02] | [0.38] | [0.21] | [0.17] | [0.18] | [0.25] | [0.46] | |
| 10~7 | β | 0.93 | 0.80 | 0.75 | 0.71 | 0.68 | 0.66 | 0.62 | F = 102.86 *** |
| | | [0.01] | [0.01] | [0.003] | [0.003] | [0.003] | [0.004] | [0.01] | DF = 6 |
| | $R^1$ | 0.35 | 0.40 | 0.43 | 0.42 | 0.41 | 0.38 | 0.34 | |

Table 4.5 Estimated bias and mean squared error for different combinations of quantiles, residual distributions, and covariate variances: QR refers to naïve QR estimates that ignores the existence of measurement errors; SIM.lin refers to SIMEX estimates based on linear extrapolants; SIM.q refers to SIMEX estimates based on quadratic extrapolants; $\tau$ refers to the quantile values; $v$ is measurement error contained in the independent variable X; $\epsilon$ is the model residual; $X = X.true + v$ ; $Y = 1 + X.true + \epsilon$

| Model | | $v \sim N(0,1)$ $\epsilon \sim N(0,1)$ | | | | $v \sim N(0,1)$ $\epsilon \sim \chi^2(2)$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\beta}_0$ | | $\hat{\beta}_1$ | | $\hat{\beta}_0$ | | $\hat{\beta}_1$ | |
| | | $\hat{bias}$ | $\hat{mse}$ | $\hat{bias}$ | $\hat{mse}$ | $\hat{bias}$ | $\hat{mse}$ | $\hat{bias}$ | $\hat{mse}$ |
| *Var (X.true) = 4* | | | | | | | | | |
| | QR | -0.031 | 0.011 | -0.199 | 0.041 | -0.040 | 0.010 | -0.200 | 0.042 |
| τ = 0.25 | SIM.lin | -0.045 | 0.012 | -0.099 | 0.012 | 0.018 | 0.009 | -0.100 | 0.012 |
| | SIM.q | -0.033 | 0.021 | -0.029 | 0.005 | 0.061 | 0.020 | -0.030 | 0.005 |
| | QR | 0.196 | 0.047 | -0.200 | 0.041 | 0.170 | 0.041 | -0.201 | 0.043 |
| τ = 0.5 | SIM.lin | 0.096 | 0.018 | -0.099 | 0.012 | 0.136 | 0.030 | -0.101 | 0.013 |
| | SIM.q | 0.025 | 0.018 | -0.030 | 0.005 | 0.086 | 0.028 | -0.032 | 0.007 |
| | QR | 0.433 | 0.197 | -0.201 | 0.042 | 0.202 | 0.071 | -0.201 | 0.047 |
| τ = 0.75 | SIM.lin | 0.247 | 0.072 | -0.101 | 0.012 | 0.110 | 0.045 | -0.101 | 0.018 |
| | SIM.q | 0.100 | 0.031 | -0.034 | 0.005 | 0.037 | 0.053 | -0.033 | 0.014 |
| *Var (X.true) = 9* | | | | | | | | | |
| | QR | -0.156 | 0.034 | -0.099 | 0.011 | -0.056 | 0.012 | -0.100 | 0.011 |
| τ = 0.25 | SIM.lin | -0.100 | 0.020 | -0.029 | 0.002 | 0.036 | 0.010 | -0.030 | 0.002 |
| | SIM.q | -0.045 | 0.022 | -0.003 | 0.002 | 0.075 | 0.023 | -0.004 | 0.002 |
| | QR | 0.101 | 0.019 | -0.100 | 0.011 | 0.184 | 0.046 | -0.101 | 0.011 |
| τ = 0.5 | SIM.lin | 0.031 | 0.009 | -0.029 | 0.002 | 0.140 | 0.032 | -0.031 | 0.002 |
| | SIM.q | 0.008 | 0.016 | -0.005 | 0.002 | 0.084 | 0.029 | -0.006 | 0.003 |
| | QR | 0.357 | 0.138 | -0.102 | 0.011 | 0.224 | 0.081 | -0.100 | 0.013 |
| τ = 0.75 | SIM.lin | 0.161 | 0.036 | -0.031 | 0.002 | 0.089 | 0.041 | -0.030 | 0.004 |
| | SIM.q | 0.057 | 0.024 | -0.007 | 0.002 | 0.013 | 0.056 | -0.006 | 0.006 |

Table 4.6 Estimated bias and mean squared error for different combinations of quantiles, residual distributions, and covariate variances: QR refers to naïve QR estimates that ignores the existence of measurement errors; SIM.lin refers to SIMEX estimates based on linear extrapolants; SIM.q refers to SIMEX estimates based on quadratic extrapolants; $Y = 1 + X_{1.\text{true}} + X_{2.\text{true}} + \epsilon$ , $X_1 = X_{1.\text{true}} + v_1$ , and $X_2 = X_{2.\text{true}} + v_2$

| Model | | $\hat{\beta}_0$ | | $\hat{\beta}_1$ | | $\hat{\beta}_2$ | |
|---|---|---|---|---|---|---|---|
| | | $\hat{bias}$ | $\hat{mse}$ | $\hat{bias}$ | $\hat{mse}$ | $\hat{bias}$ | $\hat{mse}$ |
| $X_{1.\text{true}} \sim N(1,9)$ | $v_1 \sim N(0,1)$ | $X_{2.\text{true}} \sim N(1,4)$ | | $v_2 \sim N(0,1)$ | | | |
| | QR | -0.132 | 0.034 | -0.100 | 0.011 | -0.201 | 0.043 |
| τ = 0.25 | SIM.lin | -0.136 | 0.035 | -0.030 | 0.002 | -0.100 | 0.013 |
| | SIM.q | -0.088 | 0.044 | -0.006 | 0.003 | -0.033 | 0.007 |
| | QR | 0.298 | 0.103 | -0.101 | 0.011 | -0.199 | 0.042 |
| τ = 0.5 | SIM.lin | 0.128 | 0.031 | -0.030 | 0.002 | -0.099 | 0.012 |
| | SIM.q | 0.033 | 0.031 | -0.006 | 0.002 | -0.028 | 0.006 |
| | QR | 0.729 | 0.548 | -0.100 | 0.011 | -0.199 | 0.042 |
| τ = 0.75 | SIM.lin | 0.391 | 0.170 | -0.029 | 0.002 | -0.099 | 0.013 |
| | SIM.q | 0.161 | 0.063 | -0.005 | 0.003 | -0.030 | 0.007 |
| $X_{1.\text{true}} \sim N(1,9)$ | $v_1 \sim N(0,1)$ | $X_{2.\text{true}} \sim N(1,16)$ | | $v_2 \sim N(0,2)$ | | | |
| | QR | -0.400 | 0.180 | -0.101 | 0.012 | -0.112 | 0.013 |
| τ = 0.25 | SIM.lin | -0.305 | 0.112 | -0.031 | 0.003 | -0.036 | 0.002 |
| | SIM.q | -0.166 | 0.069 | -0.007 | 0.004 | -0.008 | 0.002 |
| | QR | 0.212 | 0.061 | -0.100 | 0.011 | -0.111 | 0.013 |
| τ = 0.5 | SIM.lin | 0.065 | 0.021 | -0.029 | 0.002 | -0.035 | 0.002 |
| | SIM.q | 0.012 | 0.034 | -0.005 | 0.003 | -0.005 | 0.002 |
| | QR | 0.830 | 0.709 | -0.099 | 0.011 | -0.111 | 0.013 |
| τ = 0.75 | SIM.lin | 0.442 | 0.215 | -0.028 | 0.003 | -0.036 | 0.002 |
| | SIM.q | 0.200 | 0.083 | -0.003 | 0.004 | -0.006 | 0.002 |

Table 4.7 Probabilities of growing at or above the 25th, 50th, and 75th conditional percentiles for one year, two years, and three years.

| | $\hat{P}_{g8\|g7} \geq p$ | | | $(\hat{P}_{g8\|g7} \geq p) \cap (\hat{P}_{g9\|g8} \geq p)$ | | | $(\hat{P}_{g8\|g7} \geq p) \cap (\hat{P}_{g9\|g8} \geq p) \cap (\hat{P}_{g10\|g9} \geq p)$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $p$ | Theoretical | Empirical (Simple QR) | Empirical (SIMEX) | Theoretical | Empirical (Simple QR) | Empirical (SIMEX) | Theoretical | Empirical (Simple QR) | Empirical (SIMEX) |
| 25 | 0.758 | 0.755 | 0.717 | 0.575 | 0.515 | 0.447 | 0.436 | 0.349 | 0.275 |
| 50 | 0.505 | 0.505 | 0.502 | 0.255 | 0.192 | 0.177 | 0.129 | 0.075 | 0.059 |
| 75 | 0.253 | 0.255 | 0.287 | 0.064 | 0.032 | 0.037 | 0.016 | 0.005 | 0.005 |

Table 4.8 Percentages of students whose growth rates from consecutive years are different by less than 25, between 25 and 50, between 50 and 75, and above 75

| | | Percentage | | | Percentage |
|---|---|---|---|---|---|
| $\left\|\hat{P}_{g9\|g8}-\hat{P}_{g8\|g7}\right\|<25$ | Simple QR | 36.31 | $\left\|\hat{P}_{g10\|g9}-\hat{P}_{g9\|g8}\right\|<25$ | Simple QR | 37.20 |
| | SIMEX | 31.86 | | SIMEX | 33.41 |
| $25\le\left\|\hat{P}_{g9\|g8}-\hat{P}_{g8\|g7}\right\|<50$ | Simple QR | 29.94 | $25\le\left\|\hat{P}_{g10\|g9}-\hat{P}_{g9\|g8}\right\|<50$ | Simple QR | 30.43 |
| | SIMEX | 27.11 | | SIMEX | 27.84 |
| $50\le\left\|\hat{P}_{g9\|g8}-\hat{P}_{g8\|g7}\right\|<75$ | Simple QR | 21.69 | $50\le\left\|\hat{P}_{g10\|g9}-\hat{P}_{g9\|g8}\right\|<75$ | Simple QR | 21.53 |
| | SIMEX | 23.06 | | SIMEX | 22.78 |
| $\left\|\hat{P}_{g9\|g8}-\hat{P}_{g8\|g7}\right\|\ge 75$ | Simple QR | 12.07 | $\left\|\hat{P}_{g10\|g9}-\hat{P}_{g9\|g8}\right\|\ge 75$ | Simple QR | 10.84 |
| | SIMEX | 17.97 | | SIMEX | 15.97 |

Figure 4.1 (a) Scatter plot of grade 8 reading scores against grade 7 reading scores with seven quantile regression lines at $(\tau = 0.03, 0.1, 0.25, 0.5, 0.75, 0.9, 0.97)$ (b) grade 9 reading scores plotted against grade 8 reading scores with quantile regression lines at the same quantiles (c) grade 10 reading scores plotted against grade 9 scores with seven quantile regression lines. The vertical and horizontal green lines mark the proficiency cut scores in the corresponding grades.

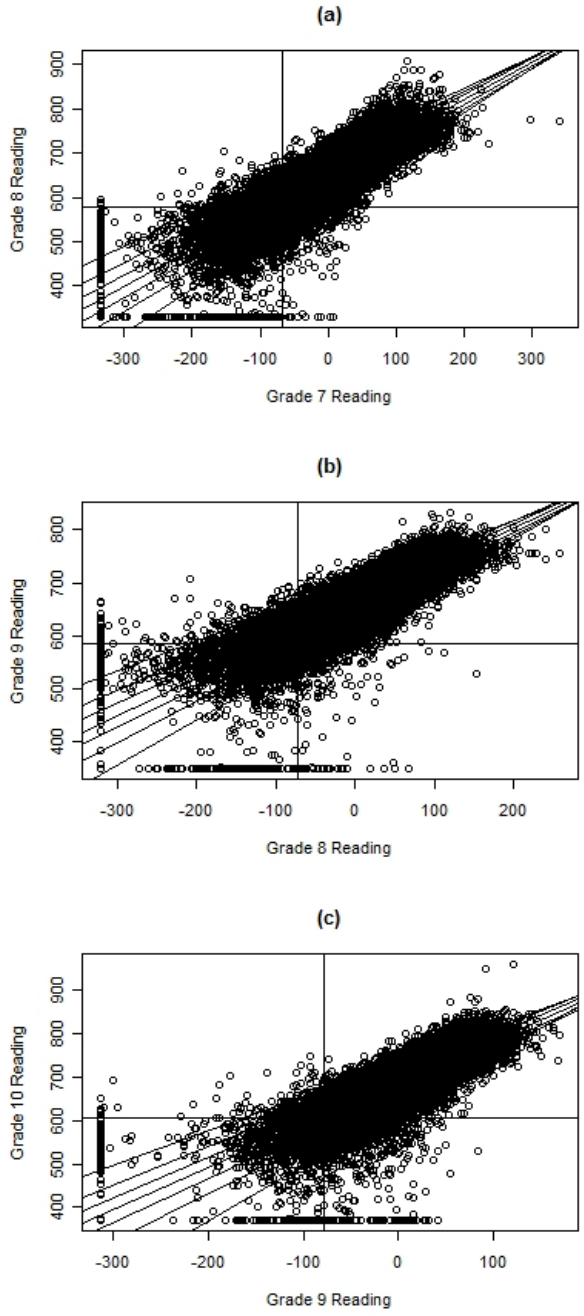


Figure 4.2 Estimated parameters and 95% confidence intervals of the lag-1 Quantile Regressions plotted against the quantile values. The horizontal lines are estimated parameters from OLS models with 95% confidence intervals.

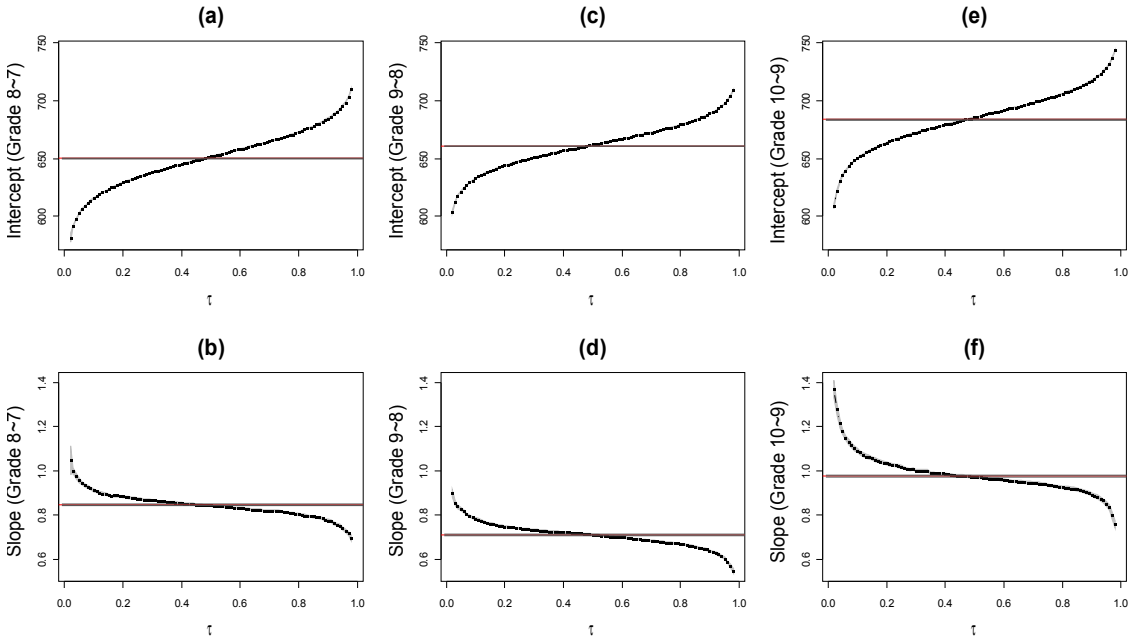
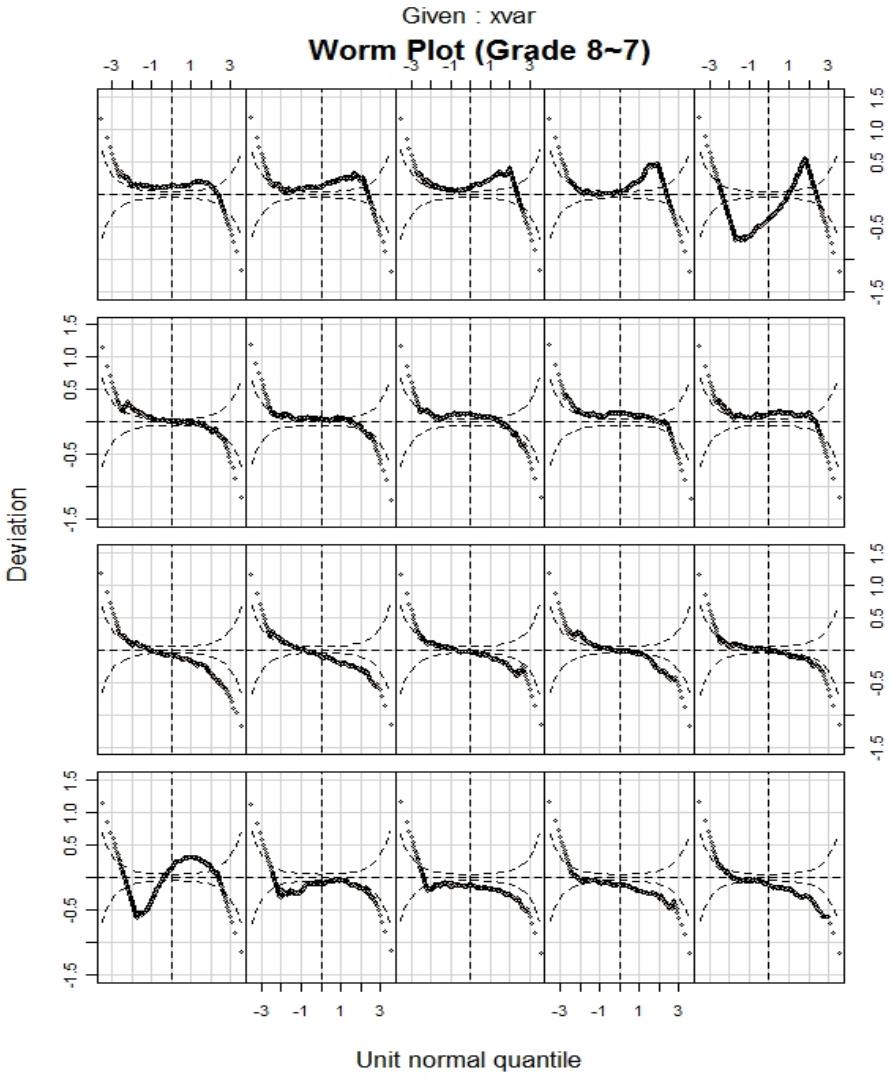


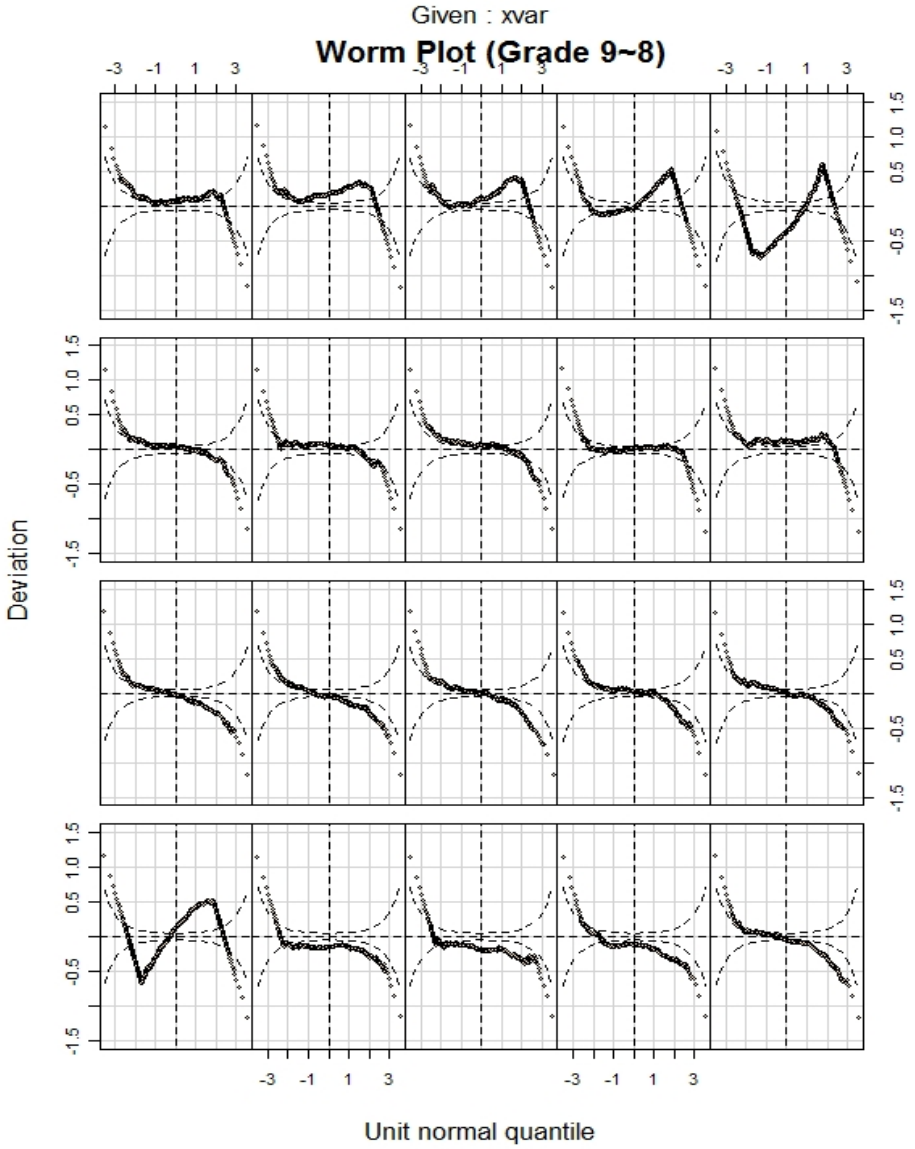
Figure 4.3

(a) Diagnostic plot for goodness-of-fit of the Quantile Regression of grade 8 reading scores on grade 7 reading scores



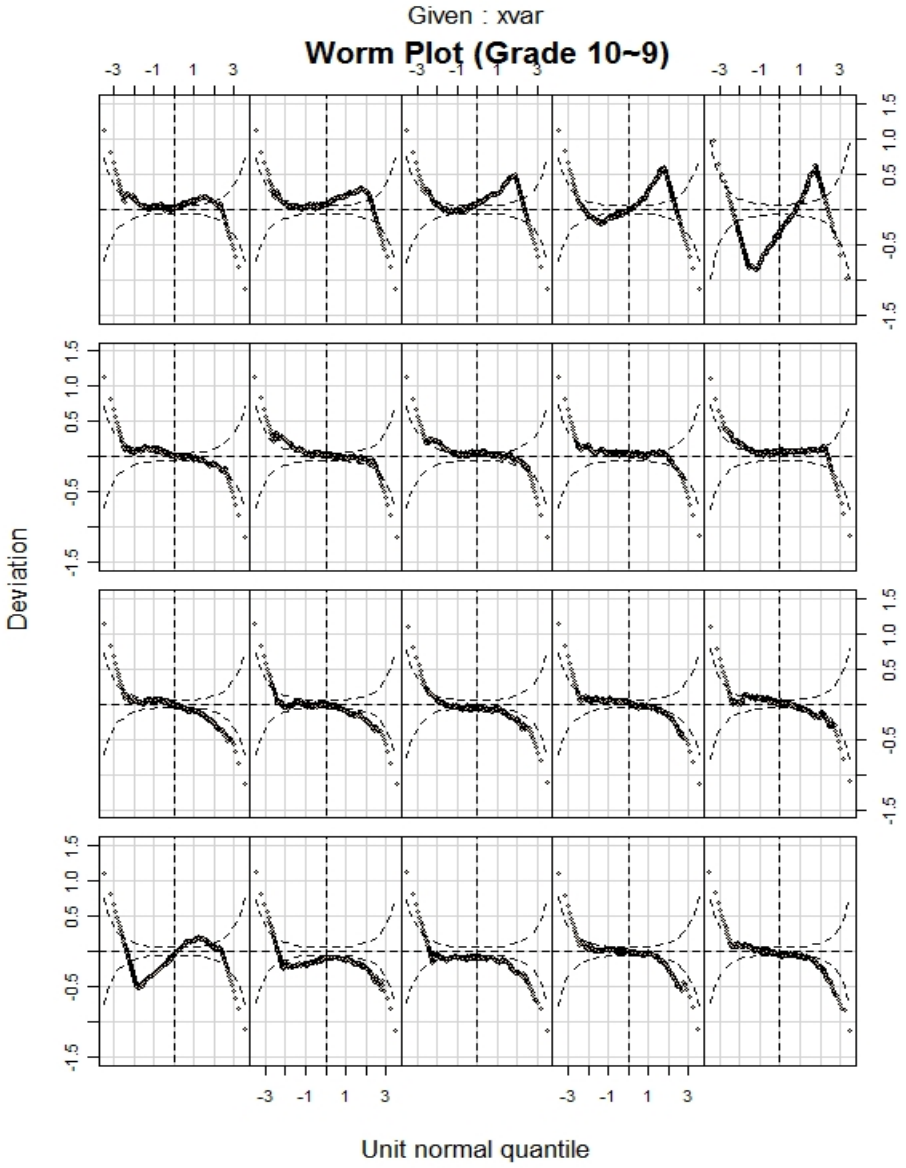

| Worm Plot X Range for figure (a) (sample sizes in brackets) | | | | |
|---|---|---|---|---|
| 678-686 [2691] | 687-696 [2753] | 697-707 [2502] | 708-724 [2703] | 725-930 [2803] |
| 640-646 [2526] | 647-654 [2870] | 655-661 [2579] | 662-669 [2788] | 670-677 [2598] |
| 597-607 [2725] | 608-616 [2541] | 617-624 [2520] | 625-632 [2686] | 633-639 [2506] |
| 320-519 [2422] | 520-549 [2299] | 550-569 [2577] | 570-584 [2622] | 585-596 [2580] |

(b) Diagnostic plot for goodness-of-fit of the Quantile Regression of grade 9 reading scores on grade 8 reading scores




| Worm Plot X Range for figure (b) (sample sizes in brackets) | | | | |
|---|---|---|---|---|
| 698-705 [2316] | 706-715 [2525] | 716-727 [2415] | 728-745 [2397] | 746-889 [2586] |
| 661-667 [2323] | 668-674 [2299] | 675-681 [2339] | 682-689 [2580] | 690-697 [2532] |
| 618-627 [2567] | 628-636 [2541] | 637-645 [2655] | 646-653 [2489] | 654-660 [2359] |
| 344-549 [2586] | 550-576 [2370] | 577-593 [2364] | 594-606 [2347] | 607-617 [2420] |

(c) Diagnostic plot for goodness-of-fit of the Quantile Regression of grade 10 reading scores on grade 9 reading scores




| Worm Plot X Range for figure c (sample sizes in brackets) | | | | |
|---|---|---|---|---|
| 703-709 [2205] | 710-717 [2210] | 718-727 [2122] | 728-744 [2226] | 745-824 [1251] |
| 673-678 [2242] | 679-684 [2282] | 685-690 [2267] | 691-696 [2221] | 697-702 [2144] |
| 641-648 [2278] | 649-655 [2284] | 656-661 [2048] | 662-667 [2266] | 668-672 [1928] |
| 362-584 [2106] | 585-607 [2192] | 608-621 [2161] | 622-631 [2006] | 632-640 [2258] |

Figure 4.4 Density plots of the 2000 naïve and SIMEX estimates for different quantile models and covariate variances for $\epsilon \sim N(0,1)$ : the black vertical line in each plot represents the true value of the corresponding parameter.

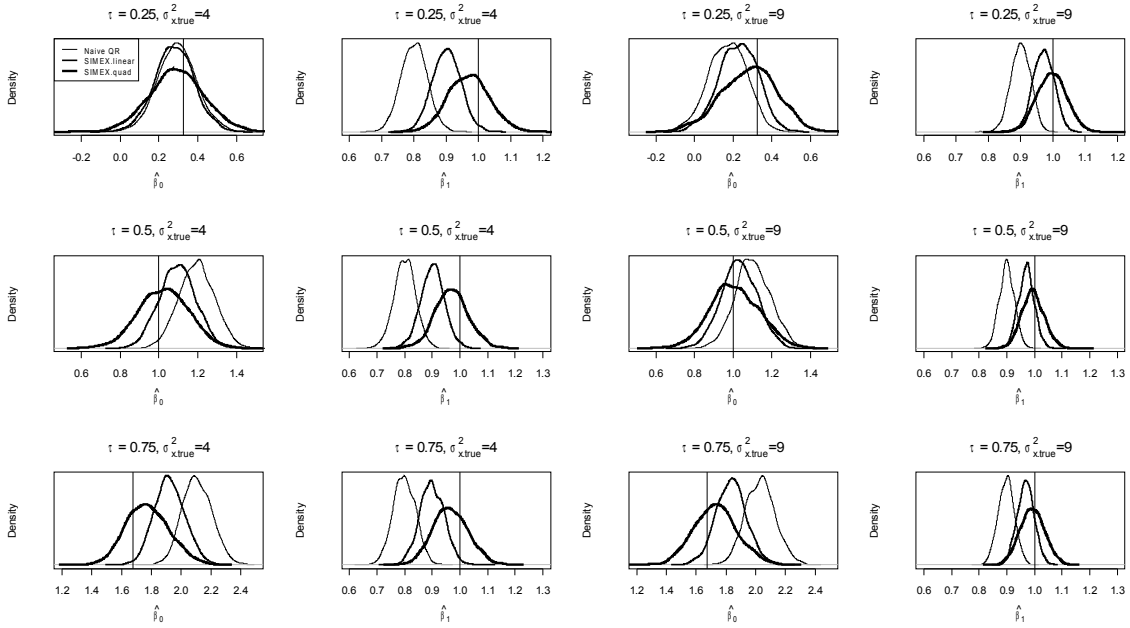


Figure 4.5 Comparison of the SIMEX estimates to the naïve estimates of lag-1 QR lines

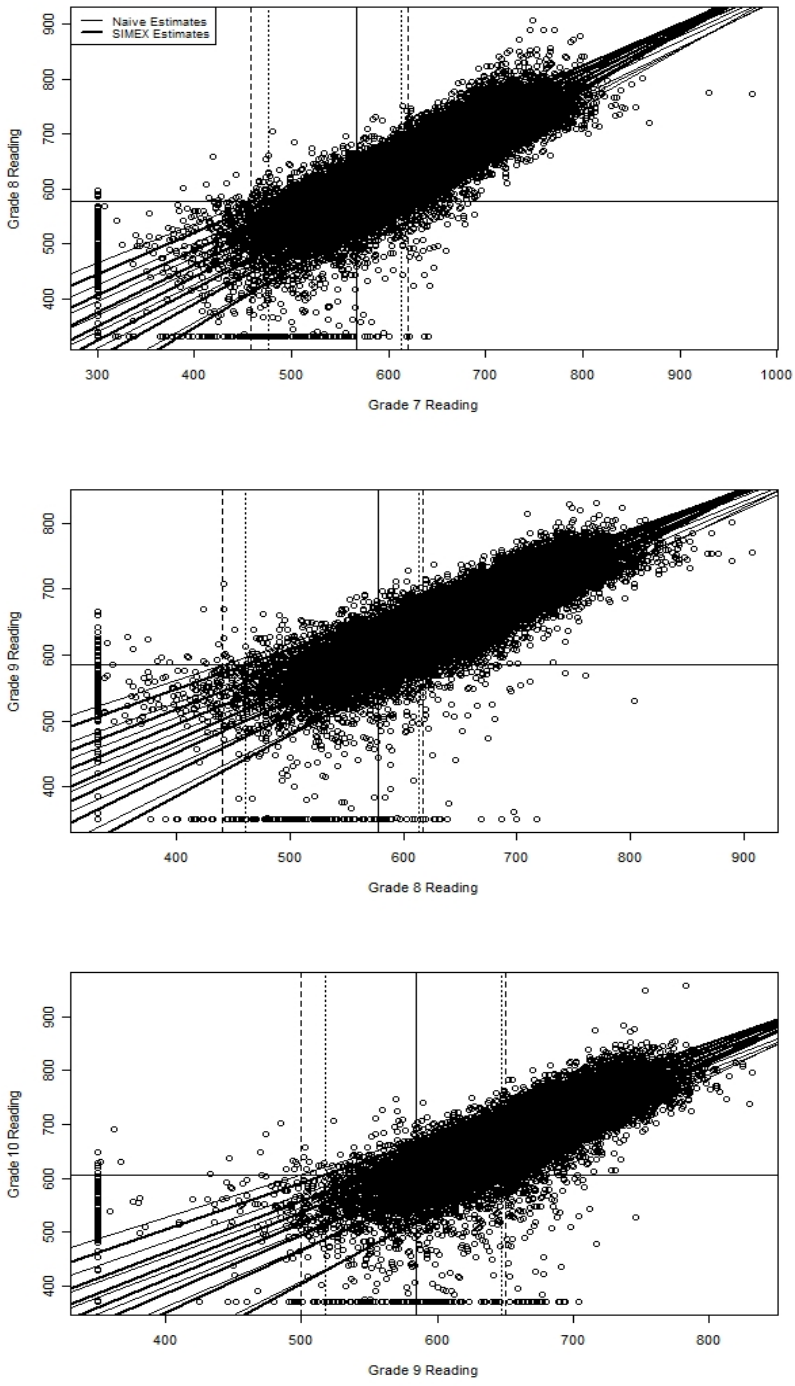


Figure 4.6 Conditional growth charts based on fitted values of the simple QR and SIMEX models. Each plot has the starting point at the 25th unconditional percentiles of grade 7 scores in 2003, and follows the growth paths at the 25th, 50th, and 75th conditional percentiles consistently for three years.

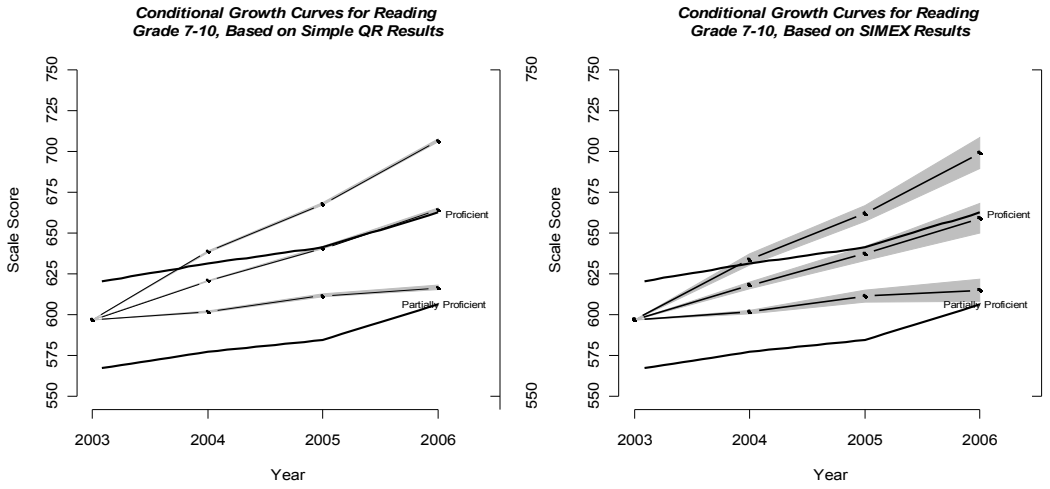



Figure 4.7 Density of lag-1 growth percentiles

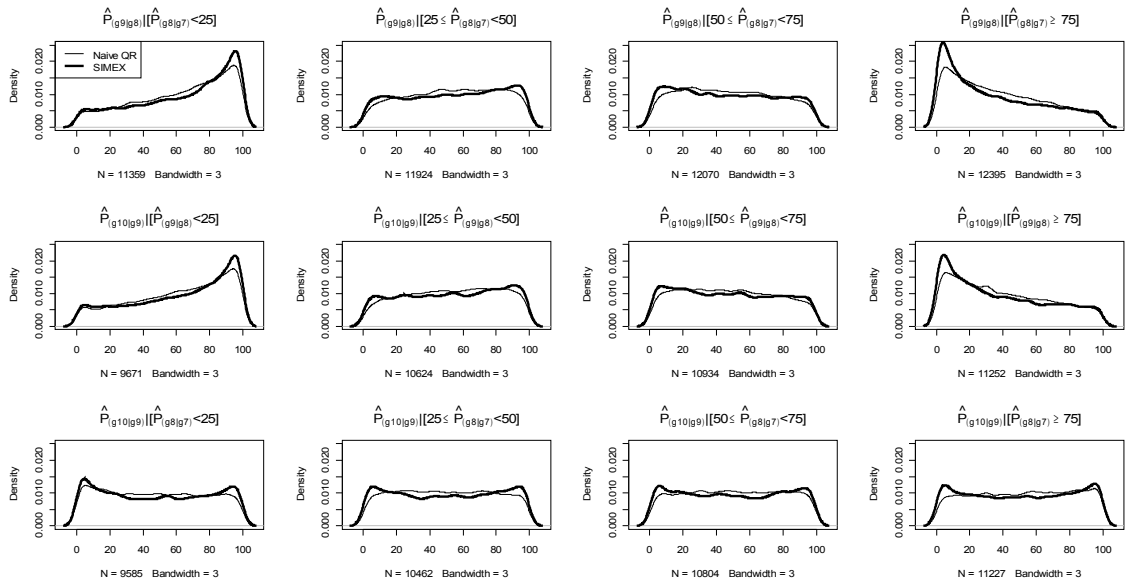


Figure 4.8 Density of growth percentiles conditioning on grade 7 scores

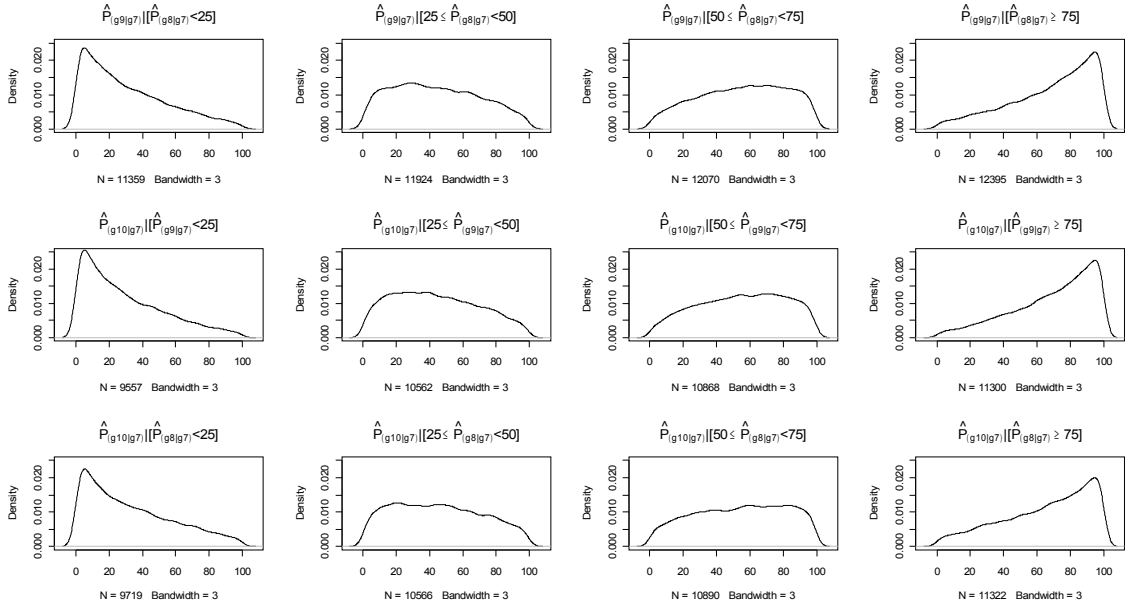



Figure 4.9 The two groups of students presented in the two plots fulfill the standards $\left|\hat{P}_{g9|g8}-\hat{P}_{g8|g7}\right|\geq 50$ and $\left|\hat{P}_{g10|g9}-\hat{P}_{g9|g8}\right|\geq 50$, respectively. The distribution of difference between $\hat{P}_{g9|g7}$ and $\hat{P}_{g8|g7}$ for the first group is drawn in the left plot, and the distribution of difference between $\hat{P}_{g10|g7}$ and $\hat{P}_{g9|g7}$ for the second group is drawn in the right plot.

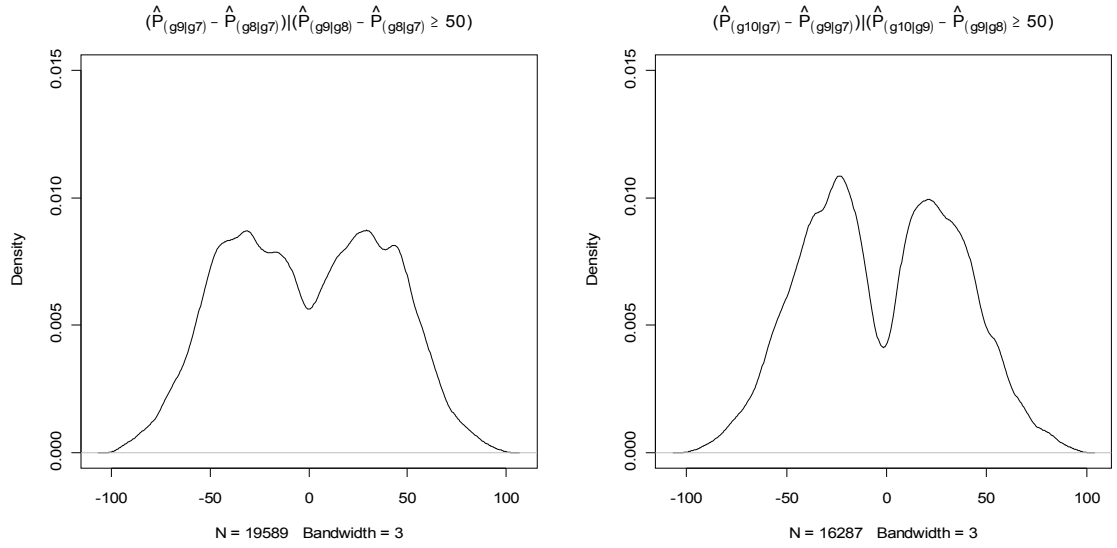


Figure 4.10

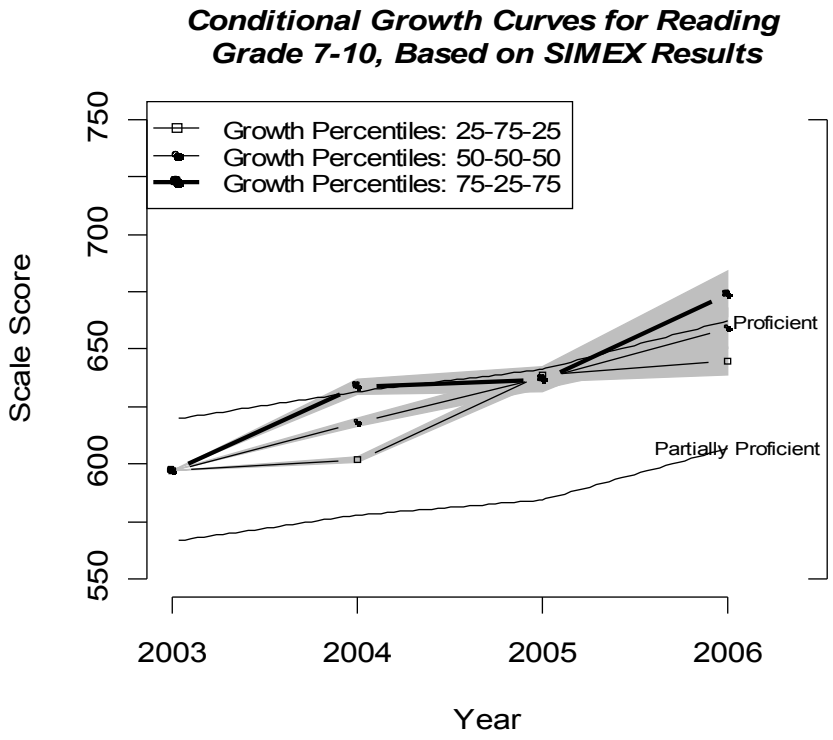

Conditional Growth Curves for Reading
Grade 7-10, Based on SIMEX Results
Growth Percentiles: 25-75-25
Growth Percentiles: 50-50-50
Growth Percentiles: 75-25-75
Scale Score
550
600
650
700
750
Proficient
Partially Proficient
2003
2004
2005
2006
Year

Figure 4.11 Unconditional and conditional growth chart to screen the academic growth of student 251337

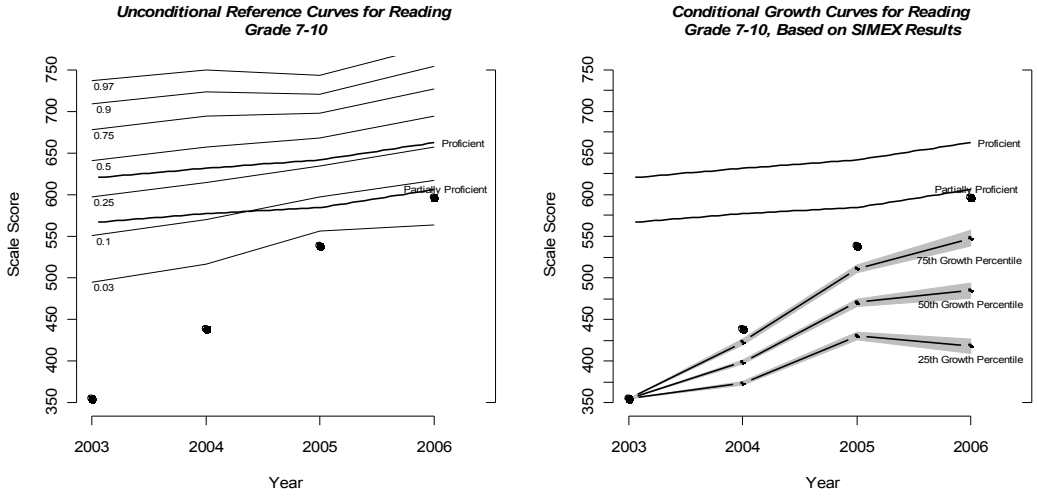



Figure 4.12 Unconditional and conditional growth chart to screen the academic growth of student 561315

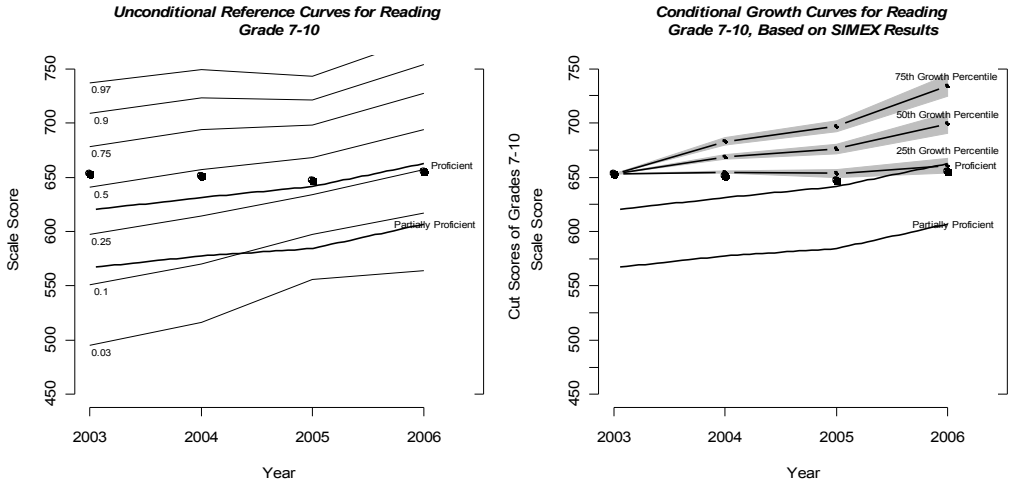


Figure 4.13 Boxplots of students' growth percentiles for a sample of schools

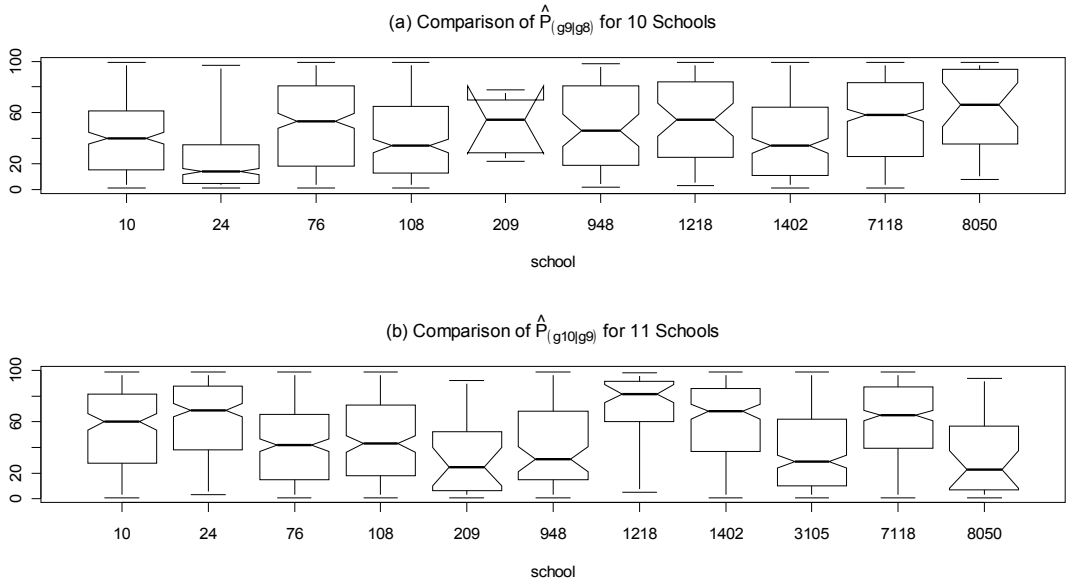



Figure 4.14 Scatter plot of the 9th grade and the 10th grade scores of the students in schools 3105 and 10. The seven black lines in the plot are the SIMEX quantile regression lines estimated with the whole cohort's data—exactly the same as the blue lines in the third plot of figure 4.5

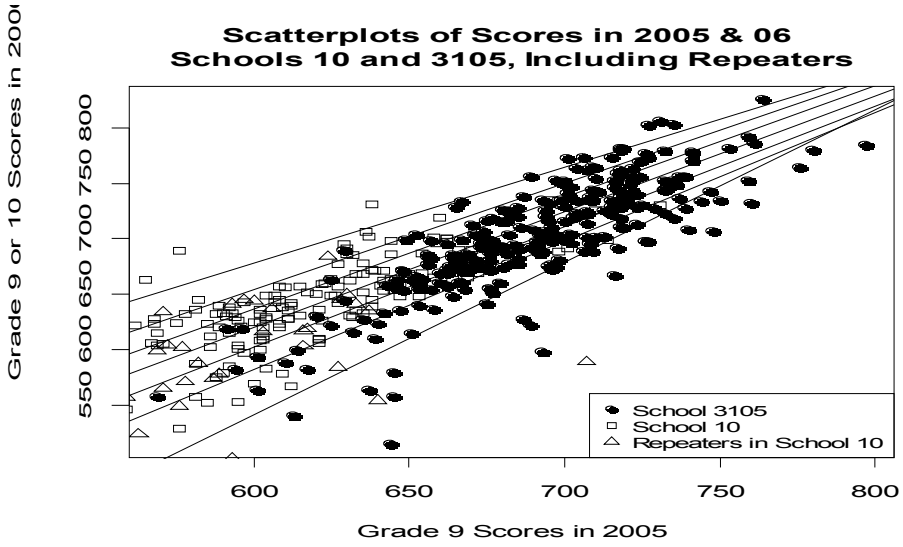


Figure 4.15 Distribution of lag-1 and lag-2 growth percentiles of the students in schools 3105 and 10.