U.S. Department of Education
January 2014

# Testing the importance of individual growth curves in predicting performance on a high-stakes reading comprehension test in Florida

Yaacov Petscher
Sarah Kershaw
Sharon Koon
Barbara R. Foorman

Florida Center for Reading Research at the Florida State University

## Key findings

To what extent does individual student change (growth) over the academic year statistically explain why students differ in end-of-year performance after accounting for performance on interim assessments? The four growth estimates examined in this report (simple difference, average difference, ordinary least squares, and empirical Bayes) all contributed significantly to predicting performance on the end-of-year criterion-referenced reading test when performance on the initial (fall) interim assessment was used as a covariate. The simple difference growth estimate was the best predictor when controlling for mid-year (winter) status, and all but the simple difference estimate contributed significantly when controlling for final (spring) status. Quantile regression suggested that the relations between growth and the outcome were conditional on the outcome, implying that traditional linear regression analyses could mask the predictive relations.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

REL SOUTHEAST
Regional Educational Laboratory
At Florida State University

REL 2014–006

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

January 2014

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-12-C-0011 by Regional Educational Laboratory Southeast administered by Florida Center for Reading Research, Florida State University. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Petscher, Y., Kershaw, S., Koon, S., & Foorman, B. R. (2014). *Testing the importance of individual growth curves in predicting performance on a high-stakes reading comprehension test in Florida* (REL 2014–006). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs.

This report is available on the Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

# Summary

Districts and schools use progress monitoring to assess student progress, to identify students who fail to respond to intervention, and to further adapt instruction to student needs. Researchers and practitioners often use progress monitoring data to estimate student achievement growth (slope) and evaluate changes in performance over time for individual students and groups of students.

The literature reports mixed findings on whether measuring individual student change over time on an interim progress monitoring assessment adds value to understanding student differences in future performance on an assessment. Specifically, to what extent does change over the academic year statistically explain why students differ in end-of-year performance after accounting for performance at the fall, winter, or spring assessment period (status variable). Some studies suggest that individual growth during the year does statistically predict variable differences in future performance on an assessment (Kim, Petscher, Schatschneider, & Foorman, 2010). Others find no contribution beyond that predicted by performance on an interim assessment (Schatschneider, Wagner, & Crawford, 2008; Yeo, Fearrington, & Christ, 2012).

Monitoring student progress is central to accountability systems in general and is useful for measuring how well students respond to instruction or intervention. Progress monitoring entails tracking individual growth across the academic year. Thus, it is important to understand why individual students differ on an outcome beyond what can be known by accounting for performance on a status assessment.

This study examines the relations among descriptive measures of growth (simple difference and average difference) and inferential measures (ordinary least squares and empirical Bayes) for students in grades 3–10 and considers how well such measures statistically explain differences in end-of-year reading comprehension after controlling for student performance on a mid-year status assessment. The study also looks at how the results change when controlling for initial (fall) and final (spring) status and when the relations among individual growth curves, status, and end-of-year reading comprehension performance depend on end-of-year reading comprehension performance.

Using archival data for 2009/10, the study analyzes a stratified random sample of 800,000 Florida students in grades 3–10: their fall, winter, and spring reading comprehension scores on the Florida Assessments for Instruction in Reading (FAIR) and their reading comprehension scores on the 2010 end-of-year state accountability assessment, the Florida Comprehensive Assessment Test (FCAT). Student differences in reading comprehension performance were explained by the four growth estimates (measured by the coefficient of determination, $R^2$) and differed by status variable used (performance on the fall, winter, or spring FAIR reading comprehension screen).

All four growth estimates significantly contributed to the prediction of FCAT performance when controlling for fall status, as did all but the simple difference estimate when controlling for spring status. But only the simple difference score was a good predictor when controlling for winter status. Quantile regression suggested that the relations between growth estimates and FCAT performance were conditional on the FCAT, implying that traditional linear regression analyses could mask the predictive relations.

# Contents

## Figures

## Tables

## Why this study?

When an interim progress monitoring assessment is administered multiple times over an academic year, data from a sample of individual students are typically used to evaluate changes in performance over time. These data can be used to estimate what is called an individual growth curve, which is the estimated amount of change a student is expected to make over time. An individual growth curve can be used to test whether a student's performance is improving (growing), as well as whether the student is growing faster or slower than other students in the sample. It can also be used to identify students whose score at the first assessment period and whose growth trends differ from the sample mean. For example, it might be possible to identify a set of students who began the school year at a similarly low level of reading performance but who display different individual growth curves, with some improving faster than others.

In addition to revealing changes in an individual student's progress and comparing changes across students, individual growth curves can be used to identify students who fail to demonstrate adequate progress relative to an aggregate mean (for example, the classroom or school mean) or to a benchmark assessment standard. Widely used curriculum-based measurement programs for reading, such as the Dynamic Indicators of Basic Early Literacy Skills assessments (DIBELS; Good & Kaminski, 2002), include criteria for fluency rates that are aligned with a set of predictive validity analyses. The cutpoints for risk between any two assessment points (for example, 77 correct words per minute in the fall of grade 3 and 92 correct words per minute in the winter) can be used to calculate the between-assessment change needed to achieve the appropriate benchmark (92–77 = an increase of 15 correct words per minute). Interim progress monitoring/benchmarking assessments administered between the two points could then be used to determine whether a student is on track to meet the later benchmark.

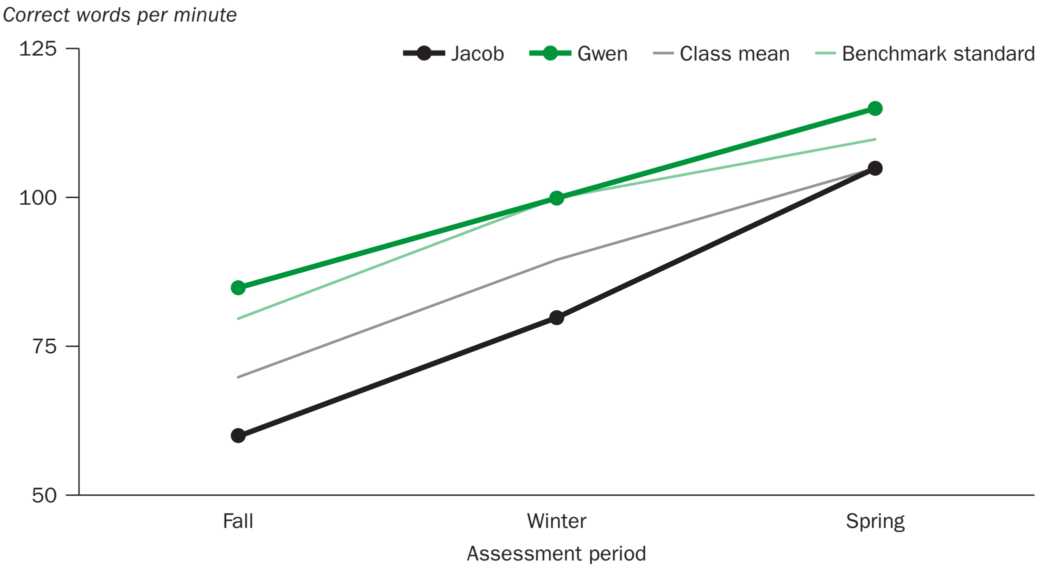
***Individual growth curves can be used to identify students who fail to demonstrate adequate progress relative to an aggregate mean or to a benchmark assessment standard***

The four types of score evaluations—individual student growth, student-to-student comparison, student-to-sample comparison, and student-to-benchmark comparison—can be depicted graphically. Figure 1 documents the progress of two hypothetical students, Jacob and Gwen, on a measure of oral reading fluency (correct words per minute) at three points during the year relative to the classroom mean and the benchmark goals at each point. These graphs provide a wealth of information about student performance and growth. First, they show that Jacob made greater gains from winter to spring (gain of 25 correct words per minute) than from fall to winter (gain of 20 correct words), while Gwen's progress was stable (gain of 15 correct words per minute in both periods). Second, comparing Jacob's progress with Gwen's shows that Gwen began the year with a higher score but that Jacob improved faster, narrowing the initial gap by the spring assessment. Third, Gwen consistently performed above the classroom mean (the sample), while Jacob performed below it until the spring assessment. Finally, Gwen's performance was at or above the benchmark standard during the period, while Jacob's was consistently below it.

### Using growth estimates from progress monitoring to identify students who are not responding to instruction

This example shows how growth can be used as an isolated measure of responsiveness to intervention. Recently, studies within the response to intervention framework have tested whether individual student growth on progress monitoring assessments by a particular

**Figure 1. Sample individual progress plot demonstrating the potential for four types of score evaluation**



**Source:** Authors' illustration.

population statistically explained sample differences in selected outcome performance beyond what could be predicted by performance at a single point in time (initial, middle, or final performance; Kim et al., 2010; Schatschneider et al., 2008; Yeo et al., 2012; Zumeta, Compton, & Fuchs, 2012).

Yeo et al. (2012) used latent parallel process growth models to test how well intercepts and individual growth estimates from curriculum-based measures of reading fluency and maze (a multiple-choice cloze task[1]) explained sample differences on the Tennessee Comprehensive Assessment Program. They found that individual growth curves from the structural portion of the model did not statistically explain differences beyond the first assessment. Zumeta et al. (2012) used nonlinear individual growth curve analysis and multiple regression to evaluate the correlations between growth in word identification fluency and several outcome measures, including the Woodcock Reading Mastery Test-Revised (Woodcock, 1998) and both the sight word and decoding portions of the Test of Word Reading Efficiency (Torgesen, Wagner, & Rashotte, 1999). Weak to moderate correlations between the measures of growth and the selected outcomes were observed.

Because Yeo et al. (2012) and Zumeta et al. (2012) used different samples and different measures of reading skills, their conflicting findings cannot be compared. However, two other studies using the same measures with approximately the same sample also yielded conflicting findings (Kim et al., 2010; Schatschneider et al., 2008).

Kim et al. (2010) used a combination of growth curve modeling and dominance analysis (Azen, 2013) to test whether growth in oral reading fluency as measured by the DIBELS assessment (Good & Kaminski, 2002) explained variations in student scores on the Stanford Achievement Test, 10th edition (SAT-10; Harcourt Brace, 2003) for a cohort of students followed from grade 1 through grade 3. The study reported that about 15 percent of the growth occurred during grade 1, 15 percent during grade 2, and 6 percent during grade 3. Individual student growth in oral reading fluency during grade 2 explained 7 percent of the variance in SAT-10 scores at the end of the school year but did not explain grade 3 sample differences in SAT-10 performance when other variables were controlled for. Finally, growth in oral reading fluency during grade 3 accounted for approximately 6 percent of sample differences in grade 3 SAT-10 performance.

In a similar study using DIBELS as a measure of growth in oral reading fluency, Schatschneider et al. (2008) used a combination of linear analysis of individual growth curves and multiple regression to predict grade 1 performance on the SAT-10. The study found that growth in oral reading fluency did not explain variation in student performance on the SAT-10 after controlling for initial status. Although form effects are often an issue when oral reading fluency is used as a measure of growth (Ardoin & Christ, 2009; Cummings, Park, & Schaper, 2013; Francis et al., 2008; Hintze & Christ, 2004; Petscher, Cummings, Biancarosa, & Fien, 2013; Petscher & Kim, 2011), Kim et al. (2010) and Schatschneider et al. (2008) used approximately the same sample and oral reading fluency probes, so the sharp contrast in their conclusions needs to be explained.

***The choice of controlling for the first or the last assessment point affects the understanding of how individual growth curves can account for individual differences in an outcome***

## Limitations of prior research

All these studies show how growth on interim progress monitoring measures might expand the understanding of why students vary in their performance on selected outcome measures. Each study had specific limitations in study design and sampling, but three methodological differences across the studies are especially worth noting: the status variable used as a covariate in predicting the outcome, the type of growth estimate used to predict the outcome, and the achievement level of the sample.

- *The status variable used.* The predictive studies by Kim et al. (2010), Yeo et al. (2012), and Zumeta et al. (2012) used the first assessment point (student performance in the fall) as the status variable. Schatschneider et al. (2008) used the final assessment point (spring). Each study addressed the broad question of how well individual growth curves explain differences in selected outcomes beyond what can be explained by a status variable. Schatschneider et al. framed the research questions around how individual growth curves uniquely predicted outcomes beyond predictions based on end-of-year status, while the other three studies looked at using individual growth curves to explain differences in outcome performance beyond the contributions based on beginning-of-year status. The choice of the first or the last assessment point affects the understanding of how individual growth curves can account for individual differences in an outcome, controlling for fall or spring status. The use of different status covariates in these studies means that the results are not directly comparable.

  An ancillary consideration is that none of these predictive models used the mid-year assessment as a status variable. The mid-year has appeal both instructionally and practically. From an instructional perspective it marks the first time that a learning gain within the same school year can be evaluated. Knowing the unique contribution of gains from the fall to the mid-year for predicting outcomes at the end of the school year could enable teachers to modify instruction accordingly. From a practical perspective, using mid-year status makes more sense than using beginning-of-year status, when no growth has yet taken place, or end-of-year status, when teachers can no longer adapt instruction to individual differences in gains over the year.

- *Type of growth estimate used.* All the studies but Yeo et al. (2012) used individual growth curves estimated with ordinary least squares (OLS) regression to predict the selected outcomes; Yeo et al. used latent growth curves with a maximum likelihood estimator. None of the studies used an empirical Bayes slope (model-based estimate of individual growth curves), often considered a best practice for estimating individual growth (Singer & Willett, 2003) because it combines OLS estimates with the grand mean (population mean). The empirical Bayes slope shrinks an OLS estimate toward the grand mean by a factor proportional to its individual unreliability; thus, individual OLS growth values at the tails of the slope distribution get pulled much closer to the grand mean because they are typically less reliable—and therefore less likely to reflect the true slope. Although the empirical Bayes estimate yields a more reliable slope, it often comes at a cost: biased estimates (Singer & Willett, 2003). (The next section on estimation frameworks discusses the differences between OLS and empirical Bayes estimates.)
- *Achievement level of the sample.* A final methodological consideration concerns the nature of the sample. Yeo et al. (2012), noting that the lack of predictive validity of individual growth curves might have been related to their sample not consisting predominantly of students at high risk of low performance, conjectured that individual growth curves might be more predictive for these students. Zumeta et al. (2012) found that growth on the word identification fluency task was more strongly associated with outcomes for the low-performing subsample than for the average and high-performing subsamples. Because frequent progress monitoring often focuses on students with the highest risk of low performance, the types of regression models typically used to evaluate differential predictive validity might fail to adequately capture how well individual growth curves explain differences in FCAT performance for individuals at the low end of the achievement distribution for the dependent variable.

***Because frequent progress monitoring often focuses on students with the highest risk of low performance, regression models typically used to evaluate differential predictive validity might fail to capture how well individual growth curves explain differences in performance for students at the low end of the achievement distribution for the dependent variable***

Despite study differences in status variables, estimators, and samples, the broad research question was the same: To what extent are individual growth curves related to the outcomes? Another way to frame this query in order to test the conclusion by Yeo et al. (2012) that individual growth curves might be more predictive for students at high risk of low performance: Does the relation between the individual growth curve and the outcome for students at the lower end of the distribution on the outcome differ from the relation for students at the middle and upper ends of the distribution?

Traditional fixed- and random-effects regression models are often underpowered for addressing this type of question, as these models require splitting up the sample to test whether one point of the distribution differs from another. This approach produces a restricted range of values for the dependent variable and reduces the sample to observations above or below the selected cutpoint in the distribution. A conditional median model such as quantile regression (Koenker, 2005) can overcome these limitations. Quantile regression includes no assumptions about the sample distribution (for example, normality) and uses the entire sample to compute coefficients at each quantile (Koenker, 2005; Petscher & Logan, in press; Petscher, Logan, & Zhou, 2013). Rather than asking the question, “To what extent do individual growth curves explain sample differences in outcome performance?” quantile regression asks, “To what extent do individual growth curves explain sample differences in outcome performance based on the outcome performance itself?” The section on estimation frameworks expands on this analytic tool’s benefits and limitations.

## Study questions

The growing reliance on interim progress monitoring assessments in both response to intervention and broader accountability systems (such as those mandated by the No Child Left Behind Act of 2001) elevates the importance of studying how well individual growth curves predict performance on a selected outcome beyond what can be accounted for by the status variable alone. The literature has produced mixed findings, with some studies showing that growth in reading statistically explains such differences (Kim et al., 2010) but others finding it does not (Schatschneider et al., 2008; Yeo et al., 2012).

Differences in the type of growth estimates used in the study (OLS or maximum likelihood) and in the status variable used as a covariate (fall or spring assessment) have been proposed as the reason for the conflicting results. Thus, two immediate goals of this study were to evaluate the extent to which different approaches to estimating individual growth curves differentially predict an outcome beyond what is predicted using a single status variable (for example, results of the fall interim assessment), as well as the extent to which the statistical significance of the individual growth curve might vary when the status variable changes (for example, from fall to spring). In addition, the study sought to expand the research base by testing the extent to which individual growth curves predict performance beyond what is predicted by a mid-year (winter) status variable and by studying what the unique relations might look like in understudied populations (for example, students in secondary grades).

***Two immediate goals of this study were to evaluate the extent to which different approaches to estimating individual growth curves differentially predict an outcome beyond what is predicted using a single status variable and the extent to which the statistical significance of the individual growth curve might vary when the status variable changes***

K–3 students have been the population of interest in many response to intervention studies, with less focus on middle and secondary school students (Barth et al., 2012; Espin, Wallace, Lembke, Campbell, & Long, 2010; Pyle & Vaughn, 2012; Tichá, Espin, & Wayman, 2009). The National Center on Response to Intervention noted that most states use response to intervention as a prevention/intervention model, while some use it for identifying students with learning disabilities. Recent research has found secondary school students to be responsive to targeted, intensive literacy interventions (Calhoon, 2005; Calhoon & Petscher, 2013; Edmonds et al., 2009; Vaughn et al., 2010, 2011, 2012). That makes it important to characterize the extent to which growth in measures of reading comprehension and related skills explain variation in scores on outcomes for both primary and secondary school students. That is especially relevant considering studies such as Silberglitt and Hintze (2007) that find differences in expected growth rates as grade level rises, with average growth slowing from grade 2 to grade 6 on interim progress monitoring assessments (administered three times a year).

In a typical response to intervention framework, progress monitoring assessments are administered once or twice a week. Considering the practical obstacles of scheduling weekly or more frequent assessments, recent research has examined the viability of shifting to fewer assessments (Jenkins, Graff, & Miglioretti, 2009). Most recently, Ardoin, Christ, Morena, Cormier, and Klingbeil (2013) used simulation to study the validity and reliability of growth estimates dependent on the schedule and duration of progress monitoring assessments, as well as the dataset quality. They found growth estimates from monthly assessments over a 17-week period to be sufficiently valid and reliable for low-stakes decisions when the dataset was of very high quality.

Florida administers interim progress monitoring assessments three times a year, as well as the end-of-year FCAT. This provides an opportunity to study how well individual growth curve estimates explain individual differences in the end-of-year reading comprehension test with a large, diverse sample. Shaped by the limitations of previous research and gaps in developmental research on progress monitoring, the following research questions consider student growth for grades 3–10 in 2009/10:

- What are the relations among descriptive measures of student change (simple difference and average difference) and inferential measures of individual growth curves (OLS and empirical Bayes)?
- Controlling for students' mid-year status, how well do the measures of student change and individual growth curves explain differences in end-of-year reading comprehension performance?
- Controlling for students' initial or final status, how well do the measures of student change and individual growth curves explain differences in end-of-year reading comprehension performance?
- Controlling for the type of estimator, how well do the measures of student change and individual growth curves explain differences in end-of-year reading comprehension performance conditional on end-of-year reading comprehension performance?

Before the findings are presented, the next three sections outline the theoretical framework for the analyses and discuss the sample, the measures used in the study, and the methods used to analyze the data.

***With interim progress monitoring assessments administered three times a year, Florida provides an opportunity to study how well individual growth curve estimates explain individual differences in the end-of-year reading comprehension test with a large, diverse sample***

## Specifying the estimation framework

The four types of estimates for measuring student change range from descriptive, computationally simple measures of change (simple difference and average difference scores) to complex, inferential measures (OLS and empirical Bayes).

### Descriptive measures of change

*Simple difference score.* The simple difference score, one of the earliest methods of analyzing data over multiple assessments (Lord, 1956), is depicted by:

$$\Delta X_i = X_{2i} - X_{1i},$$

which captures the change, $\Delta X_i$, for student $i$ from time $X_{1i}$ to time $X_{2i}$. The ease of calculating the simple difference score makes it appealing, but the measure fell out of favor when many researchers—for example, Cronbach and Furby (1970)—incorrectly considered it unreliable and thus invalid. Rogosa, Brandt, and Zimowski (1982) refuted those arguments, showing that the supposed unreliability of the change score was often a function of a lack of individual differences in change over time, as measured by

$$\rho_{(\Delta X)} = \frac{\sigma^2_{\Delta T}}{\sigma^2_{\Delta T} + \sigma^2_{\Delta e}},$$

where $\sigma^2_{\Delta T}$ is the true score variance and $\sigma^2_{\Delta e}$ is the error score variance. If there is no variance in $\sigma^2_{\Delta T}$, the reliability of the change score is 0. This does not mean that meaningful change does not occur, but rather that it is impossible to detect *reliable* change when there is no variance in the estimated change score because all students change by approximately the same amount (positive or negative). As the rate of change varies across students, reliability will increase when the error score variance is low.

*Average difference scores.* When a student's performance is measured on more than two occasions, the simple difference score cannot capture change over the academic year because growth rates might vary across semesters (Ardoin & Christ, 2008). Benchmark/interim assessments are typically administered to students three times a year: at the beginning of the school year in the fall, at the mid-point in the winter, and at the end in the spring. The average difference score would represent the average amount of change observed across the two change scores (the change from fall to winter plus the change from winter to spring divided by two). By drawing on more information than the simple difference, the average difference score may better reflect student progress across the year.

***Individual growth curve analyses are used for estimating change when there are more than two waves of assessments***

### Inferential measures of change

An alternative method for estimating change when there are more than two waves of assessments is student growth curve analysis. This framework uses inferential models to characterize change.

*Ordinary least squares.* OLS models provide a simple structure for the data, allowing individual trajectories to be estimated for each student based on the student's assessment data. A structural form of the OLS regression model for growth is

$$Y_{ti} = \beta_{0i} + \beta_{1i}(TIME)_{ti} + e_{ti}\,,$$

where $Y_{ti}$ is the predicted score for student $i$ at time $t$ and is a function of the intercept ($\beta_{0i}$), slope ($\beta_{1i}$), and residual ($e_{ti}$). OLS regression is useful for characterizing change because it is readily available in most statistical software packages, which routinely provide summary statistics ($R^2$ and residual variance) that can be used for evaluating the goodness of fit for each student.

Despite the ease of testing, this model has several limitations. First, OLS regression assumes that the errors between the observations are not correlated and that the residuals have constant variance. As Singer and Willett (2003) note, these assumptions are often untenable with longitudinal designs because errors over repeated observations for an individual are frequently correlated. A second limitation is the need for a nearly complete dataset to estimate an empirical slope. Missing data pose a particular problem with OLS because growth curves cannot be estimated for individuals with fewer than two data points. Despite these concerns, running OLS regressions is often worthwhile, because it is practical and produces unbiased estimates of the intercept and slope (Singer & Willett, 2003). Although the estimated individual growth curves for some individuals will appear to depart significantly from the rest of the distribution, they remain unbiased as long as the model assumptions are met.

*Empirical Bayes.* Using a multilevel approach can avoid the OLS problems of poor reliability and precision. Multilevel modeling (or random-effects mixed modeling) is a flexible framework that allows fitting individual growth curves to multiple waves of data over time. One of the most basic individual growth curve specifications is the linear growth model:

$$\text{Level 1: } Y_{ti} = \pi_{0i} + \pi_{1i}(TIME)_{ti} + e_{ti}$$

$$\begin{aligned} \text{Level 2: } \pi_{0i} &= \beta_{00} + r_{0i}\,. \\ \pi_{1i} &= \beta_{00} + r_{1i} \end{aligned}$$

In level 1, $Y_{ti}$ is score $Y$ for student $i$ at time $t$. The score is predicted by intercept $\pi_{0i}$ for each student (status), slope coefficient $\pi_{1i}$ for each student (growth), and an occasion-level residual $e_{ti}$, which is the difference between a student's score and predicted score at time $t$. In level 2, $\beta_{00}$ is the overall mean intercept, and $\beta_{10}$ is the overall mean slope; $r_{0i}$ and $r_{1i}$ are the deviations of the student's intercept and slope from the overall mean, where $r_{0i} \sim (0,\tau_{00})$ and $r_{1i} \sim (0,\tau_{11})$.[2]

***The empirical Bayes approach does not require complete data, and its estimate yields a more reliable estimate of growth by bringing the ordinary least squares slope closer to the mean***

For estimating individual growth curves the multilevel model has two main advantages over OLS: it does not require complete data, and it estimates an individual slope more precisely.

Multilevel models use an empirical Bayes procedure to estimate individual growth curves, creating a composite slope from the average slope for the sample and an individual's predicted slope (Raudenbush & Bryk, 2002). As a result, the empirical Bayes curve will generally fall somewhere between an individual's OLS growth curve and the sample average growth curve. As a weighted average based on OLS and the average trajectory for a sample, an empirical Bayes trajectory will vary across individuals within a sample, but its variance will likely be smaller than that of the OLS because the empirical Bayes includes the average trajectory as part of its estimate (Singer & Willett, 2003). The farther an estimated slope is from the average trajectory, the lower the reliability of that slope. Thus, the empirical Bayes estimate yields a more reliable estimate of growth by bringing the OLS slope closer to the mean. Figure 2 illustrates this advantage by graphing the performance of a grade 4 student in Florida whose reading comprehension was assessed three times a year using the FAIR. This graph shows how the empirical Bayes brings the OLS growth curve closer to the sample mean trajectory.

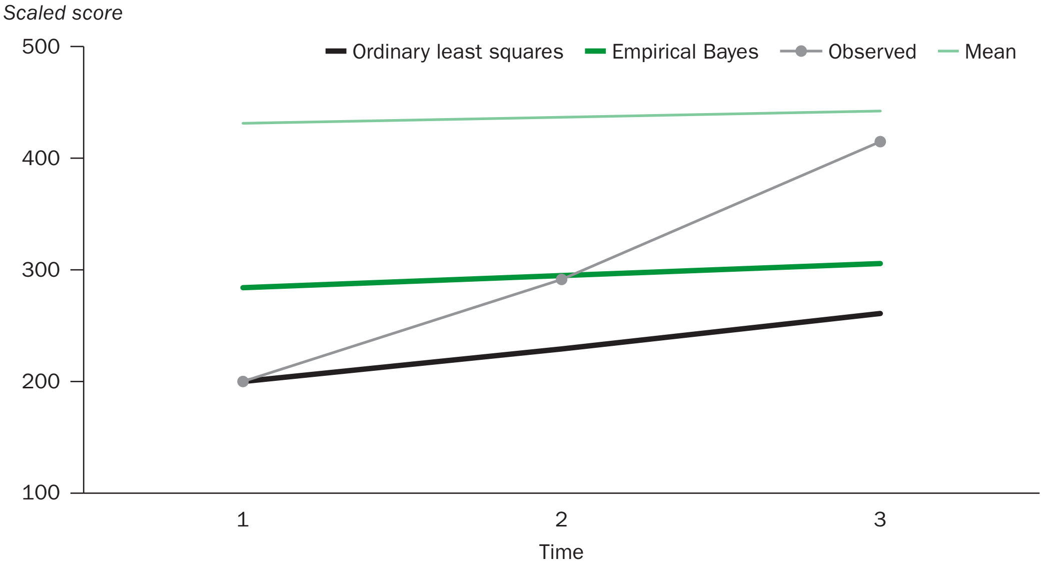
A second advantage of the multilevel model is that the empirical Bayes approach often provides a more precise estimate of an individual slope. The multilevel model estimate of change assumes that the level-1 residual variance is the same for everyone, whereas the OLS model estimates a variance for each student. Reducing the number of variances estimated confers greater reliability to the individual growth curves.

The empirical Bayes individual growth curve also has limitations, the most notable being bias because the individual's growth curve estimate is weighted by the sample mean. When using such estimates from multilevel models in a secondary analysis, researchers must decide which attribute is more important to the analysis: a more reliable slope (empirical Bayes) or a more unbiased estimate of it (OLS).

## Selecting the appropriate growth measure

Each of the four growth measures outlined here has conceptual or statistical properties that can influence a decision on how to estimate growth. One measure's statistical merits might

**Figure 2. Estimates of individual student growth using observed data, the sample mean, the ordinary least squares trajectory, and the empirical Bayes trajectory**



**Source:** Authors' illustration.

need to be weighed against its computational demands. For example, local or state education agency personnel might prefer the descriptive measures of growth (simple and average difference scores), because they allow teachers to estimate student change across the year by calculating simple scores. But a researcher interested in using individual growth curves for analysis might be drawn to an inferential measure that maximizes reliability (empirical Bayes) or provides an unbiased estimate of change (OLS). Researchers will typically opt for the more reliable estimate of growth, as minimizing error in measured variables is always desirable. Further, a score's validity depends on the degree of error in the measure. Thus, differences in the predictive validity of the individual growth curves for each growth measure are important to consider.

### Using quantile regression for analyzing the relation between individual growth curves and the selected outcome, conditional on the outcome

The fourth research question is concerned with how well individual growth curves and a status measure explain differences in the selected outcome, conditional on the outcome. Conventional regression models, such as multiple regression, are conditional means models—when the relation between two variables (*X* and *Y*) is being estimated, the resulting coefficients for the independent variables are interpreted as mean effects. As a conditional median model, quantile regression models examine the relation between *X* and *Y* conditional on the score of *Y* (the *p*th quantile). The conceptual underpinnings of a quantile are similar to those of a percentile: the *p*th quantile refers to the place in a distribution of scores where the proportion of the population below that value is *p*. Thus, the .25 quantile is similar to the 25th percentile; both refer to the point below which 25 percent of the sample distribution lies. (For more detail on quantile regression, see Koenker, 2005, and Petscher et al., 2013.)

More broadly, multiple regression answers the question: "What is the relation between *X* and *Y*?" Quantile regression answers the question: "What is the relation between *X* and *Y* for those who vary in their score on Y?" This approach illuminates relations throughout the distribution of the outcome variable that conditional means models cannot detect.

In linear regression, an estimated value of Y is calculated based on the corresponding X value:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i\,, \tag{1}$$

where $\beta_0$ is the intercept, $\beta_1$ is the slope, $X_i$ is the score for individual $i$ on independent variable X, and $\varepsilon_i$ is the error term, which is assumed to be identically, independently, and normally distributed, with a mean of 0 and variance of $\sigma^2$. The intercept and slope portions of the model are estimated such that the values are relative to the mean of Y, given X. The relation of X to Y is estimated by minimizing the squared difference between the predicted value of Y and the observed value of Y (the sum of the squared error). The result of the prediction equation can be represented by a single line through a scatterplot of points.

***Quantile regression answers the question: "What is the relation between X and Y for those who vary in their score on Y?"***

Similarly, quantile regression can be used to estimate the relation of X to Y at a given quantile within the distribution of Y. This can be done by identifying the sample score in the distribution associated with the quantile ($\tau$) of interest and estimating the coefficients for the independent variables. However it may seem, quantile regression is not akin to dividing the sample into multiple subgroups based on percentiles or cutpoints of the dependent variable and then fitting a linear regression to each subgroup. Rather, the selection of a given quantile occurs through minimization of the sum of absolute residuals, which is dependent on the given quantile. The minimization function is represented by:

$$\hat{Q}_Y(\tau) = \underset{\xi_\tau \in \mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{i\in\{i\,|\,Y_i \geq \xi_\tau\}} \tau\,|\,Y_i - \xi_\tau\,| + \sum_{i\in\{i\,|\,Y_i < \xi_\tau\}} (1-\tau)\,|\,Y_i - \xi_\tau\,| \right\}, \tag{2}$$

where $Y_i$ is the vector of independent variables, $\xi_\tau$ is the dependent variable, and $\tau$ is the quantile to be estimated. The relation between the independent and dependent variables can then be expressed as:

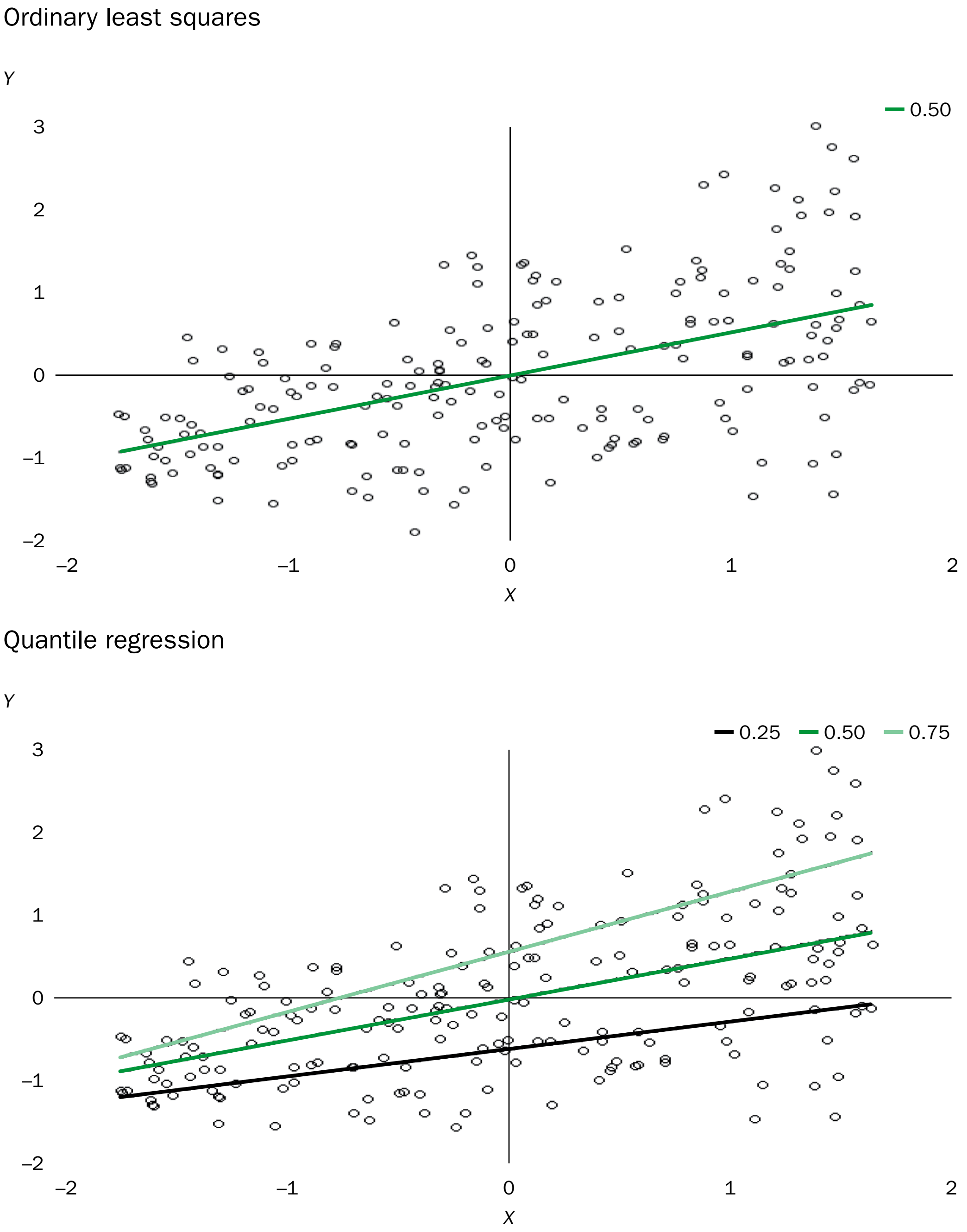
$$Y_i = \beta_0^{(p)} + \beta_1^{(p)} X_i + \varepsilon_i^{(p)}\,. \tag{3}$$

Equation 3 is structurally similar to equation 1, with the addition of a superscript $p$ (the $p$th quantile) above the intercept, slope, and error parameters. A distinguishing feature of equation 3 (quantile estimation) is that no assumption is made about the distributional form for $\varepsilon_i^{(p)}$ (for example, normal, poisson), while the corresponding $\varepsilon_i$ in equation 1 (typical linear regression) is assumed to be normally distributed. This critical difference allows quantile regression equations to be fitted to dependent or independent variables with non-normal distributions.

Just as with linear regression, equation 3 would also be represented by a single line through a scatterplot, but that line would be unique to the specified quantile rather than the average for the entire distribution. Although quantile regression can be viewed as an extension of median regression (a regression where $\tau$ = .5), the quantile approach can extend beyond the median through the asymmetric weighting system outlined in equation 2. Positive residuals would be given a weight of $\tau$; negative residuals, a weight of $1-\tau$.

One way to illustrate differences between OLS and quantile regression is through an example of both methods using a simulated dataset: 200 participants with scores on a dependent variable, *Y*, and one independent predictor, *X*. Variables *X* and *Y* have a mean of 0 and a standard deviation of 3. Figure 3, top panel, displays a scatterplot of *X* and *Y* with the OLS regression line, where the fit line represents the minimization of the sum of squared residuals. Figure 3, bottom panel, represents the results of a quantile regression on

**Figure 3. Comparison of scatterplots and fit lines from ordinary least squares and quantile regressions at the .25, .50, and .75 quantiles, using simulated data**



*n* = 200.

**Source:** Authors' illustration.

the same data, displaying three fit lines: one each for the .25, .50, and .75 quantiles. The fit line for the .50 quantile (the median) is very similar to that for the OLS regression (see figure 3, top panel). The line for the .25 quantile is not as steep as that for the .50 or .75 quantile, suggesting that scores at the 25th percentile of *Y* demonstrate a weaker relation between *X* and *Y* than do scores at or above the median.

## Constructing the sample and describing the measures

Data for this study are from the Progress Monitoring and Reporting Network (PMRN), a database hosted and maintained by the Florida Department of Education. The measures used are the results on the state achievement test, the FCAT, and on the FAIR, administered three times a year for progress monitoring.

### Constructing a stratified subsample

The study drew on archival data from the PMRN on 1,132,263 students in grades 3–10 for 2009/10. The PMRN contains progress monitoring data in reading reported three times a year, as well as outcome data for the FCAT. A key consideration was that the findings reflect Florida's student population. As such, it was important to compare the demographics and academic achievement of students in the PMRN data with those of the student population in Florida as a whole. An initial investigation revealed that the PMRN sample did not precisely reflect the achievement distribution of all grade 3–10 students in the state. To correct for differences, a stratified subsample was constructed to reflect the observed achievement distribution across the five FCAT proficiency levels (see section on measures). State-aggregated data on the population distribution on the FCAT (see bottom of table 1) were used as known parameters for constructing the stratified random sample.

***A stratified subsample was constructed to reflect the observed achievement distribution across the five FCAT proficiency levels to more closely match the state population***

From the full PMRN sample (1,132,263 students), a stratified random sample of 800,000 students (100,000 per grade) was created. The achievement distribution for the stratified PMRN sample (see table 1) more closely matched the state population. The demographic characteristics of the stratified PMRN sample matched those of the state population as well: 51 percent male, 48 percent White, 24 percent Hispanic, 19 percent Black, 4 percent more than one race/ethnicity, 2 percent Asian, and less than 1 percent other (table 2). Approximately 7 percent of students were identified as English language learners, and 56 percent were eligible for free or reduced-price lunch, a proxy for low-income status.

### Explaining the measures

*Florida Comprehension Assessment Test.* The FCAT is part of Florida's effort to assess student achievement in reading, writing, math, and science, as outlined in Florida's Sunshine State Standards (Florida Department of Education, 2001). The FCAT reading comprehension subtest is an end-of-year, group-administered, criterion-referenced test consisting of informational and narrative reading passages with multiple-choice questions (Florida Department of Education, 2005). Students receive a developmentally scaled score and a proficiency level score, with level 1 the lowest proficiency and level 5 the highest. Students meet grade-level standards if they score at level 3 or higher. The current study used FCAT developmental scale scores from the end of the 2009/10 school year.

## Table 1. Proportion of students in the full PMRN sample, stratified random PMRN sample, and state population scoring at proficiency levels 1–5 on the 2010 FCAT, by grade

*(percent)*

| FCAT | Grade | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| proficiency level[a] | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| PMRN sample | | | | | | | | |
| 1 | 16 | 17 | 17 | 20 | 17 | 21 | 24 | 37 |
| 2 | 12 | 14 | 16 | 18 | 20 | 30 | 33 | 30 |
| 3 | 33 | 32 | 32 | 31 | 34 | 32 | 26 | 16 |
| 4 | 31 | 28 | 27 | 22 | 21 | 14 | 11 | 6 |
| 5 | 8 | 10 | 8 | 8 | 8 | 3 | 6 | 10 |
| Stratified PMRN sample | | | | | | | | |
| 1 | 16 | 16 | 15 | 17 | 14 | 17 | 21 | 32 |
| 2 | 12 | 13 | 15 | 16 | 17 | 27 | 30 | 29 |
| 3 | 33 | 32 | 33 | 32 | 34 | 34 | 28 | 18 |
| 4 | 31 | 29 | 28 | 26 | 24 | 17 | 13 | 8 |
| 5 | 8 | 11 | 9 | 9 | 10 | 4 | 7 | 13 |
| State population | | | | | | | | |
| 1 | 16 | 16 | 15 | 17 | 14 | 17 | 21 | 32 |
| 2 | 12 | 13 | 15 | 16 | 17 | 27 | 30 | 29 |
| 3 | 33 | 32 | 33 | 32 | 34 | 34 | 28 | 18 |
| 4 | 31 | 29 | 28 | 26 | 24 | 17 | 13 | 8 |
| 5 | 8 | 11 | 9 | 9 | 10 | 4 | 7 | 14 |

PMRN is Progress Monitoring and Reporting Network. FCAT is Florida Comprehensive Assessment Test.

**a.** Of the five proficiency levels on the FCAT, 1 is the lowest and 5 is the highest. Students are designated as meeting grade-level standards if they score at level 3 or higher.

**Source:** Authors' analysis based on data from Florida Department of Education (2010) and http://fcat.fldoe.org/results/default.asp.

## Table 2. Student demographics for the stratified PMRN sample, by grade, 2009/10

*(percent)*

| | | Grade | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Variable | Average | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Male | 51 | 51 | 51 | 51 | 52 | 51 | 51 | 52 | 50 |
| Race/ethnicity[a] | | | | | | | | | |
| White | 48 | 44 | 49 | 50 | 49 | 48 | 50 | 49 | 50 |
| Hispanic | 24 | 28 | 23 | 23 | 23 | 23 | 23 | 23 | 23 |
| Black | 19 | 22 | 21 | 20 | 19 | 22 | 22 | 22 | 21 |
| More than one | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 3 |
| Asian | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 3 |
| Other | 1 | 0.4 | 0.4 | 0.3 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 |
| English language learner student[b] | 7 | 11 | 8 | 7 | 6 | 5 | 5 | 5 | 6 |
| Student eligible for free or reduced-price lunch | 56 | 62 | 59 | 58 | 58 | 57 | 55 | 50 | 46 |

PMRN is Progress Monitoring and Reporting Network.

**a.** Unless otherwise noted, Hispanic includes Latino and Black includes African-American.

**b.** Students identified as English language learners took all assessments in English.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

Reported reliability for the FCAT is high, at .90 (Florida Department of Education, 2005). Moreover, content validity and concurrent validity of test scores have been established through a series of expert panel reviews and data analyses (Florida Department of Education, 2001). The validity of the FCAT as a comprehensive assessment of reading outcomes received strong empirical support in an analysis of its correlations with a variety of other reading comprehension, language, and basic reading measures (Schatschneider et al., 2004).

*Florida Assessments for Instruction in Reading.* The FAIR consists of interim reading assessments given three times a year in kindergarten through grade 10 (Florida Department of Education, 2009). In grades 3–10 students take a computer-adaptive reading comprehension screen consisting of up to three passages with multiple-choice questions similar in format to those on the FCAT. Performance is reported as an ability score (a developmental scaled score that can track growth from grade 3 through grade 10). The current study used the FAIR ability scores from the fall, winter, and spring assessments for the 2009/10 school year.

***Performance on the FAIR can help explain individual student differences on the FCAT beyond those explained by performance on the prior-year FCAT***

Reported reliability for the ability scores from the reading comprehension screen is at least .90 for 60 percent of students and at least .80 for 98 percent of students (Florida Department of Education, 2009). Recent technical reporting on the FAIR showed strong correlations ($r > .66$) across assessment points (fall, winter, and spring administration) for the FAIR reading comprehension screen for students in grades 3–10 (Foorman & Petscher, 2010a). In addition, the screen has been shown to explain individual differences in FCAT reading performance beyond that predicted using prior-year performance on the FCAT (average $\Delta R^2$ = 3.7 percent; Foorman & Petscher, 2010b). Together, these reports indicate that performance on the FAIR can help explain individual student differences on the FCAT beyond those explained by data for the prior-year FCAT.

*Missing data.* The amount of data missing in the stratified sample increased with grade level and decreased across FAIR assessment points within grade for all grades (table 3). Because all students are required to take the FCAT, missingness was not related to the outcome variable (end-of-year reading comprehension). Thus, the data were assumed to be missing at random.[3]

**Table 3. Rates of missing data for the three FAIR assessment points, by grade, 2009/10**

*(percent)*

| Grade | Fall | Winter | Spring |
|---|---|---|---|
| 3 | 4.0 | 2.8 | 4.7 |
| 4 | 4.6 | 4.1 | 5.2 |
| 5 | 5.0 | 4.3 | 5.3 |
| 6 | 9.4 | 6.6 | 6.4 |
| 7 | 10.0 | 7.2 | 7.3 |
| 8 | 10.7 | 8.1 | 7.2 |
| 9 | 13.4 | 10.0 | 9.4 |
| 10 | 26.6 | 21.3 | 22.1 |

FAIR is Florida Assessments for Instruction in Reading.

**Note:** The Florida Comprehension Assessment Test had 0 percent missing data due to the stratified random sampling procedure.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

As noted, a disadvantage of using OLS to estimate growth is that complete data are necessary for estimating parameters, whereas empirical Bayes can be used to estimate individual growth curves using full information maximum likelihood so missing data are not an impediment. To compare these measures of growth, therefore, it was necessary to evaluate both complete case and missing data conditions. All missing data were imputed using PROC MI in SAS.[4]

## Analyzing the data

This section describes the data analysis for each research question, looking at student growth in grades 3–10.

### What are the relations among descriptive measures of student change and inferential measures of individual growth curves?

To explore the relations among the four measures, each was used to calculate and estimate growth. For the two descriptive measures (simple difference and average difference) the observed measures of reading comprehension ability (fall, winter, and spring FAIR ability scores) were used. Calculating the simple difference score allows teachers to estimate how much change in reading performance relative to instruction has occurred and to compare student change. The simple difference score was calculated as the change occurring between the fall and winter assessments (the first estimate of change that can be calculated during the academic year using interim/benchmark assessments). The average difference score was calculated as the difference between the fall and spring assessments divided by the number of change scores (two) during the year. The simple difference between winter and spring was not calculated because it cannot be an actionable score for modifying instruction to help students meet an end-of-year benchmark for a state achievement test.

***The simple difference score is calculated as the change occurring between the fall and winter assessments. The average difference score is calculated as the difference between the fall and spring assessments divided by the number of change scores (two) during the year***

The two inferential measures of growth (OLS and empirical Bayes) were estimated using a multilevel growth model in HLM6 software (Raudenbush, Bryk, Cheong, & Congdon, 2004). Growth curve analyses were run for each grade, and the residual files were retained so that the OLS and empirical Bayes estimates could be used in the secondary multiple regression analysis. The individual growth curves were used to estimate the means and variances for each measure by grade and to evaluate the distribution of each measure. Simple bivariate correlations and scatterplots were used to examine the relations among growth measures.

### Controlling for students' mid-year status, how well do the measures of student change and individual growth curves explain differences in end-of-year reading comprehension performance?

A series of hierarchical multiple regression analyses were run for each grade level using the generated measures of growth. The baseline regression model included the intercept and the winter FAIR ability score. The resulting $R^2$ conveyed how much of the individual variation in the FCAT was explained by the winter FAIR reading comprehension ability score. To estimate a total $R^2$ based on both the winter FAIR and each growth measure, four additional models were run iteratively, with each measure of growth entered as a second independent variable. The difference between the total $R^2$ and the $R^2$ for each of the added-growth models was used to evaluate which measures of growth best explained differences in the FCAT for each grade level. Although there are methods for testing whether two $R^2$ values are statistically differentiated (Alf & Graf, 1999), such analysis would not yield meaningful information with a sample so large. Instead, the difference in $R^2$ values between estimators was compared using Cohen's (1988) criteria: $\Delta R^2$ of 2–12 percent is considered a small yet practically important effect; 13–25 percent, a moderate effect; and 26 percent or greater, a large effect.[5]

### Controlling for students' initial or final status, how well do the measures of student change and individual growth curves explain differences in end-of-year reading comprehension performance?

As with the previous research question, a series of hierarchical multiple regression analyses were run for each grade level, but for this question the baseline regression model was changed. The regression models were run using the fall FAIR reading comprehension ability score rather than the winter score as the status variable before including each growth measure. A second set of regressions was then run using the spring FAIR score as the status variable.

### Controlling for the type of estimator, how well do the measures of student change and individual growth curves explain differences in end-of-year reading comprehension performance conditional on end-of-year reading comprehension performance?

Quantile regressions were run using the quantreg procedure in SAS 9.3 for each grade level (SAS Institute Inc., 2012). Modeling procedures for the quantile analysis were similar to those for the previous research questions: a baseline model was iteratively run using the fall, winter, or spring assessment results as an independent variable but with four additional hierarchical multiple quantile regressions added, sequentially changing the slope coefficient as a predictor.

*All four growth estimates contributed significantly to the prediction of FCAT performance when controlling for initial (fall) status, only the simple difference was a good predictor when controlling for mid-year (winter) status, and all but the simple difference estimate when controlling for final (spring) status*

## Explaining the results

The extent to which individual differences in student FCAT performance were explained by each of the four growth estimates differs by status variable (performance on the fall, winter, or spring FAIR) and measure of growth used. All four growth estimates contributed significantly to the prediction of FCAT performance when controlling for initial (fall) status, as did all but the simple difference estimate when controlling for final (spring) status. But only the simple difference score (difference between the fall and winter test administrations) was a good predictor when controlling for mid-year (winter) status. Quantile regression suggested that the relations between growth estimates and FCAT scores were conditional on the outcome, implying that traditional linear regression analyses could mask the predictive relations.

### Descriptive analyses of reading comprehension test results

In 2009/10 the FAIR reading comprehension ability scores across grades 3–10 ranged from 200 to 800, and the FCAT developmental scale score ranged from 86 to 3008 (table 4). The mean FAIR reading comprehension ability score rose from fall to spring for all students across grades.

Table 4. Mean, standard deviation, minimum, and maximum for the fall, winter, and spring FAIR reading comprehension ability scores and the FCAT developmentally scaled scores, by grade, 2009/10

| Variable | Mean | Standard deviation | Minimum | Maximum |
|---|---|---|---|---|
| Grade 3 | | | | |
| Fall FAIR | 353.95 | 109.65 | 200 | 784 |
| Winter FAIR | 387.51 | 109.02 | 200 | 708 |
| Spring FAIR | 412.12 | 111.58 | 200 | 784 |
| FCAT | 1,386.86 | 371.22 | 86 | 2,514 |
| Grade 4 | | | | |
| Fall FAIR | 431.14 | 109.82 | 200 | 766 |
| Winter FAIR | 449.10 | 107.81 | 200 | 766 |
| Spring FAIR | 467.41 | 107.32 | 200 | 766 |
| FCAT | 1,599.32 | 334.03 | 295 | 2,638 |
| Grade 5 | | | | |
| Fall FAIR | 478.51 | 106.51 | 200 | 799 |
| Winter FAIR | 490.34 | 105.15 | 200 | 799 |
| Spring FAIR | 503.94 | 104.65 | 200 | 799 |
| FCAT | 1,653.21 | 334.88 | 474 | 2,713 |
| Grade 6 | | | | |
| Fall FAIR | 500.11 | 104.31 | 200 | 800 |
| Winter FAIR | 502.87 | 112.37 | 200 | 800 |
| Spring FAIR | 515.03 | 113.11 | 200 | 800 |
| FCAT | 1,727.62 | 342.57 | 539 | 2,758 |
| Grade 7 | | | | |
| Fall FAIR | 518.58 | 109.09 | 200 | 793 |
| Winter FAIR | 520.70 | 116.02 | 200 | 793 |
| Spring FAIR | 531.67 | 115.58 | 200 | 793 |
| FCAT | 1,832.20 | 301.73 | 671 | 2,767 |
| Grade 8 | | | | |
| Fall FAIR | 544.99 | 95.23 | 200 | 793 |
| Winter FAIR | 550.02 | 101.48 | 200 | 793 |
| Spring FAIR | 560.40 | 102.88 | 200 | 793 |
| FCAT | 1,893.81 | 237.71 | 886 | 2,790 |
| Grade 9 | | | | |
| Fall FAIR | 563.15 | 96.07 | 200 | 800 |
| Winter FAIR | 564.76 | 103.27 | 200 | 800 |
| Spring FAIR | 572.36 | 103.93 | 200 | 800 |
| FCAT | 1,948.38 | 269.92 | 772 | 2,943 |
| Grade 10 | | | | |
| Fall FAIR | 588.17 | 93.03 | 317 | 800 |
| Winter FAIR | 587.33 | 100.33 | 317 | 800 |
| Spring FAIR | 592.98 | 100.24 | 317 | 800 |
| FCAT | 1,970.62 | 330.04 | 844 | 3,008 |

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

## Reading comprehension performance remained fairly stable across the three assessment points of the Florida Assessments for Instruction in Reading, and correlations with the Florida Comprehensive Assessment Test were strong

The FAIR reading comprehension ability scores were strongly and positively correlated across the three assessment points, indicating that performance remained fairly stable within grades (table 5). Correlations of the fall and winter FAIR with the FCAT were strong within and across grades as well ($r$ = .70–.75 across grades and time points; table 6). Moreover, the concurrent correlation between the spring FAIR and the FCAT ranged from .70 to .76 across grades.

*What are the relations among descriptive measures of student change and inferential measures of individual growth curves?* The simple difference score was calculated as the difference between the fall and winter FAIR reading comprehension ability scores, whereas the average difference score was calculated as the difference between the spring and fall scores divided by two. The two inferential measures of growth were the residuals from the multilevel model nesting time within student.[6]

***The FAIR reading comprehension ability scores were strongly and positively correlated across the three assessment points, indicating that performance remained fairly stable within grades***

**Table 5. Correlations among the three FAIR reading comprehension ability scores, by grade, 2009/10**

| Grade | Fall to winter | Winter to spring | Fall to spring |
|---|---|---|---|
| 3 | .74 | .75 | .70 |
| 4 | .70 | .73 | .67 |
| 5 | .71 | .68 | .73 |
| 6 | .74 | .71 | .74 |
| 7 | .73 | .70 | .73 |
| 8 | .74 | .70 | .73 |
| 9 | .73 | .69 | .72 |
| 10 | .73 | .72 | .69 |

FAIR is Florida Assessments for Instruction in Reading.

**Note:** All correlations are significant at $p < .001$.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Table 6. Correlations among the three FAIR reading comprehension ability scores and the FCAT developmental scaled score, by grade, 2009/10**

| Grade | Fall FAIR and FCAT | Winter FAIR and FCAT | Spring FAIR and FCAT |
|---|---|---|---|
| 3 | .73 | .75 | .76 |
| 4 | .70 | .73 | .75 |
| 5 | .71 | .73 | .74 |
| 6 | .75 | .74 | .74 |
| 7 | .73 | .71 | .71 |
| 8 | .74 | .73 | .73 |
| 9 | .72 | .70 | .70 |
| 10 | .74 | .71 | .71 |

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** All correlations are significant at $p < .001$.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

Change-based means on the descriptive measures revealed that simple difference scores across grades ranged from –0.84 point in grade 10 to 33.56 points in grade 3 (table 7). A negative average score may appear counterintuitive considering that a developmentally scaled score should produce average increases over time. Indeed, comparing the observed means in table 4 (588.17 in the fall and 587.33 in the winter for grade 10) shows that the decrease from fall to winter was negligible (Cohen's *d* = 0.009). A similar pattern was

**Table 7. Mean scores, standard deviations, minimum scores, and maximum scores for descriptive and inferential growth estimates, by grade, 2009/10**

| Variable | Mean | Standard deviation | Minimum | Maximum |
|---|---|---|---|---|
| Grade 3 | | | | |
| Simple difference | 33.56 | 77.58 | –472.00 | 446.00 |
| Average difference | 29.09 | 42.64 | –224.50 | 254.00 |
| Ordinary least squares | 0.00 | 12.46 | –68.54 | 65.58 |
| Empirical Bayes | 0.00 | 2.19 | –11.61 | 11.45 |
| Grade 4 | | | | |
| Simple difference | 17.96 | 88.84 | –442.00 | 505.00 |
| Average difference | 18.13 | 44.32 | –204.50 | 245.00 |
| Ordinary least squares | 0.00 | 13.03 | –61.37 | 66.96 |
| Empirical Bayes | 0.00 | 2.43 | –11.47 | 12.56 |
| Grade 5 | | | | |
| Simple difference | 11.83 | 80.49 | –456.00 | 464.00 |
| Average difference | 12.71 | 42.43 | –244.50 | 249.00 |
| Ordinary least squares | 0.00 | 12.53 | –74.54 | 68.56 |
| Empirical Bayes | 0.00 | 2.08 | –12.30 | 11.46 |
| Grade 6 | | | | |
| Simple difference | 2.76 | 79.17 | –600.00 | 488.00 |
| Average difference | 7.46 | 41.96 | –246.00 | 227.50 |
| Ordinary least squares | 0.00 | 13.03 | –70.77 | 63.81 |
| Empirical Bayes | 0.00 | 1.92 | –8.58 | 7.40 |
| Grade 7 | | | | |
| Simple difference | 2.12 | 82.77 | –494.00 | 593.00 |
| Average difference | 6.54 | 43.77 | –296.50 | 296.50 |
| Ordinary least squares | 0.00 | 13.64 | –83.68 | 83.15 |
| Empirical Bayes | 0.00 | 1.77 | –10.13 | 9.41 |
| Grade 8 | | | | |
| Simple difference | 5.03 | 71.89 | –487.00 | 416.00 |
| Average difference | 7.71 | 38.55 | –196.50 | 246.00 |
| Ordinary least squares | 0.00 | 12.05 | –59.59 | 66.29 |
| Empirical Bayes | 0.00 | 1.85 | –8.28 | 7.85 |
| Grade 9 | | | | |
| Simple difference | 1.62 | 73.37 | –457.00 | 483.00 |
| Average difference | 4.61 | 39.70 | –224.00 | 205.50 |
| Ordinary least squares | 0.00 | 12.62 | –66.15 | 56.40 |
| Empirical Bayes | 0.00 | 1.90 | –8.96 | 7.09 |
| Grade 10 | | | | |
| Simple difference | –0.84 | 71.52 | –457.00 | 447.00 |
| Average difference | 2.40 | 38.16 | –228.50 | 231.00 |
| Ordinary least squares | 0.00 | 12.27 | –67.33 | 65.20 |
| Empirical Bayes | 0.00 | 1.81 | –8.66 | 7.80 |

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

observed for the average difference score ranged from a between-assessment gain of 2.40 points in grade 10 to 29.09 points in grade 3. The pattern of average change across grades was fairly systematic: the largest gains were made by students in the lowest grades.

While the OLS and empirical Bayes both had a mean of 0 across grades, their standard deviations differed. Across grades 3–10 standard deviations ranged from 12.05 to 13.64 for OLS-based individual growth curves and from 1.77 to 2.43 for empirical Bayes–based individual growth curves (see table 7). This difference was no surprise considering the shrinkage in growth when estimated using empirical Bayes (the individual slope is weighted by the average of the sample).

To better understand the relations among the four slope scores, as well as their bivariate correlations with the FCAT, figure 4 depicts a matrix scatterplot that includes Pearson correlations (upper diagonal), histograms (diagonal), and scatterplots (lower diagonal). The histograms for all variables within grades show that the scores follow a fairly normal distribution.

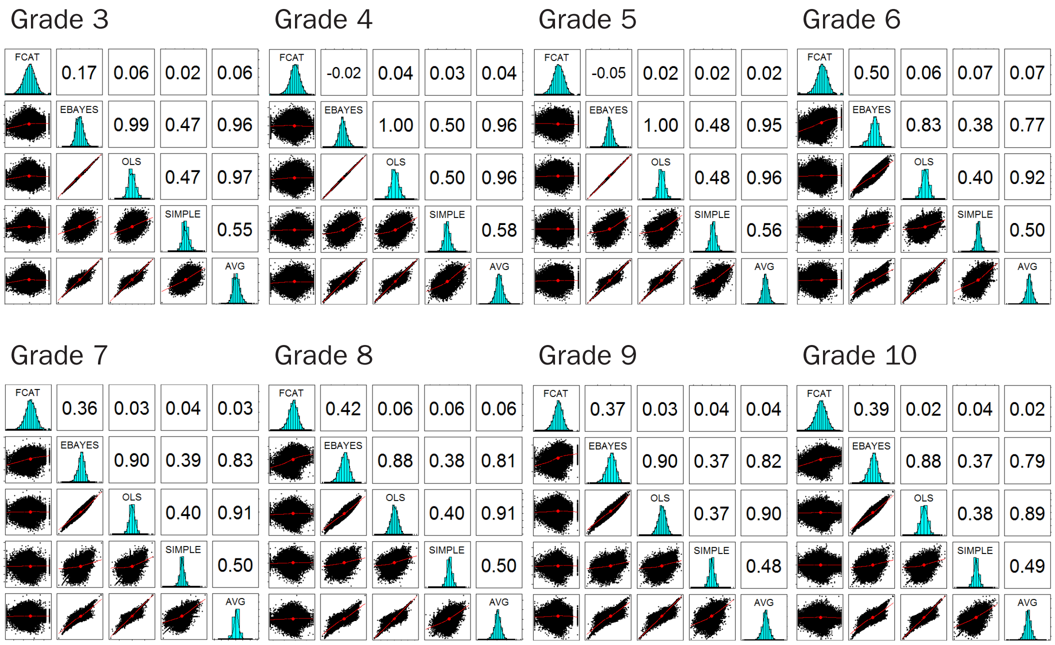
Several correlational trends are worth noting. First, the growth measures are shown to be moderately to perfectly correlated across grades. For students in grades 3–5 a nearly perfect correlation is observed between empirical Bayes and OLS ($r$ = .99–1.00), empirical Bayes and average difference ($r$ = .95–.96), and OLS and average difference ($r$ = .96–.97). For students in grades 6–10 the correlations decrease slightly but remain strong ($r$ = .83–.90 between empirical Bayes and OLS, $r$ = .77–.83 between empirical Bayes and average difference, and $r$ = .89–.92 between OLS and average difference). Further, for students in grades 3–10 moderate correlations are observed between simple difference and empirical Bayes ($r$ = .38–.50), simple difference and OLS ($r$ = .37–.51), and simple difference and average difference ($r$ = .48–.58).

***For students in grades 3–5 a nearly perfect correlation is observed between empirical Bayes and OLS, empirical Bayes and average difference, and OLS and average difference; for students in grades 6–10 the correlations decrease slightly but remain strong***

Second, as already noted, a limitation of the progress monitoring literature is that many of the studies evaluating the relation of growth with outcomes used samples that did not consist predominately of low-ability/high-risk students. The same criticism can be leveled here, as the sample of 100,000 students in each grade contained students with low, average, and above average reading ability (see table 1). The scatterplots in figure 4 corroborate this in that the correlations among growth measures are stronger at the lowest end of the reading ability distribution. That being the case, the Pearson correlations shown in figure 4 would mask the correlations among variables for students at the lowest end of the reading ability distribution because the coefficients were estimated in a conditional means model (which estimates the average relation).

To evaluate whether correlations at the lowest end of the ability distribution were being masked by the Pearson correlations displayed in figure 4, simple quantile regressions were run at the .25 quantile (25th percentile) for the correlations among growth measures. A comparison of the bivariate quantile correlations (table 8) with the Pearson correlations (figure 4) shows that the correlations did not differ more than .03 for any grade between empirical Bayes and OLS, empirical Bayes and average difference, or OLS and average difference. In addition, the correlations between simple difference and all other measures of growth did not increase more than .06, suggesting that the Pearson correlation adequately captures the correlations among growth measures for students at the lowest end of the reading ability distribution.

**Figure 4. Matrix scatterplot depicting Pearson correlations (upper diagonal), histograms (diagonal), and scatterplots (lower diagonal) for the FCAT and growth measures, by grade, 2009/10**



FCAT is Florida Comprehensive Assessment Test. EBAYES is empirical Bayes. OLS is ordinary least squares. SIMPLE is simple difference. AVG is average difference.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Table 8. Correlations among growth measures at the .25 quantile, by grade, 2009/10**

| Grade | Empirical Bayes and ordinary least squares | Empirical Bayes and average difference | Empirical Bayes and simple difference | Ordinary least squares and average difference | Ordinary least squares and simple difference | Average difference and simple difference |
|---|---|---|---|---|---|---|
| 3 | 1.00 | .98 | .51 | .97 | .47 | .55 |
| 4 | 1.00 | .96 | .50 | .96 | .50 | .59 |
| 5 | 1.00 | .96 | .48 | .96 | .49 | .57 |
| 6 | .84 | .78 | .43 | .92 | .44 | .54 |
| 7 | .91 | .84 | .44 | .92 | .44 | .55 |
| 8 | .89 | .82 | .44 | .92 | .44 | .55 |
| 9 | .92 | .83 | .43 | .90 | .42 | .54 |
| 10 | .90 | .82 | .41 | .89 | .41 | .52 |

**Note:** All correlations were significant at $p < .001$.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

Third, unlike the correlations among the growth measures, the correlations between the growth measures and the FCAT were near zero for all grades (see figure 4), with the exception of the correlation between empirical Bayes and the FCAT for students in grades 6–10 ($r$ = .50 in grade 6; $r$ = .36 in grade 7; $r$ = .42 in grade 8; $r$ = .37 in grade 9; $r$ =.39 in grade 10). The lack of a correlation between most of the slope measures and the FCAT suggests that differences in the individual growth curves would not explain differences in the FCAT outcome but that once status was included as a predictor, individual growth curves could contribute.

*Controlling for students' mid-year status, how well do the measures of student change and individual growth curves explain differences in end-of-year reading comprehension performance?* Hierarchical multiple regressions used to estimate how well the four measures of growth explain differences in the FCAT after accounting for mid-year (winter) status on the FAIR found that the measures accounted for 49–56 percent of the variance for students in grades 3–10 in the baseline model (table 9).[7] Adding the average difference score and the OLS slope did not explain FCAT differences among students beyond that accounted for by the winter assessment, and adding the empirical Bayes explained only a negligible amount of additional variance ($\Delta R^2$ = 0–2 percent). By contrast, adding the simple difference score (change from the fall FAIR to the winter FAIR) explained an additional 6–10 percent of the variance in FCAT scores across grades 3–10. Based on Cohen's criteria for evaluating the strength of an $R^2$, these measures added small yet practically important effects.

*Controlling for students' initial or final status, how well do the measures of student change and individual growth curves explain differences in end-of-year reading comprehension performance?* Results for the hierarchical multiple regressions controlling for students' initial (fall) status showed that the base model accounted for 48–56 percent of the variance in FCAT performance for students in grades 3–10 (table 10).[8] In general, all growth measures explained additional variance in FCAT performance for all grades after controlling for fall status. Simple difference explained an additional 7–12 percent of variance, average difference explained 7–14 percent, OLS explained 5–12 percent, and empirical Bayes explained 6–12 percent. As was the case for the base (winter status variable) model, the amount of variance explained by the four growth measures was interpreted as a small effect.

***Results for the hierarchical multiple regressions controlling for students' initial (fall) status showed that the base model accounted for 48–56 percent of the variance in FCAT performance for students in grades 3–10***

Results for the hierarchical multiple regressions controlling for students' final (spring) status showed that the base model accounted for 49–58 percent of the variance in FCAT performance for students in grades 3–10 (table 11).[9] In general, average difference, OLS, and empirical Bayes measures of growth explained significant variance in FCAT performance after controlling for fall status, whereas the simple difference measure of growth explained no additional variance. Average difference added 7–12 percentage points, OLS added 7–10 percentage points, and empirical Bayes added 5–8 percentage points. Auxiliary comparisons between growth measures controlling for spring status showed that average

**Table 9. Proportion of variance in FCAT scores explained by growth measures after controlling for mid-year (winter) status (base model) on the FAIR, by grade, 2009/10**

| | | $\Delta R^2$ from base model | | | |
|---|---|---|---|---|---|
| Grade | Base | Simple difference | Average difference | Ordinary least squares | Empirical Bayes |
| 3 | 56 | 7 | 0 | 0 | 1 |
| 4 | 54 | 6 | 0 | 0 | 0 |
| 5 | 54 | 7 | 0 | 0 | 0 |
| 6 | 54 | 9 | 0 | 0 | 2 |
| 7 | 51 | 9 | 0 | 0 | 1 |
| 8 | 53 | 9 | 0 | 0 | 1 |
| 9 | 49 | 9 | 0 | 0 | 1 |
| 10 | 51 | 10 | 0 | 0 | 1 |

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** $\Delta R^2 \geq 1$ percent is significant at $p < .001$.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

### Table 10. Proportion of variance in the FCAT scores explained by growth measures after controlling for initial (fall) status (base model), by grade, 2009/10

| | | $\Delta R^2$ from base model | | | |
|---|---|---|---|---|---|
| Grade | Base | Simple difference | Average difference | Ordinary least squares | Empirical Bayes |
| 3 | 53 | 10 | 12 | 11 | 11 |
| 4 | 48 | 12 | 14 | 12 | 12 |
| 5 | 50 | 11 | 12 | 10 | 10 |
| 6 | 56 | 8 | 9 | 7 | 9 |
| 7 | 53 | 7 | 8 | 6 | 7 |
| 8 | 54 | 8 | 9 | 7 | 8 |
| 9 | 52 | 7 | 8 | 6 | 7 |
| 10 | 55 | 7 | 7 | 5 | 6 |

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** $\Delta R^2 \geq 1$ percent is significant at $p < .001$.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

***Results for the hierarchical multiple regressions controlling for students' final (spring) status showed that the base model accounted for 49–58 percent of the variance in FCAT performance for students in grades 3–10***

### Table 11. Proportion of variance in the FCAT scores explained by growth after controlling for final (spring) status (base model), by grade, 2009/10

| | | $\Delta R^2$ from base model | | | |
|---|---|---|---|---|---|
| Grade | Base | Simple difference | Average difference | Ordinary least squares | Empirical Bayes |
| 3 | 58 | 0 | 7 | 8 | 7 |
| 4 | 56 | 0 | 7 | 7 | 7 |
| 5 | 54 | 0 | 8 | 8 | 8 |
| 6 | 55 | 0 | 10 | 9 | 5 |
| 7 | 51 | 0 | 11 | 9 | 7 |
| 8 | 53 | 0 | 10 | 9 | 6 |
| 9 | 49 | 0 | 11 | 9 | 6 |
| 10 | 50 | 0 | 12 | 10 | 7 |

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** $\Delta R^2 \geq 1$ percent is significant at $p < .001$.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

difference, OLS, and empirical Bayes explained the same amount of variance in FCAT scores for students in grades 3–5. However, in grades 6 and 8 average difference and OLS explained 3–5 percentage points more variance than empirical Bayes. Further, in grades 7, 9, and 10 average difference explained 2 percentage points more variance than OLS, which explained 2–3 percentage points more than empirical Bayes. Except for the simple difference score, the amount of variance explained by the growth measures was interpreted as a small effect.

*Controlling for the type of estimator, how well do the measures of student change and individual growth curves explain differences in end-of-year reading comprehension performance conditional on end-of-year reading comprehension performance?* This research question was explored using the same hierarchical multiple regression models tested for the other three questions but in a quantile regression framework. Appendixes B, C, and D contain quantile process plots that depict the relation between growth and reading comprehension conditional on FCAT after controlling for status. For each measure of growth the appendixes contain plots for the intercept and for the two predictors. The intercept plot shows that students at lower quantiles of reading comprehension have lower reading comprehension ability scores. The other two plots reflect the slope coefficients for status and growth conditional on reading comprehension ability. The growth plots show that the slope coefficient decreases as reading comprehension ability increases.

Several trends emerged across grades, statuses, and growth measures. When controlling for initial (fall) status (appendix B), growth had a stronger relation to the FCAT score beyond that predicted by fall status for students at the lowest comprehension ability level. Across most grades and measures of growth the coefficient for growth was larger for students whose fall FAIR scores were below the .20 quantile (20th percentile).

This trend was amplified when controlling for mid-year (winter) status (appendix C) for grades 6–10; however, the direction of the coefficient for growth was dependent on end-of-year reading comprehension performance: students with low FCAT scores had a positive slope coefficient, and students with high FCAT scores had a negative slope coefficient.

When controlling for final (spring) status (appendix D), average difference, OLS, and empirical Bayes demonstrated a trend of equal coefficients for growth across the quantiles, though small increases were observed from the lower to the upper ends of the distribution. The trend for simple difference across grades, however, consistently showed a negative coefficient across the quantiles that was stronger for students in the middle of the distribution and weaker for those at the tails.

***The extent to which individual growth curves explain variance beyond what can be explained by any one status variable depends on grade level, growth measure, and the status variable controlled for***

## Conclusions, implications, and limitations

The results from a combination of multilevel growth curves, hierarchical multiple regression, and quantile regression suggest that the extent to which individual growth curves explain variance beyond what can be explained by any one status variable depends on grade level, growth measure, and the status variable controlled for.

Although grade-level differences were not specifically examined, the data suggest that grade level interacts with the status variable when the effect of the individual growth curve is compared for elementary school and either middle or high school. In grades 3–5 the average amount of variance in FCAT scores explained by the individual growth curve was 11 percent when controlling for the fall FAIR score and 6 percent when controlling for the spring FAIR score, while in grades 6–10 it was 8 percent when controlling for either the fall or the spring score.

The status covariate was found to affect how individual growth curves relate to FCAT performance. When the fall score was used as the status variable in the base model, all four growth measures added a small yet practically important contribution (average 9–11 percent) to the prediction of the FCAT score. When the spring FAIR score was used, the simple difference growth measure was no longer predictive, and when the winter score was used, the simple difference growth measure was the only one that consistently added a practically important contribution to the explanation of differences in the outcome across grades.

How well the type of growth estimate explained student FCAT scores was most strongly associated with the status variable covariate. Bivariate correlations demonstrated that the slope coefficients were at least moderately associated with one another (see figure 1). The average correlation among the four growth measures across grades was $r$ = .70, suggesting that the stability or rank ordering of students by slope was fairly consistent across the estimates. Although the correlations among the slope estimates were large, the correlations of the slope estimates and the outcome depended entirely on the status variable (except for the empirical Bayes slope). This finding is consistent with estimates reported by Zumeta et al. (2012), who found near-zero correlations for slope with the selected outcomes of decoding and reading fluency for the representative sample of 25 percent students of low ability, 50 percent of average ability, and 25 percent of high ability.

Traditional regression analysis might mask predictive correlations with an outcome. Consider that the analysis of the effect of individual growth curves when controlling for winter status on the FAIR suggested that the average difference, OLS, and empirical Bayes growth measures consistently did not explain differences in the FCAT. However, this result might reflect only the average for the sample. Quantile regression (appendix C) yielded several results that were small but statistically different from zero for several slope scores in grades 6 and higher. This finding needs to be contextualized. Many of the observed negative or nonzero effects for slope occurred at quantiles greater than .80 and less than .20. A body of research in quantile regression summarized by Petscher et al. (2013) noted that such values should be interpreted with caution because there may be fewer individuals at the extreme quantiles. Thus, although the average amount of variance explained in FCAT scores may be zero, it might well be that this would change for students with low or high reading comprehension ability. More research is needed to confirm such observations from the present sample.

***When considering how well growth explains differences in outcomes, it is important to think about how to characterize growth and about which status variable is most appropriate***

When considering how well growth explains differences in outcomes, it is important to think about how to characterize growth and about which status variable is most appropriate. One option is to take a developmental progression perspective. Statistical models are agnostic to the data, and it is up to the user to define a model that is both statistically and theoretically sound. When the fall score is used as the status variable, any growth has yet to take place, so using the estimated slope yields little practical information, on average, for practitioners. When the winter score is used in the baseline model, the simple difference growth measure may be the most meaningful because the average difference, OLS, and empirical Bayes growth measures all incorporate information on students that is not yet available when controlling for the status variable. The only status variable for which a within-year growth estimate should be used is the spring score; from a practical perspective, however, the spring score is less useful for teachers and practitioners looking for a growth measure to use during the year to identify students requiring intervention.

Statistically, the developmental progression perspective resolves many issues with the apparently conflicting results based on centering. Consider the results for grade 3 students. The statistical outcomes for the fall model showed that 53 percent of the variance in the FCAT outcome was explained by the base model and 10 percent by individual growth curves, for a total of 63 percent. In the winter model, the base model accounted for 56 percent of the variance, an increase of 3 percentage points (56 percent minus 53 percent), but adding individual growth curves would not explain FCAT differences. The idea of losing explanatory power as the criterion is approached does not sit well: It is difficult to explain how a 63 percent total variance explained in the fall drops to 56 percent in the winter and then increases to 65 percent in the spring (57 percent base model plus 8 percent additional variance). Taking a developmental perspective instead shows that 53 percent of the variance is explained in the fall (no slope added), 63 percent in the winter (simple difference score added), and 65 percent in the spring (average difference, OLS, or empirical Bayes score added).

Taken as a whole, these findings suggest several broad recommendations, with the noted limitation that such generalizations relate solely to the data used in this study:

- When evaluating the within-year effects of individual growth curves, using OLS- or empirical Bayes–estimated individual growth curves is relevant from a developmental perspective only when the analysis controls for the final (spring) assessment rather than the initial or mid-year assessment. Although it was observed that the OLS and empirical Bayes individual growth curves explained student differences in the FCAT after controlling for the fall assessment, this appears to be due to including more statistical information in the model than a developmental perspective would allow. An individual growth curve with fall status in the statistical model contains information on growth that occurs during the academic year. Thus, how well individual growth curves explain differences in the FCAT beyond the fall assessment is based on the fact that growth estimates inherently include information on student performance in the winter and spring. It is natural to expect that an individual growth curve would predict beyond the fall status variable. Thus, from a developmental perspective, the information that individual growth curves provide to explain student differences on FCAT is potentially misleading.
- For both researchers and practitioners the simple difference score may provide valuable information on student differences in an outcome beyond that provided by the mid-year status variable. The results in this study showed a statistically significant effect for the simple difference score. It is developmentally appropriate to include this measure of student change in the statistical model because it includes mid-year status. The simple difference score does not contain additional information on future performance when mid-year status is included. Thus, its statistical relevance, coupled with its ease of calculation, suggests that this score type may be a useful measure of student change to explain differences in an outcome beyond a status variable.
- The average difference score may provide information on student performance differences on the FCAT beyond that of the spring assessment. This score type explained student differences in the FCAT at a level comparable to the OLS and empirical Bayes individual growth curves. Because this score type was statistically relevant in explaining student differences beyond the spring assessment and is computationally simple for researchers and practitioners, it should be further explored as an explanatory variable of student outcomes in conjunction with the simple difference score.

***The statistical relevance, coupled with ease of calculation, suggests that the simple difference score may be a useful measure of student change to explain differences in an outcome beyond a status variable***

While the findings of this study expand on the previous research on the value of interim assessments beyond primary grades (Kim et al., 2010; Schatschneider et al., 2008; Yeo et al., 2012) to middle and secondary school grades using a large sample, they are limited by the measures used in the population, the subject matter assessed (reading comprehension), the frequency of assessments, and the type of student growth estimates used. The findings might differ if the number of interim assessments changed or if other growth measures were used. Future work could examine the reliability of the score types and how prior-year individual growth curves could be used to inform predictions beyond those of the fall and winter status variables and prior-year FCAT performance. In that way, individual growth curves could be informative predictors beyond those assessment periods.

# Appendix A. Unstandardized regression coefficients for each model by grade and controlling for status

**Table A1. Summary of hierarchical multiple regressions using different measures of growth to predict end-of-year reading comprehension scores, controlling for status, grade 3, 2009/10**

| Model | Variable | B | Standard error | *t*-value |
|---|---|---|---|---|
| 1 | Intercept | 516.50 | 2.73 | 189.43 |
| | Fall FAIR | 2.46 | 0.01 | 334.17 |
| 2a | Intercept | 427.66 | 2.44 | 175.17 |
| | Fall FAIR | 2.71 | 0.01 | 410.45 |
| | Empirical Bayes | 57.74 | 0.33 | 174.38 |
| 2b | Intercept | 375.10 | 2.54 | 147.44 |
| | Fall FAIR | 2.86 | 0.01 | 414.21 |
| | Ordinary least squares | 10.30 | 0.06 | 169.46 |
| 2c | Intercept | 257.08 | 2.73 | 94.14 |
| | Fall FAIR | 2.92 | 0.01 | 429.14 |
| | Average difference | 3.26 | 0.02 | 186.31 |
| 2d | Intercept | 312.52 | 2.73 | 114.64 |
| | Fall FAIR | 2.89 | 0.01 | 406.79 |
| | Simple difference | 1.58 | 0.01 | 163.24 |
| 1 | Intercept | 400.31 | 2.88 | 139.08 |
| | Winter FAIR | 2.55 | 0.01 | 356.05 |
| 2a | Intercept | 412.49 | 2.88 | 143.28 |
| | Winter FAIR | 2.51 | 0.01 | 351.41 |
| | Empirical Bayes | 13.01 | 0.36 | 36.46 |
| 2b | Intercept | 400.25 | 2.87 | 139.58 |
| | Winter FAIR | 2.55 | 0.01 | 357.42 |
| | Ordinary least squares | 1.72 | 0.06 | 27.55 |
| 2c | Intercept | 392.58 | 2.91 | 134.97 |
| | Winter FAIR | 2.54 | 0.01 | 355.90 |
| | Average difference | 0.32 | 0.02 | 17.27 |
| 2d | Intercept | 312.52 | 2.83 | 114.64 |
| | Winter FAIR | 2.89 | 0.01 | 409.79 |
| | Simple difference | –1.30 | 0.01 | –134.96 |
| 1 | Intercept | 346.59 | 2.93 | 118.44 |
| | Spring FAIR | 2.52 | 0.01 | 368.30 |
| 2a | Intercept | 120.79 | 3.09 | 39.05 |
| | Spring FAIR | 3.07 | 0.01 | 420.14 |
| | Empirical Bayes | –53.63 | 0.37 | –143.72 |
| 2b | Intercept | 178.68 | 2.89 | 61.90 |
| | Spring FAIR | 2.93 | 0.01 | 431.16 |
| | Ordinary least squares | –8.98 | 0.06 | –147.39 |
| 2c | Intercept | 257.08 | 2.73 | 94.14 |
| | Spring FAIR | 2.92 | 0.01 | 429.14 |
| | Average difference | –2.58 | 0.02 | –144.94 |
| 2d | Intercept | 347.98 | 2.93 | 118.75 |
| | Spring FAIR | 2.53 | 0.01 | 368.33 |
| | Simple difference | –0.08 | 0.01 | –8.05 |

FAIR is Florida Assessments for Instruction in Reading.

**Note:** All coefficients are significant at $p < .001$.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Table A2. Summary of hierarchical multiple regressions using different measures of growth to predict end-of-year reading comprehension scores, controlling for status, grade 4, 2009/10**

| Model | Variable | B | Standard error | *t*-value |
|---|---|---|---|---|
| 1 | Intercept | 688.01 | 3.08 | 223.58 |
| | Fall FAIR | 2.11 | 0.01 | 305.59 |
| 2a | Intercept | 450.61 | 3.04 | 148.05 |
| | Fall FAIR | 2.66 | 0.01 | 386.86 |
| | Empirical Bayes | 53.16 | 0.31 | 170.79 |
| 2b | Intercept | 488.79 | 2.93 | 167.13 |
| | Fall FAIR | 2.58 | 0.01 | 389.92 |
| | Ordinary least squares | 9.73 | 0.06 | 174.77 |
| 2c | Intercept | 392.88 | 3.02 | 130.22 |
| | Fall FAIR | 2.67 | 0.01 | 409.04 |
| | Average difference | 3.17 | 0.02 | 196.15 |
| 2d | Intercept | 462.03 | 3.01 | 153.47 |
| | Fall FAIR | 2.58 | 0.01 | 387.13 |
| | Simple difference | 1.49 | 0.01 | 171.27 |
| 1 | Intercept | 580.55 | 3.08 | 188.36 |
| | Winter FAIR | 2.69 | 0.01 | 339.94 |
| 2a | Intercept | 574.95 | 3.09 | 186.19 |
| | Winter FAIR | 2.28 | 0.01 | 341.08 |
| | Empirical Bayes | 6.05 | 0.30 | 20.38 |
| 2b | Intercept | 578.99 | 3.07 | 188.44 |
| | Winter FAIR | 2.27 | 0.01 | 341.52 |
| | Ordinary least squares | 1.42 | 0.06 | 25.83 |
| 2c | Intercept | 576.43 | 3.09 | 186.59 |
| | Winter FAIR | 2.27 | 0.01 | 340.16 |
| | Average difference | 0.26 | 0.02 | 15.80 |
| 2d | Intercept | 462.03 | 3.01 | 153.47 |
| | Winter FAIR | 2.58 | 0.01 | 387.13 |
| | Simple difference | –1.08 | 0.01 | –126.61 |
| 1 | Intercept | 514.24 | 3.14 | 163.56 |
| | Spring FAIR | 2.32 | 0.01 | 354.11 |
| 2a | Intercept | 385.31 | 3.02 | 127.65 |
| | Spring FAIR | 2.60 | 0.01 | 411.64 |
| | Empirical Bayes | –39.03 | 0.28 | –140.03 |
| 2b | Intercept | 352.97 | 3.11 | 113.51 |
| | Spring FAIR | 2.67 | 0.01 | 409.72 |
| | Ordinary least squares | –7.40 | 0.05 | –137.98 |
| 2c | Intercept | 392.88 | 3.02 | 130.22 |
| | Spring FAIR | 2.67 | 0.01 | 409.04 |
| | Average difference | –2.16 | 0.02 | –137.15 |
| 2d | Intercept | 513.89 | 3.14 | 163.49 |
| | Spring FAIR | 2.33 | 0.01 | 353.99 |
| | Simple difference | –0.07 | 0.01 | –7.71 |

FAIR is Florida Assessments for Instruction in Reading.

**Note:** All coefficients are significant at $p < .001$.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Table A3. Summary of hierarchical multiple regressions using different measures of growth to predict end-of-year reading comprehension scores, controlling for status, grade 5, 2009/10**

| Model | Variable | B | Standard error | *t*-value |
|---|---|---|---|---|
| 1 | Intercept | 586.13 | 3.44 | 170.57 |
| | Fall FAIR | 2.23 | 0.01 | 318.13 |
| 2a | Intercept | 344.41 | 3.44 | 100.02 |
| | Fall FAIR | 2.74 | 0.01 | 387.48 |
| | Empirical Bayes | 56.76 | 0.36 | 156.76 |
| 2b | Intercept | 383.79 | 3.31 | 115.82 |
| | Fall FAIR | 2.65 | 0.01 | 391.05 |
| | Ordinary least squares | 9.25 | 0.06 | 160.44 |
| 2c | Intercept | 303.86 | 3.37 | 90.23 |
| | Fall FAIR | 2.74 | 0.01 | 408.43 |
| | Average difference | 3.04 | 0.02 | 180.71 |
| 2d | Intercept | 359.72 | 3.35 | 107.29 |
| | Fall FAIR | 2.67 | 0.01 | 393.49 |
| | Simple difference | 1.47 | 0.01 | 163.37 |
| 1 | Intercept | 510.45 | 3.44 | 148.28 |
| | Winter FAIR | 2.33 | 0.01 | 339.5 |
| 2a | Intercept | 506.43 | 3.46 | 146.55 |
| | Winter FAIR | 2.34 | 0.01 | 339.32 |
| | Empirical Bayes | 4.26 | 0.35 | 12.21 |
| 2b | Intercept | 508.88 | 3.44 | 148.00 |
| | Winter FAIR | 2.33 | 0.01 | 340.37 |
| | Ordinary least squares | 1.01 | 0.06 | 17.59 |
| 2c | Intercept | 508.97 | 3.45 | 147.64 |
| | Winter FAIR | 2.33 | 0.01 | 339.55 |
| | Average difference | 0.13 | 0.02 | 7.40 |
| 2d | Intercept | 359.72 | 3.35 | 107.29 |
| | Winter FAIR | 2.67 | 0.01 | 393.49 |
| | Simple difference | –1.20 | 0.01 | –135.72 |
| 1 | Intercept | 463.63 | 3.52 | 131.84 |
| | Spring FAIR | 2.36 | 0.01 | 345.49 |
| 2a | Intercept | 312.71 | 3.34 | 93.65 |
| | Spring FAIR | 2.66 | 0.01 | 409.18 |
| | Empirical Bayes | –48.78 | 0.33 | –148.88 |
| 2b | Intercept | 274.69 | 3.44 | 79.85 |
| | Spring FAIR | 2.74 | 0.01 | 408.06 |
| | Ordinary least squares | –8.22 | 0.06 | –146.75 |
| 2c | Intercept | 303.86 | 3.37 | 90.23 |
| | Spring FAIR | 2.74 | 0.01 | 408.43 |
| | Average difference | –2.44 | 0.02 | –147.24 |
| 2d | Intercept | 462.77 | 3.52 | 131.63 |
| | Spring FAIR | 2.36 | 0.01 | 345.75 |
| | Simple difference | –0.09 | 0.01 | –10.59 |

FAIR is Florida Assessments for Instruction in Reading.

**Note:** All coefficients are significant at $p < .001$.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Table A4. Summary of hierarchical multiple regressions using different measures of growth to predict end-of-year reading comprehension scores, controlling for status, grade 6, 2009/10**

| Model | Variable | B | Standard error | *t*-value |
|---|---|---|---|---|
| 1 | Intercept | 503.56 | 3.54 | 142.32 |
| | Fall FAIR | 2.45 | 0.01 | 353.41 |
| 2a | Intercept | 653.12 | 3.31 | 197.24 |
| | Fall FAIR | 2.15 | 0.01 | 330.88 |
| | Empirical Bayes | 55.20 | 0.35 | 156.61 |
| 2b | Intercept | 388.19 | 3.36 | 115.68 |
| | Fall FAIR | 2.68 | 0.01 | 407.18 |
| | Ordinary least squares | 7.19 | 0.05 | 136.66 |
| 2c | Intercept | 331.71 | 3.33 | 99.71 |
| | Fall FAIR | 2.75 | 0.01 | 426.61 |
| | Average difference | 2.59 | 0.02 | 161.34 |
| 2d | Intercept | 369.34 | 3.34 | 110.57 |
| | Fall FAIR | 2.71 | 0.01 | 414.46 |
| | Simple difference | 1.26 | 0.01 | 146.15 |
| 1 | Intercept | 596.86 | 3.35 | 177.95 |
| | Winter FAIR | 2.25 | 0.01 | 345.43 |
| 2a | Intercept | 719.17 | 3.88 | 185.27 |
| | Winter FAIR | 2.01 | 0.01 | 264.36 |
| | Empirical Bayes | 26.46 | 0.44 | 59.65 |
| 2b | Intercept | 599.02 | 3.35 | 178.65 |
| | Winter FAIR | 2.24 | 0.01 | 344.87 |
| | Ordinary least squares | 0.87 | 0.06 | 15.45 |
| 2c | Intercept | 597.59 | 3.36 | 178.04 |
| | Winter FAIR | 2.25 | 0.01 | 343.90 |
| | Average difference | 0.09 | 0.02 | 5.31 |
| 2d | Intercept | 369.34 | 3.34 | 110.57 |
| | Winter FAIR | 2.71 | 0.01 | 414.46 |
| | Simple difference | –1.45 | 0.01 | –156.35 |
| 1 | Intercept | 573.00 | 3.40 | 168.74 |
| | Spring FAIR | 2.24 | 0.01 | 348.12 |
| 2a | Intercept | 17.50 | 5.85 | 2.99 |
| | Spring FAIR | 3.32 | 0.01 | 294.52 |
| | Empirical Bayes | –75.30 | 0.66 | –113.47 |
| 2b | Intercept | 345.58 | 3.39 | 101.92 |
| | Spring FAIR | 2.68 | 0.01 | 415.40 |
| | Ordinary least squares | –8.64 | 0.06 | –154.08 |
| 2c | Intercept | 331.71 | 3.33 | 99.71 |
| | Spring FAIR | 2.75 | 0.01 | 426.61 |
| | Average difference | –2.92 | 0.02 | –167.77 |
| 2d | Intercept | 568.80 | 3.41 | 166.59 |
| | Spring FAIR | 2.25 | 0.01 | 347.16 |
| | Simple difference | –0.10 | 0.01 | –11.20 |

FAIR is Florida Assessments for Instruction in Reading.

**Note:** All coefficients are significant at $p < .001$.

**Source:** Authors’ analysis based on data from Florida Department of Education (2010).

**Table A5. Summary of hierarchical multiple regressions using different measures of growth to predict end-of-year reading comprehension scores, controlling for status, grade 7, 2009/10**

| Model | Variable | B | Standard error | *t*-value |
|---|---|---|---|---|
| 1 | Intercept | 785.85 | 3.17 | 247.89 |
| | Fall FAIR | 2.02 | 0.01 | 337.28 |
| 2a | Intercept | 837.97 | 2.94 | 285.10 |
| | Fall FAIR | 1.92 | 0.01 | 345.54 |
| | Empirical Bayes | 46.23 | 0.34 | 135.51 |
| 2b | Intercept | 680.95 | 3.08 | 221.11 |
| | Fall FAIR | 2.22 | 0.01 | 381.34 |
| | Ordinary least squares | 5.68 | 0.05 | 122.09 |
| 2c | Intercept | 632.83 | 3.08 | 205.75 |
| | Fall FAIR | 2.29 | 0.01 | 397.15 |
| | Average difference | 2.06 | 0.01 | 143.82 |
| 2d | Intercept | 669.40 | 3.07 | 218.36 |
| | Fall FAIR | 2.24 | 0.01 | 386.80 |
| | Simple difference | 0.99 | 0.01 | 130.06 |
| 1 | Intercept | 869.46 | 3.09 | 281.80 |
| | Winter FAIR | 1.85 | 0.01 | 319.68 |
| 2a | Intercept | 920.04 | 3.34 | 275.86 |
| | Winter FAIR | 1.75 | 0.01 | 279.12 |
| | Empirical Bayes | 15.73 | 0.41 | 38.33 |
| 2b | Intercept | 869.84 | 3.09 | 281.96 |
| | Winter FAIR | 1.85 | 0.01 | 319.60 |
| | Ordinary least squares | 0.38 | 0.05 | 7.63 |
| 2c | Intercept | 869.27 | 3.09 | 281.69 |
| | Winter FAIR | 1.85 | 0.01 | 319.41 |
| | Average difference | –0.05 | 0.02 | –3.19 |
| 2d | Intercept | 669.40 | 3.07 | 218.36 |
| | Winter FAIR | 2.24 | 0.01 | 386.80 |
| | Simple difference | –1.25 | 0.01 | –153.64 |
| 1 | Intercept | 844.16 | 3.15 | 267.61 |
| | Spring FAIR | 1.86 | 0.01 | 320.53 |
| 2a | Intercept | 431.00 | 4.39 | 98.11 |
| | Spring FAIR | 2.64 | 0.01 | 322.19 |
| | Empirical Bayes | –67.25 | 0.53 | –126.21 |
| 2b | Intercept | 650.16 | 3.13 | 207.49 |
| | Spring FAIR | 2.22 | 0.01 | 384.48 |
| | Ordinary least squares | –7.32 | 0.05 | –149.45 |
| 2c | Intercept | 632.83 | 3.08 | 205.75 |
| | Spring FAIR | 2.29 | 0.01 | 397.15 |
| | Average difference | –2.51 | 0.02 | –165.05 |
| 2d | Intercept | 838.71 | 3.16 | 265.05 |
| | Spring FAIR | 1.87 | 0.01 | 321.15 |
| | Simple difference | –0.15 | 0.01 | –17.82 |

FAIR is Florida Assessments for Instruction in Reading.

**Note:** All coefficients are significant at $p < .001$.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Table A6. Summary of hierarchical multiple regressions using different measures of growth to predict end-of-year reading comprehension scores, controlling for status, grade 8, 2009/10**

| Model | Variable | B | Standard error | *t*-value |
|---|---|---|---|---|
| 1 | Intercept | 894.04 | 2.96 | 301.86 |
| | Fall FAIR | 1.83 | 0.01 | 342.67 |
| 2a | Intercept | 971.31 | 2.73 | 355.44 |
| | Fall FAIR | 1.69 | 0.01 | 342.51 |
| | Empirical Bayes | 37.72 | 0.26 | 148.06 |
| 2b | Intercept | 797.46 | 2.83 | 281.55 |
| | Fall FAIR | 2.01 | 0.01 | 392.53 |
| | Ordinary least squares | 5.32 | 0.04 | 131.33 |
| 2c | Intercept | 750.18 | 2.81 | 266.70 |
| | Fall FAIR | 2.07 | 0.01 | 411.17 |
| | Average difference | 1.94 | 0.01 | 155.77 |
| 2d | Intercept | 775.06 | 2.82 | 274.42 |
| | Fall FAIR | 2.04 | 0.01 | 401.27 |
| | Simple difference | 0.96 | 0.01 | 142.94 |
| 1 | Intercept | 954.07 | 2.83 | 336.62 |
| | Winter FAIR | 1.71 | 0.01 | 337.16 |
| 2a | Intercept | 1,023.48 | 3.13 | 327.19 |
| | Winter FAIR | 1.58 | 0.01 | 281.97 |
| | Empirical Bayes | 15.34 | 0.31 | 49.76 |
| 2b | Intercept | 955.49 | 2.83 | 337.35 |
| | Winter FAIR | 1.71 | 0.01 | 336.88 |
| | Ordinary least squares | 0.66 | 0.04 | 15.53 |
| 2c | Intercept | 954.60 | 2.84 | 336.66 |
| | Winter FAIR | 1.71 | 0.01 | 335.95 |
| | Average difference | 0.07 | 0.01 | 5.46 |
| 2d | Intercept | 775.06 | 2.82 | 274.42 |
| | Winter FAIR | 2.04 | 0.01 | 401.27 |
| | Simple difference | –1.08 | 0.01 | –150.14 |
| 1 | Intercept | 950.85 | 2.85 | 333.30 |
| | Spring FAIR | 1.68 | 0.01 | 336.07 |
| 2a | Intercept | 549.38 | 4.29 | 128.10 |
| | Spring FAIR | 2.40 | 0.01 | 315.47 |
| | Empirical Bayes | –50.65 | 0.42 | –119.57 |
| 2b | Intercept | 767.35 | 2.86 | 267.91 |
| | Spring FAIR | 2.01 | 0.01 | 398.61 |
| | Ordinary least squares | –6.38 | 0.04 | –148.26 |
| 2c | Intercept | 750.18 | 2.81 | 266.70 |
| | Spring FAIR | 2.07 | 0.01 | 411.17 |
| | Average difference | –2.20 | 0.01 | –163.95 |
| 2d | Intercept | 947.92 | 2.86 | 330.97 |
| | Spring FAIR | 1.69 | 0.01 | 335.43 |
| | Simple difference | –0.08 | 0.01 | –10.80 |

FAIR is Florida Assessments for Instruction in Reading.

**Note:** All coefficients are significant at $p < .001$.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Table A7. Summary of hierarchical multiple regressions using different measures of growth to predict end-of-year reading comprehension scores, controlling for status, grade 9, 2009/10**

| Model | Variable | B | Standard error | *t*-value |
|---|---|---|---|---|
| 1 | Intercept | 809.48 | 3.52 | 229.75 |
| | Fall FAIR | 2.02 | 0.01 | 327.93 |
| 2a | Intercept | 872.46 | 3.30 | 264.33 |
| | Fall FAIR | 1.91 | 0.01 | 330.58 |
| | Empirical Bayes | 37.52 | 0.29 | 128.50 |
| 2b | Intercept | 705.26 | 3.43 | 205.29 |
| | Fall FAIR | 2.21 | 0.01 | 366.71 |
| | Ordinary least squares | 5.25 | 0.05 | 114.53 |
| 2c | Intercept | 656.79 | 3.41 | 192.48 |
| | Fall FAIR | 2.28 | 0.01 | 383.10 |
| | Average difference | 1.99 | 0.01 | 138.21 |
| 2d | Intercept | 689.53 | 3.41 | 202.19 |
| | Fall FAIR | 2.23 | 0.01 | 373.96 |
| | Simple difference | 0.98 | 0.01 | 125.70 |
| 1 | Intercept | 913.17 | 3.38 | 269.95 |
| | Winter FAIR | 1.83 | 0.01 | 311.10 |
| 2a | Intercept | 973.04 | 3.66 | 266.24 |
| | Winter FAIR | 1.73 | 0.01 | 270.58 |
| | Empirical Bayes | 14.30 | 0.35 | 41.26 |
| 2b | Intercept | 913.67 | 3.38 | 270.24 |
| | Winter FAIR | 1.83 | 0.01 | 311.11 |
| | Ordinary least squares | 0.54 | 0.05 | 11.12 |
| 2c | Intercept | 913.31 | 3.39 | 269.84 |
| | Winter FAIR | 1.83 | 0.01 | 310.57 |
| | Average difference | 0.02 | 0.02 | 1.23 |
| 2d | Intercept | 689.53 | 3.41 | 202.19 |
| | Winter FAIR | 2.23 | 0.01 | 373.96 |
| | Simple difference | –1.25 | 0.01 | –148.76 |
| 1 | Intercept | 912.69 | 3.43 | 266.30 |
| | Spring FAIR | 1.81 | 0.01 | 307.13 |
| 2a | Intercept | 469.27 | 4.91 | 95.68 |
| | Spring FAIR | 2.58 | 0.01 | 303.64 |
| | Empirical Bayes | –55.57 | 0.47 | –119.46 |
| 2b | Intercept | 696.63 | 3.45 | 201.89 |
| | Spring FAIR | 2.19 | 0.01 | 367.59 |
| | Ordinary least squares | –7.12 | 0.05 | –145.38 |
| 2c | Intercept | 656.79 | 3.41 | 192.48 |
| | Spring FAIR | 2.28 | 0.01 | 383.10 |
| | Average difference | –2.57 | 0.02 | –164.91 |
| 2d | Intercept | 907.57 | 3.44 | 263.63 |
| | Spring FAIR | 1.82 | 0.01 | 307.15 |
| | Simple difference | –0.12 | 0.01 | –14.24 |

FAIR is Florida Assessments for Instruction in Reading.

**Note:** All coefficients are significant at $p < .001$.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Table A8. Summary of hierarchical multiple regressions using different measures of growth to predict end-of-year reading comprehension scores, controlling for status, grade 10, 2009/10**

| Model | Variable | B | Standard error | *t*-value |
|---|---|---|---|---|
| 1 | Intercept | 425.84 | 4.49 | 94.82 |
| | Fall FAIR | 2.63 | 0.01 | 348.25 |
| 2a | Intercept | 525.60 | 4.24 | 124.02 |
| | Fall FAIR | 2.46 | 0.01 | 345.05 |
| | Empirical Bayes | 46.63 | 0.37 | 127.48 |
| 2b | Intercept | 295.73 | 4.39 | 67.34 |
| | Fall FAIR | 2.85 | 0.01 | 385.79 |
| | Ordinary least squares | 6.25 | 0.06 | 111.73 |
| 2c | Intercept | 235.93 | 4.34 | 54.34 |
| | Fall FAIR | 2.94 | 0.01 | 403.92 |
| | Average difference | 2.45 | 0.02 | 137.81 |
| 2d | Intercept | 272.69 | 4.32 | 63.12 |
| | Fall FAIR | 2.89 | 0.01 | 397.57 |
| | Simple difference | 1.23 | 0.01 | 129.61 |
| 1 | Intercept | 590.82 | 4.34 | 136.17 |
| | Winter FAIR | 2.35 | 0.01 | 322.62 |
| 2a | Intercept | 666.45 | 4.78 | 139.34 |
| | Winter FAIR | 2.22 | 0.01 | 275.87 |
| | Empirical Bayes | 16.27 | 0.45 | 36.48 |
| 2b | Intercept | 590.95 | 4.34 | 136.18 |
| | Winter FAIR | 2.35 | 0.01 | 322.53 |
| | Ordinary least squares | 0.09 | 0.06 | 1.53 |
| 2c | Intercept | 589.00 | 4.34 | 135.64 |
| | Winter FAIR | 2.35 | 0.01 | 322.67 |
| | Average difference | –0.16 | 0.02 | –8.59 |
| 2d | Intercept | 272.69 | 4.32 | 63.12 |
| | Winter FAIR | 2.89 | 0.01 | 397.57 |
| | Simple difference | –1.66 | 0.01 | –163.22 |
| 1 | Intercept | 593.39 | 4.44 | 133.72 |
| | Spring FAIR | 2.32 | 0.01 | 314.76 |
| 2a | Intercept | –18.50 | 6.50 | –2.85 |
| | Spring FAIR | 3.35 | 0.01 | 307.99 |
| | Empirical Bayes | –73.71 | 0.60 | –122.24 |
| 2b | Intercept | 304.75 | 4.39 | 69.44 |
| | Spring FAIR | 2.81 | 0.01 | 384.00 |
| | Ordinary least squares | –9.33 | 0.06 | –156.02 |
| 2c | Intercept | 235.93 | 4.34 | 54.34 |
| | Spring FAIR | 2.94 | 0.01 | 403.92 |
| | Average difference | –3.43 | 0.02 | –179.63 |
| 2d | Intercept | 583.54 | 4.46 | 130.85 |
| | Spring FAIR | 2.34 | 0.01 | 315.43 |
| | Simple difference | –0.20 | 0.01 | –19.11 |

FAIR is Florida Assessments for Instruction in Reading.

**Note:** All coefficients are significant at $p < .001$.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Table A9. Comparison of $\Delta R^2$ between growth estimates when controlling for the fall FAIR, by grade, 2009/10**

| Growth comparison | | Growth comparison | |
|---|---|---|---|
| *Grade 3* | | *Grade 7* | |
| Empirical Bayes – ordinary least squares | .00 | Empirical Bayes – ordinary least squares | .01 |
| Empirical Bayes – average difference | –.01 | Empirical Bayes – average difference | –.01 |
| Empirical Bayes – simple difference | .01 | Empirical Bayes – simple difference | .00 |
| Ordinary least squares – average difference | –.01 | Ordinary least squares – average difference | –.02 |
| Ordinary least squares – simple difference | .01 | Ordinary least squares – simple difference | –.01 |
| Average difference – simple difference | .02 | Average difference – simple difference | .01 |
| *Grade 4* | | *Grade 8* | |
| Empirical Bayes – ordinary least squares | .00 | Empirical Bayes – ordinary least squares | .01 |
| Empirical Bayes – average difference | –.02 | Empirical Bayes – average difference | –.01 |
| Empirical Bayes – simple difference | .00 | Empirical Bayes – simple difference | .00 |
| Ordinary least squares – average difference | –.02 | Ordinary least squares – average difference | –.02 |
| Ordinary least squares – simple difference | .00 | Ordinary least squares – simple difference | –.01 |
| Average difference – simple difference | .02 | Average difference – simple difference | .01 |
| *Grade 5* | | *Grade 9* | |
| Empirical Bayes – ordinary least squares | .00 | Empirical Bayes – ordinary least squares | .01 |
| Empirical Bayes – average difference | –.02 | Empirical Bayes – average difference | –.01 |
| Empirical Bayes – simple difference | –.01 | Empirical Bayes – simple difference | .00 |
| Ordinary least squares – average difference | –.02 | Ordinary least squares – average difference | –.02 |
| Ordinary least squares – simple difference | –.01 | Ordinary least squares – simple difference | –.01 |
| Average difference – simple difference | .01 | Average difference – simple difference | .01 |
| *Grade 6* | | *Grade 10* | |
| Empirical Bayes – ordinary least squares | .02 | Empirical Bayes – ordinary least squares | .01 |
| Empirical Bayes – average difference | .00 | Empirical Bayes – average difference | –.01 |
| Empirical Bayes – simple difference | .01 | Empirical Bayes – simple difference | –.01 |
| Ordinary least squares – average difference | –.02 | Ordinary least squares – average difference | –.02 |
| Ordinary least squares – simple difference | –.01 | Ordinary least squares – simple difference | –.02 |
| Average difference – simple difference | .01 | Average difference – simple difference | .00 |

FAIR is Florida Assessments for Instruction in Reading.

**Source:** Authors’ analysis based on data from Florida Department of Education (2010).

**Table A10. Comparison of $\Delta R^2$ between growth estimates when controlling for the winter FAIR, by grade, 2009/10**

| Growth comparison | | Growth comparison | |
|---|---|---|---|
| *Grade 3* | | *Grade 7* | |
| Empirical Bayes – ordinary least squares | .01 | Empirical Bayes – ordinary least squares | .01 |
| Empirical Bayes – average difference | .01 | Empirical Bayes – average difference | .01 |
| Empirical Bayes – simple difference | –.06 | Empirical Bayes – simple difference | –.08 |
| Ordinary least squares – average difference | .00 | Ordinary least squares – average difference | .00 |
| Ordinary least squares – simple difference | –.07 | Ordinary least squares – simple difference | –.09 |
| Average difference – simple difference | –.07 | Average difference – simple difference | –.09 |
| *Grade 4* | | *Grade 8* | |
| Empirical Bayes – ordinary least squares | .00 | Empirical Bayes – ordinary least squares | .01 |
| Empirical Bayes – average difference | .00 | Empirical Bayes – average difference | .01 |
| Empirical Bayes – simple difference | –.06 | Empirical Bayes – simple difference | –.08 |
| Ordinary least squares – average difference | .00 | Ordinary least squares – average difference | .00 |
| Ordinary least squares – simple difference | –.06 | Ordinary least squares – simple difference | –.09 |
| Average difference – simple difference | –.06 | Average difference – simple difference | –.09 |
| *Grade 5* | | *Grade 9* | |
| Empirical Bayes – ordinary least squares | .00 | Empirical Bayes – ordinary least squares | .01 |
| Empirical Bayes – average difference | .00 | Empirical Bayes – average difference | .01 |
| Empirical Bayes – simple difference | –.07 | Empirical Bayes – simple difference | –.08 |
| Ordinary least squares – average difference | .00 | Ordinary least squares – average difference | .00 |
| Ordinary least squares – simple difference | –.07 | Ordinary least squares – simple difference | –.09 |
| Average difference – simple difference | –.07 | Average difference – simple difference | –.09 |
| *Grade 6* | | *Grade 10* | |
| Empirical Bayes – ordinary least squares | .02 | Empirical Bayes – ordinary least squares | .01 |
| Empirical Bayes – average difference | .02 | Empirical Bayes – average difference | .01 |
| Empirical Bayes – simple difference | –.07 | Empirical Bayes – simple difference | –.09 |
| Ordinary least squares – average difference | .00 | Ordinary least squares – average difference | .00 |
| Ordinary least squares – simple difference | –.09 | Ordinary least squares – simple difference | –.10 |
| Average difference – simple difference | –.09 | Average difference – simple difference | –.10 |

FAIR is Florida Assessments for Instruction in Reading.

**Source:** Authors’ analysis based on data from Florida Department of Education (2010).

**Table A11. Comparison of $\Delta R^2$ between growth estimates when controlling for the spring FAIR, by grade, 2009/10**

| Growth comparison | | Growth comparison | |
|---|---|---|---|
| *Grade 3* | | *Grade 7* | |
| Empirical Bayes – ordinary least squares | –.01 | Empirical Bayes – ordinary least squares | –.02 |
| Empirical Bayes – average difference | .00 | Empirical Bayes – average difference | –.04 |
| Empirical Bayes – simple difference | .07 | Empirical Bayes – simple difference | .07 |
| Ordinary least squares – average difference | .01 | Ordinary least squares – average difference | –.02 |
| Ordinary least squares – simple difference | .08 | Ordinary least squares – simple difference | .09 |
| Average difference – simple difference | .07 | Average difference – simple difference | .11 |
| *Grade 4* | | *Grade 8* | |
| Empirical Bayes – ordinary least squares | .00 | Empirical Bayes – ordinary least squares | –.03 |
| Empirical Bayes – average difference | .00 | Empirical Bayes – average difference | –.04 |
| Empirical Bayes – simple difference | .07 | Empirical Bayes – simple difference | .06 |
| Ordinary least squares – average difference | .00 | Ordinary least squares – average difference | –.01 |
| Ordinary least squares – simple difference | .07 | Ordinary least squares – simple difference | .09 |
| Average difference – simple difference | .07 | Average difference – simple difference | .10 |
| *Grade 5* | | *Grade 9* | |
| Empirical Bayes – ordinary least squares | .00 | Empirical Bayes – ordinary least squares | –.03 |
| Empirical Bayes – average difference | .00 | Empirical Bayes – average difference | –.05 |
| Empirical Bayes – simple difference | .08 | Empirical Bayes – simple difference | .06 |
| Ordinary least squares – average difference | .00 | Ordinary least squares – average difference | –.02 |
| Ordinary least squares – simple difference | .08 | Ordinary least squares – simple difference | .09 |
| Average difference – simple difference | .08 | Average difference – simple difference | .11 |
| *Grade 6* | | *Grade 10* | |
| Empirical Bayes – ordinary least squares | –.04 | Empirical Bayes – ordinary least squares | –.03 |
| Empirical Bayes – average difference | –.05 | Empirical Bayes – average difference | –.05 |
| Empirical Bayes – simple difference | .05 | Empirical Bayes – simple difference | .07 |
| Ordinary least squares – average difference | –.01 | Ordinary least squares – average difference | –.02 |
| Ordinary least squares – simple difference | .09 | Ordinary least squares – simple difference | .10 |
| Average difference – simple difference | .10 | Average difference – simple difference | .12 |

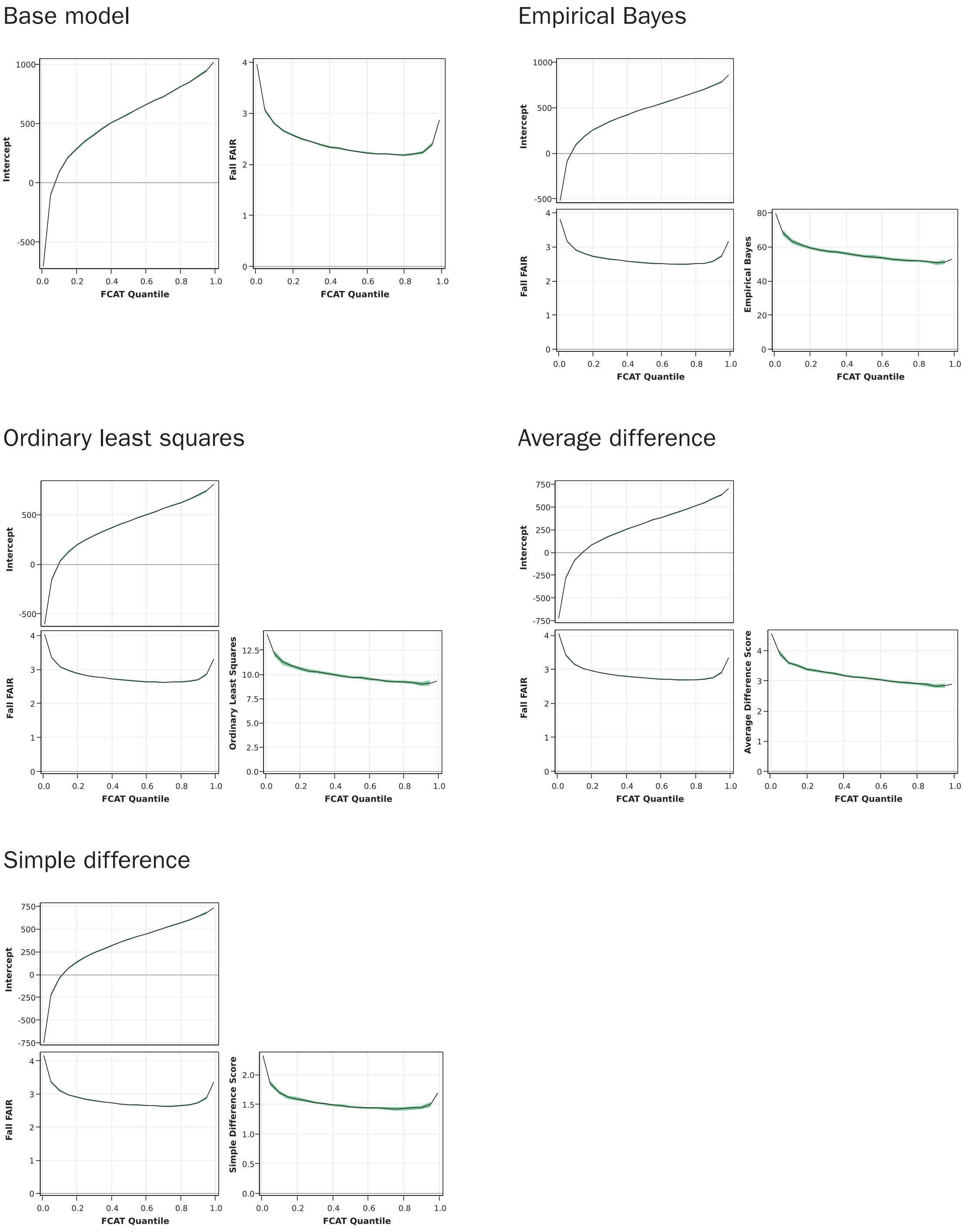
FAIR is Florida Assessments for Instruction in Reading.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

# Appendix B. Unstandardized multiple quantile regression process plots centering time at the initial (fall) status on the Florida Assessments for Instruction in Reading

This appendix contains quantile process plots that depict the relation between growth and reading comprehension performance conditional on student reading comprehension performance after controlling for initial (fall) status. Three graphs are included in each reported process plot because the models have two predictors as well as an intercept; the base model plot, which includes only one predictor (status) plus an intercept, has just two graphs. The intercept portion of the process plot displays the predicted reading comprehension score (y-axis) across the distribution of Florida Comprehensive Assessment Test (FCAT) scores when fall status and growth measure are 0 (x-axis). This plot shows that students at lower quantiles of FCAT performance have lower reading comprehension scores. The remaining plots reflect the slope coefficients for status and growth conditional on reading comprehension performance. Most pertinent is the growth plot, which shows that the coefficient decreases as FCAT performance increases.

**Figure B1. Grade 3: unstandardized multiple quantile regression process plots centering time at the fall FAIR: estimated parameter by quantile for the FCAT, 2009/10**


Base model

Empirical Bayes

Ordinary least squares

Average difference

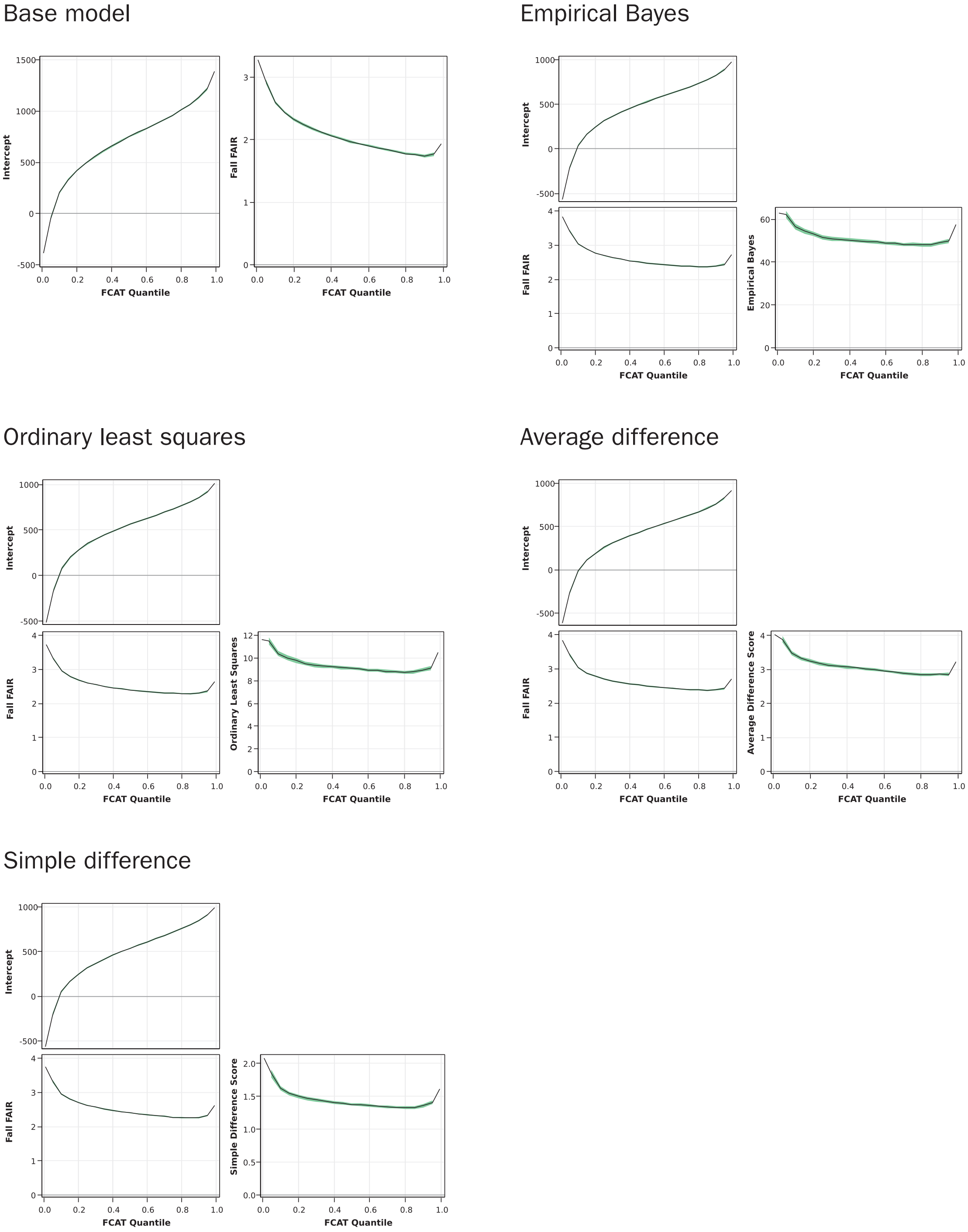
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

### Figure B2. Grade 4: unstandardized multiple quantile regression process plots centering time at the fall FAIR: estimated parameter by quantile for the FCAT, 2009/10



Base model

Empirical Bayes

Ordinary least squares

Average difference

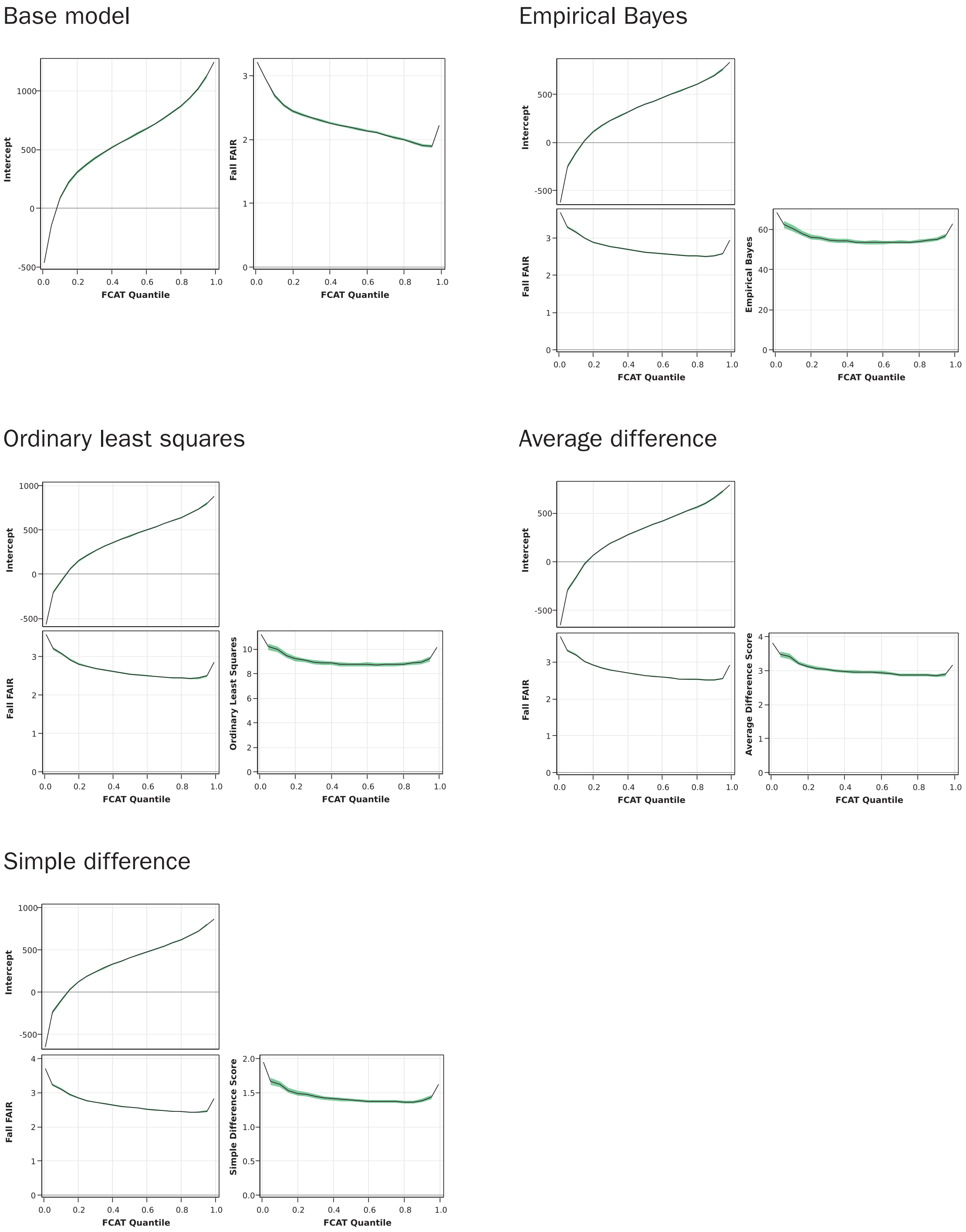
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

### Figure B3. Grade 5: unstandardized multiple quantile regression process plots centering time at the fall FAIR: estimated parameter by quantile for the FCAT, 2009/10


Base model

Empirical Bayes

Ordinary least squares

Average difference

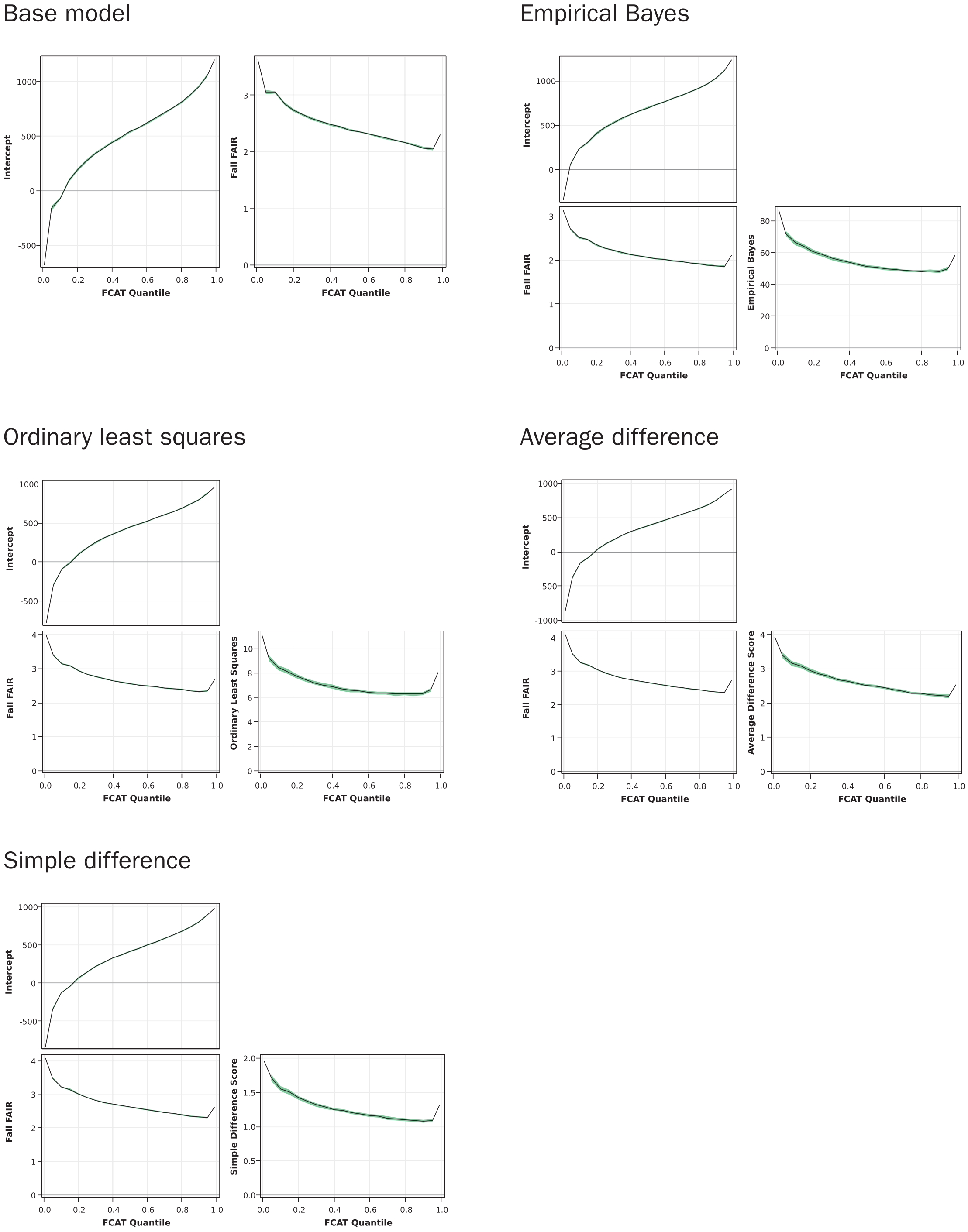
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

## Figure B4. Grade 6: unstandardized multiple quantile regression process plots centering time at the fall FAIR: estimated parameter by quantile for the FCAT, 2009/10



Base model

Empirical Bayes

Ordinary least squares

Average difference

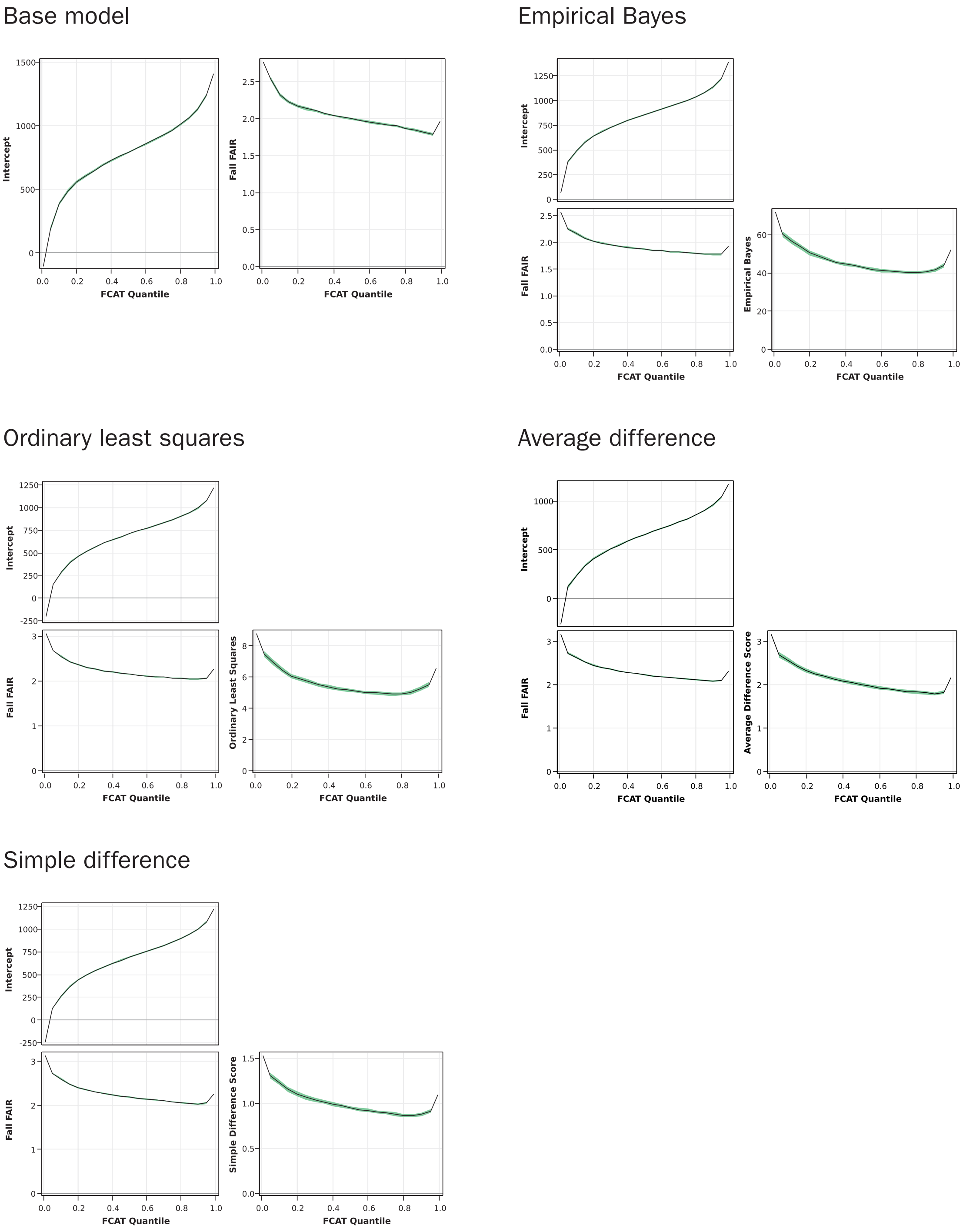
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

## Figure B5. Grade 7: unstandardized multiple quantile regression process plots centering time at the fall FAIR: estimated parameter by quantile for the FCAT, 2009/10


Base model

Empirical Bayes

Ordinary least squares

Average difference

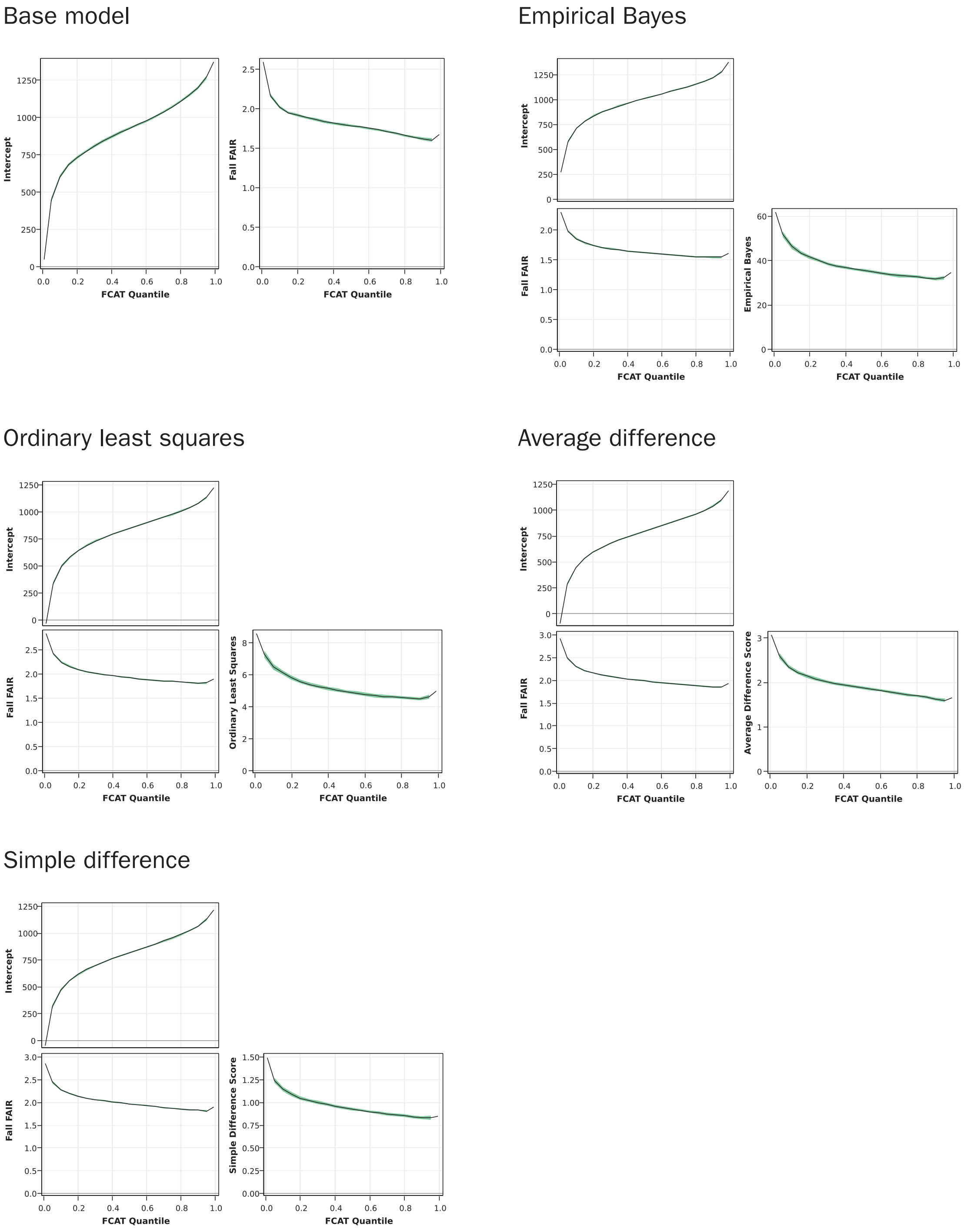
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

## Figure B6. Grade 8: unstandardized multiple quantile regression process plots centering time at the fall FAIR: estimated parameter by quantile for the FCAT, 2009/10

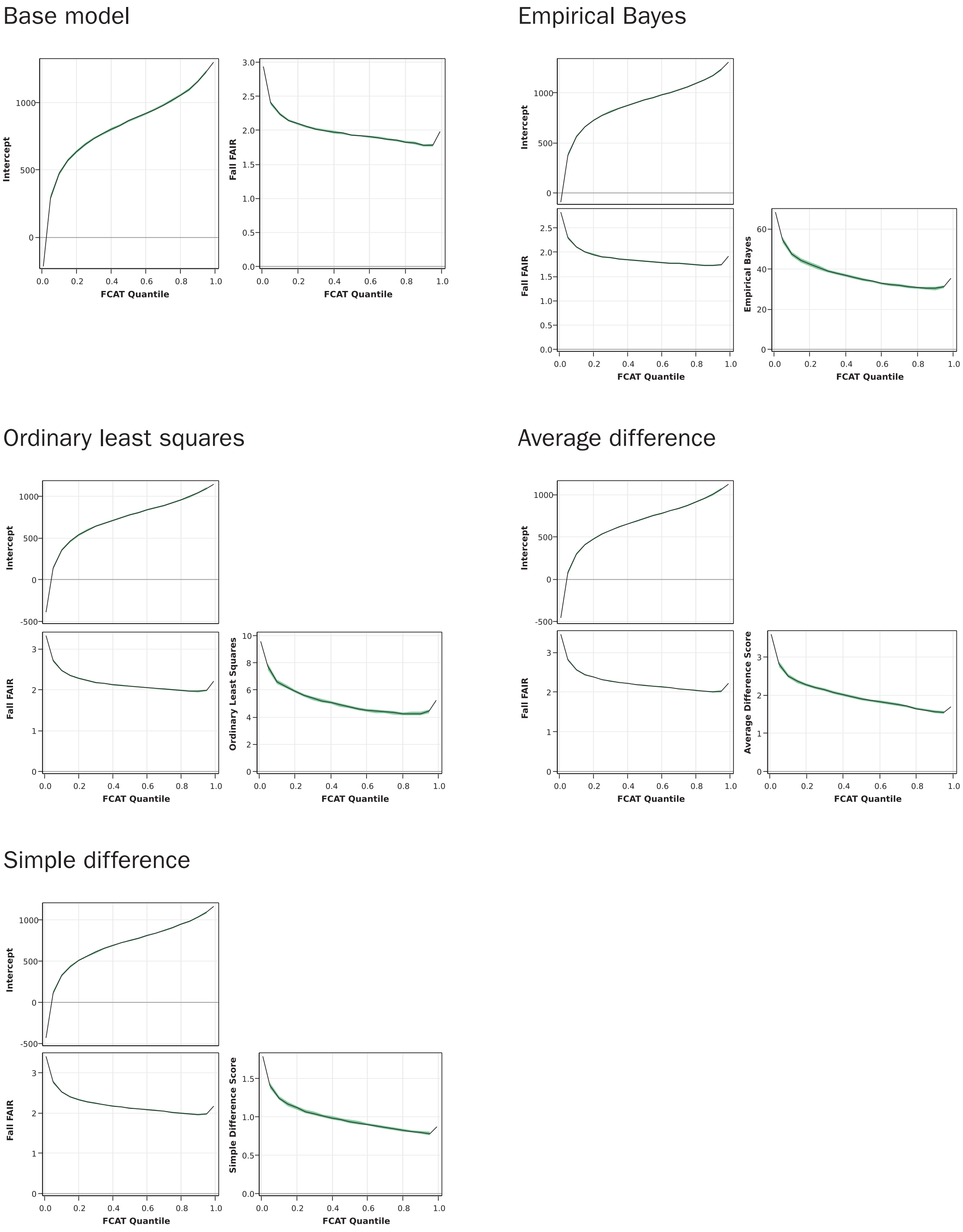


FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Figure B7. Grade 9: unstandardized multiple quantile regression process plots centering time at the fall FAIR: estimated parameter by quantile for the FCAT, 2009/10**



Base model

Empirical Bayes

Ordinary least squares

Average difference

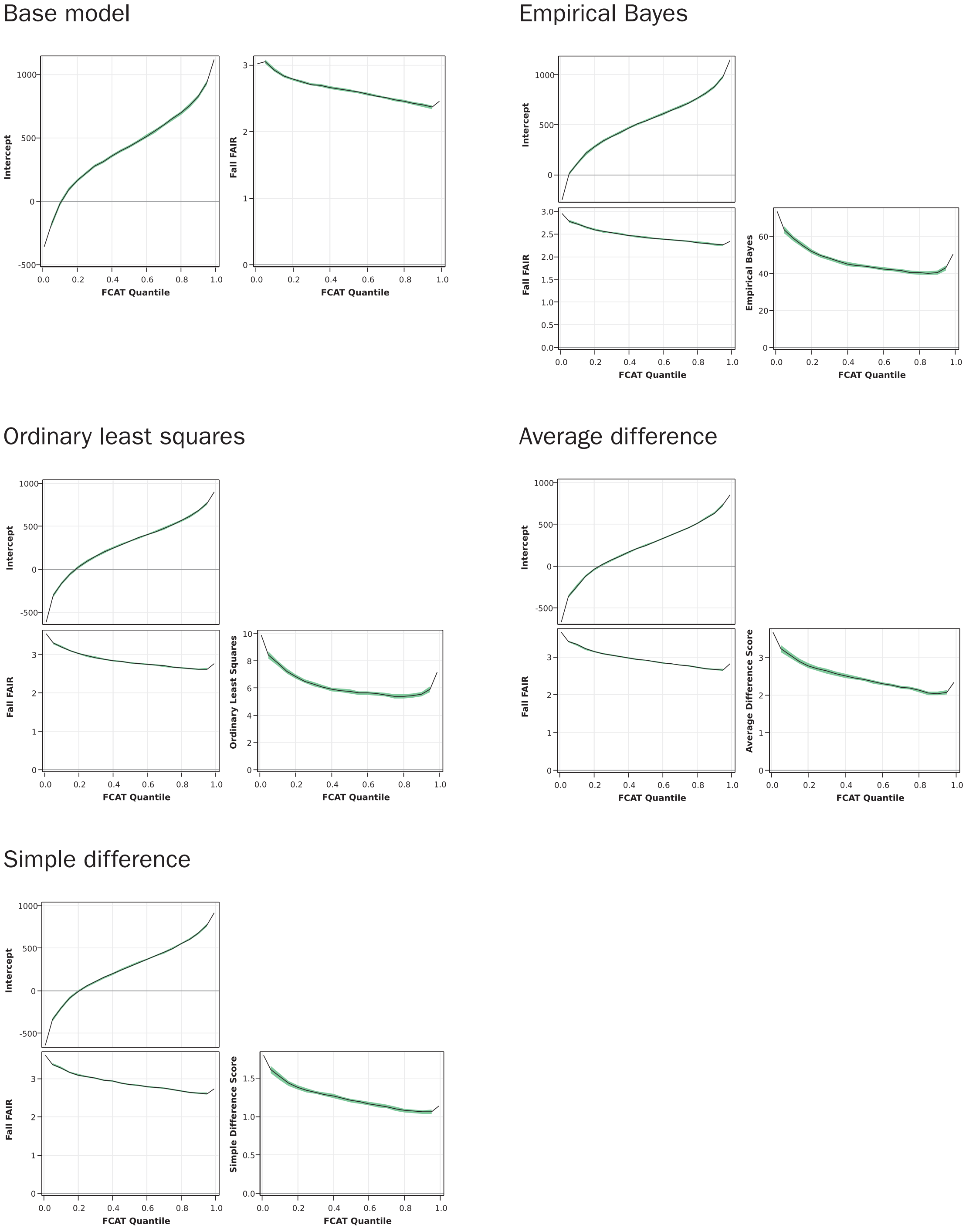
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Figure B8. Grade 10: unstandardized multiple quantile regression process plots centering time at the fall FAIR: estimated parameter by quantile for the FCAT, 2009/10**



Base model

Empirical Bayes

Ordinary least squares

Average difference

Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

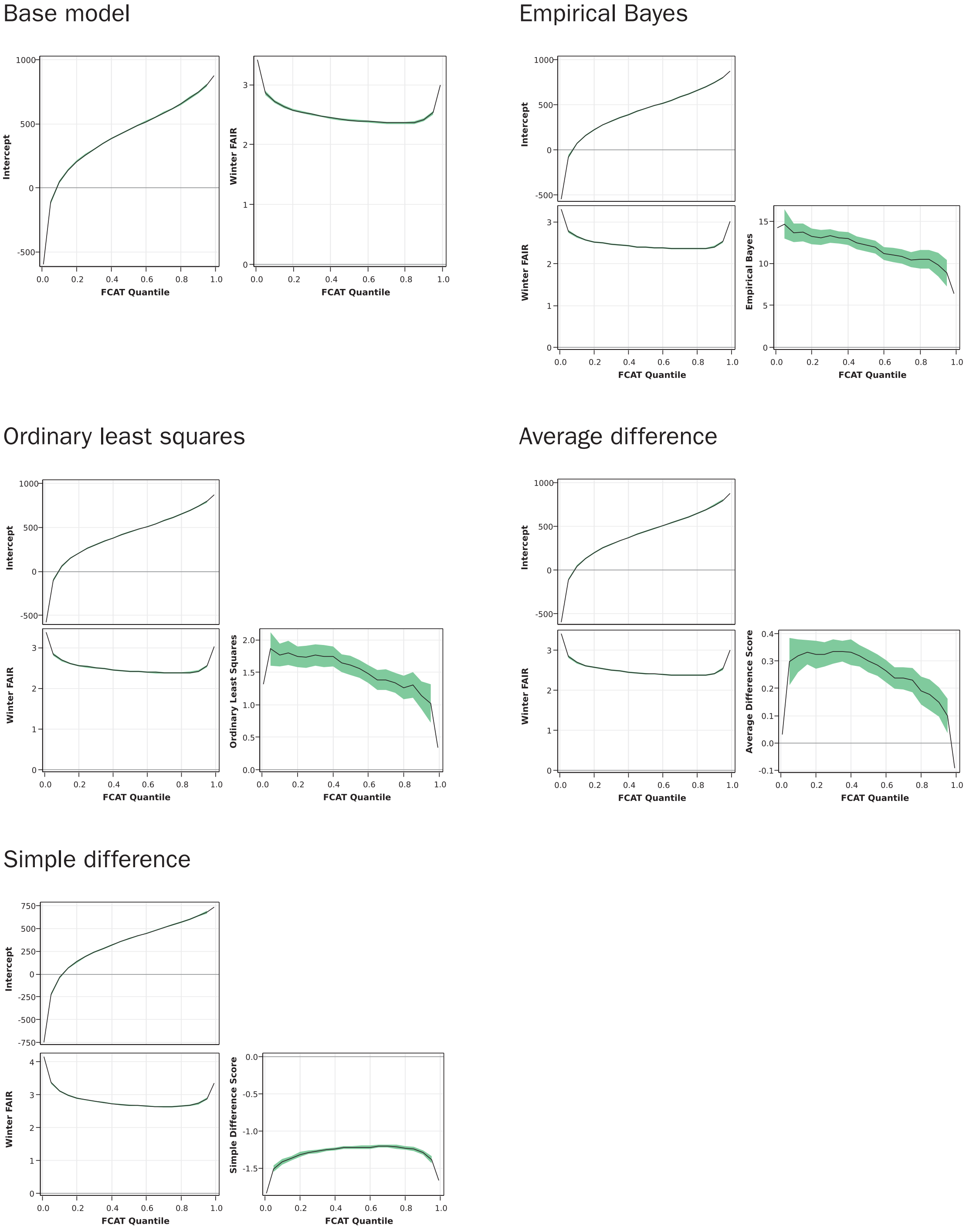
**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

# Appendix C. Unstandardized multiple quantile regression process plots centering time at the mid-year (winter) status on the Florida Assessments for Instruction in Reading

This appendix contains quantile process plots that depict the relation between growth and reading comprehension performance conditional on student reading comprehension performance after controlling for initial (fall) status. Three graphs are included in each reported process plot because the models have two predictors as well as an intercept; the base model plot, which includes only one predictor (status) plus an intercept, has just two graphs. The intercept portion of the process plot displays the predicted reading comprehension score (y-axis) across the distribution of Florida Comprehensive Assessment Test (FCAT) scores when winter status and growth measure are 0 (x-axis). This plot shows that students at lower quantiles of FCAT performance have lower reading comprehension scores. The remaining plots reflect the slope coefficients for status and growth conditional on reading comprehension performance. Most pertinent is the growth plot, which shows that the coefficient decreases as FCAT performance increases.

## Figure C1. Grade 3: unstandardized multiple quantile regression process plots centering time at the winter FAIR: estimated parameter by quantile for the FCAT, 2009/10



Base model

Empirical Bayes

Ordinary least squares

Average difference

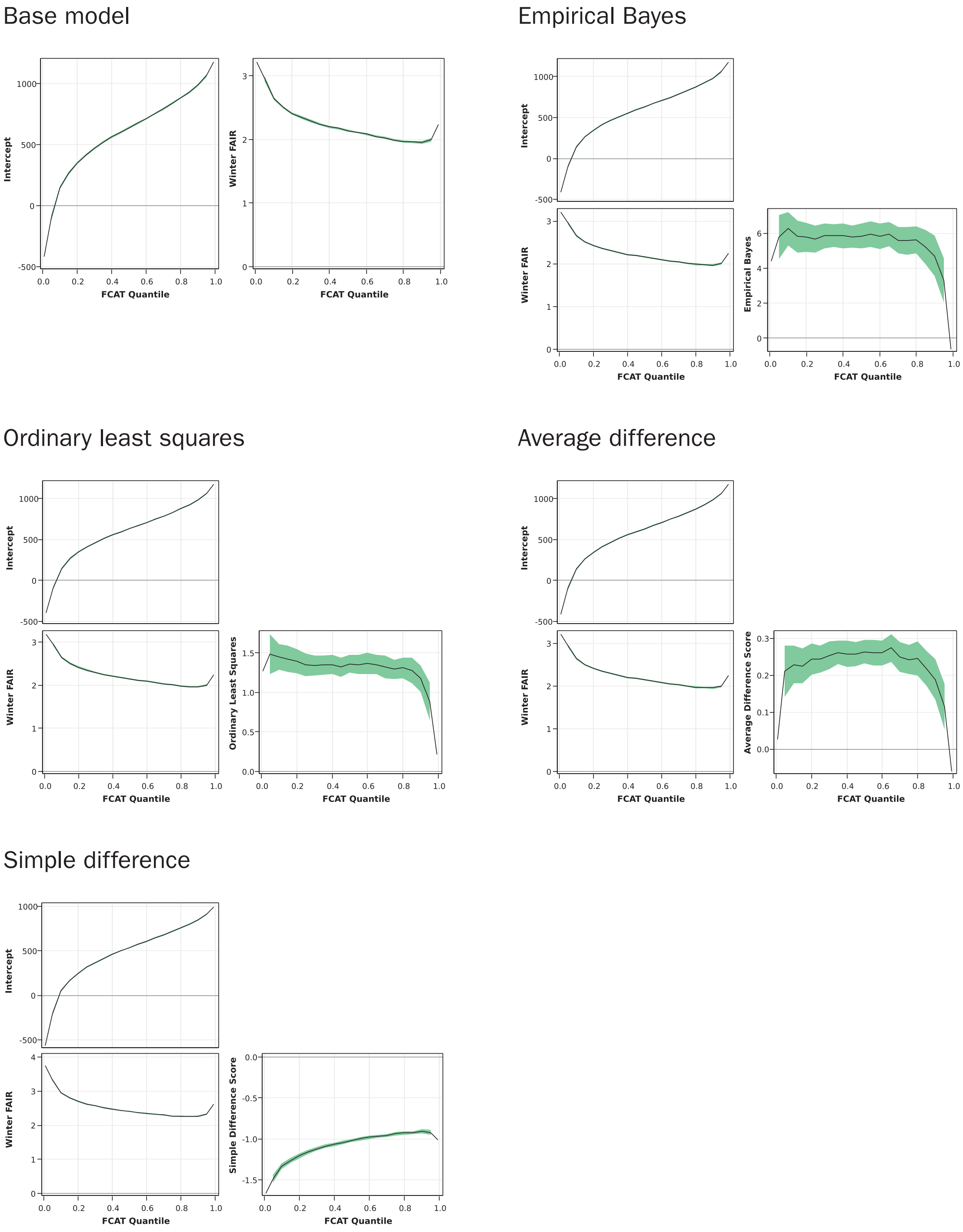
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Figure C2. Grade 4: unstandardized multiple quantile regression process plots centering time at the winter FAIR: estimated parameter by quantile for the FCAT, 2009/10**



Base model

Empirical Bayes

Ordinary least squares

Average difference

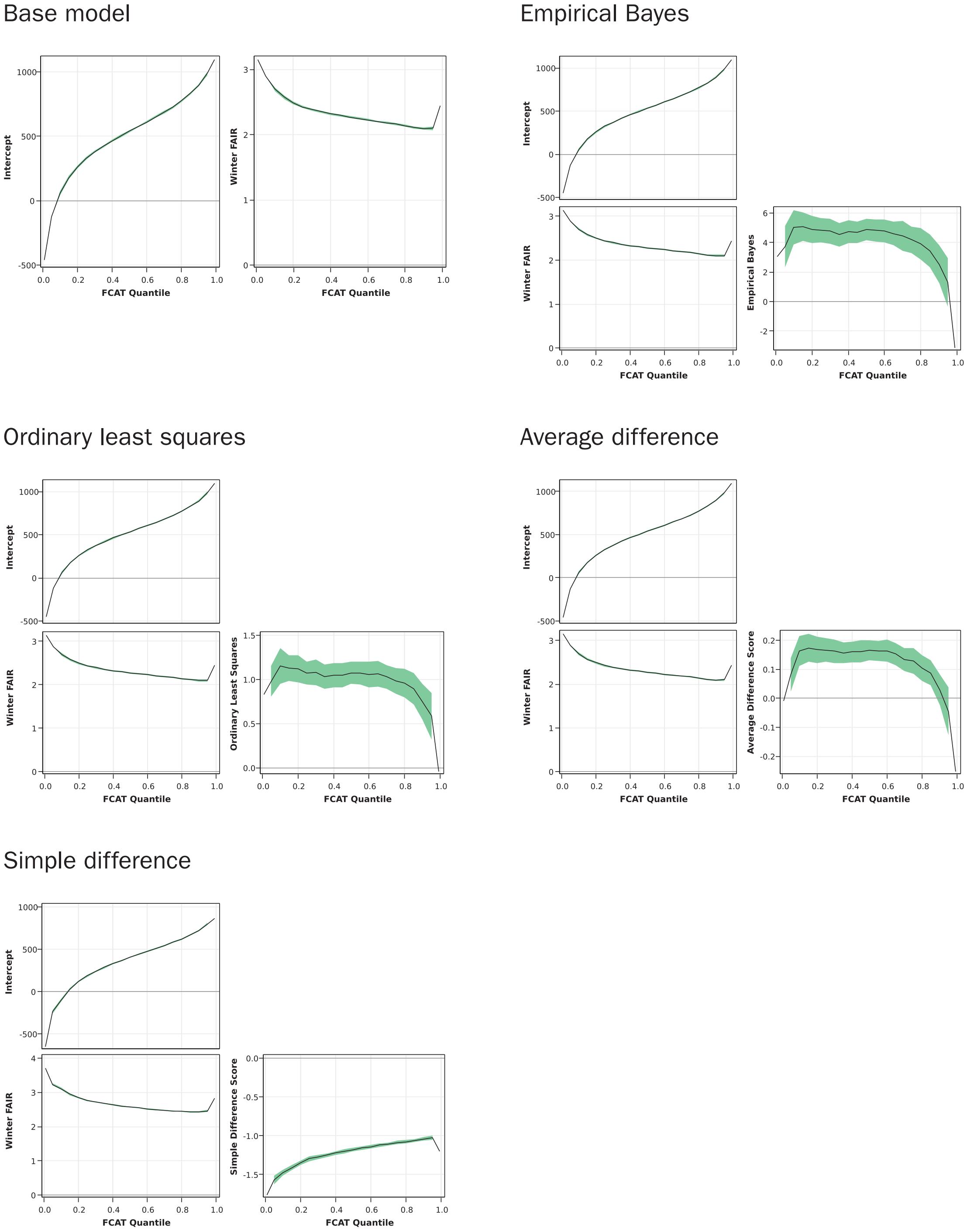
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

## Figure C3. Grade 5: unstandardized multiple quantile regression process plots centering time at the winter FAIR: estimated parameter by quantile for the FCAT, 2009/10



Base model

Empirical Bayes

Ordinary least squares

Average difference

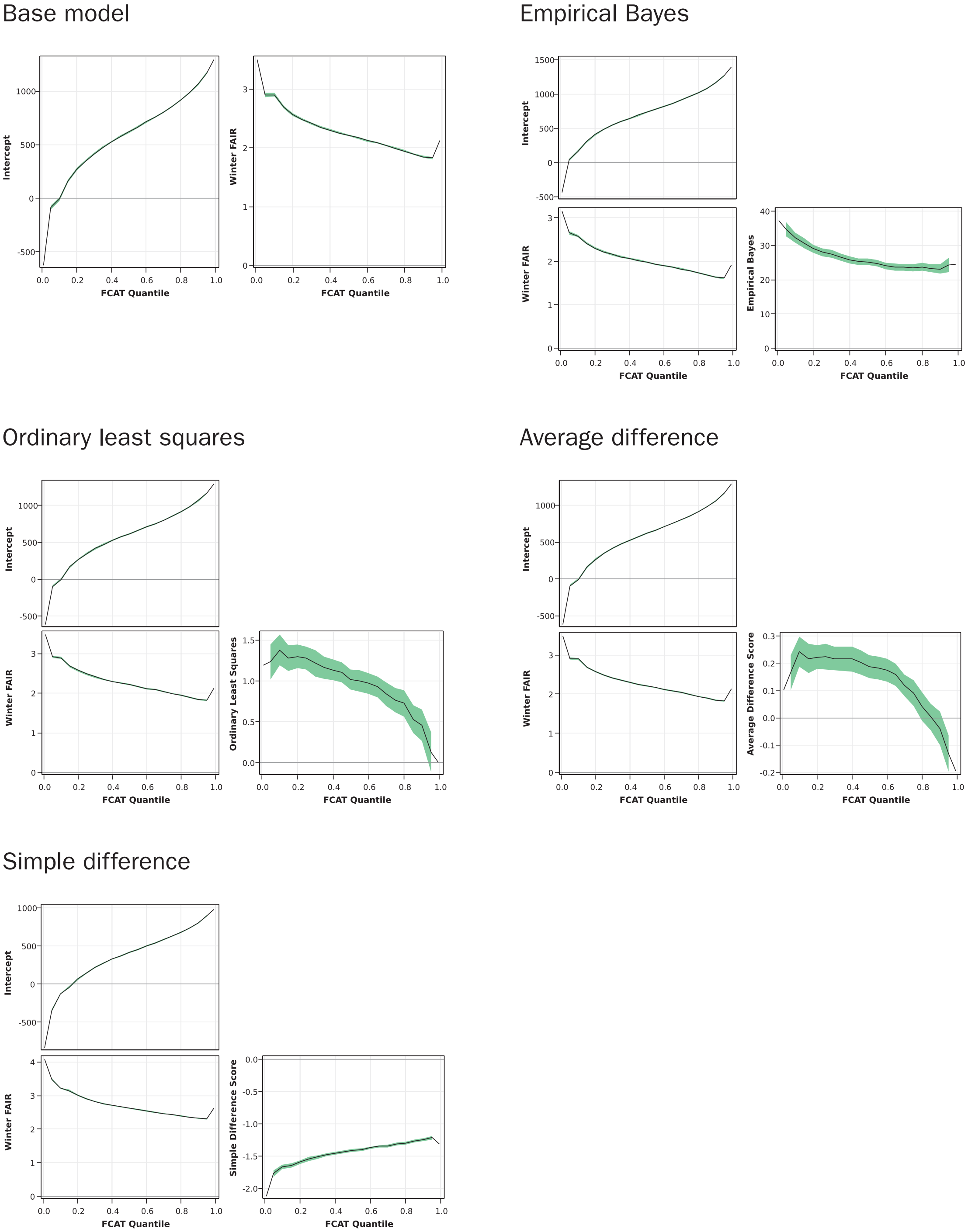
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

## Figure C4. Grade 6: unstandardized multiple quantile regression process plots centering time at the winter FAIR: estimated parameter by quantile for the FCAT, 2009/10



Base model

Empirical Bayes

Ordinary least squares

Average difference

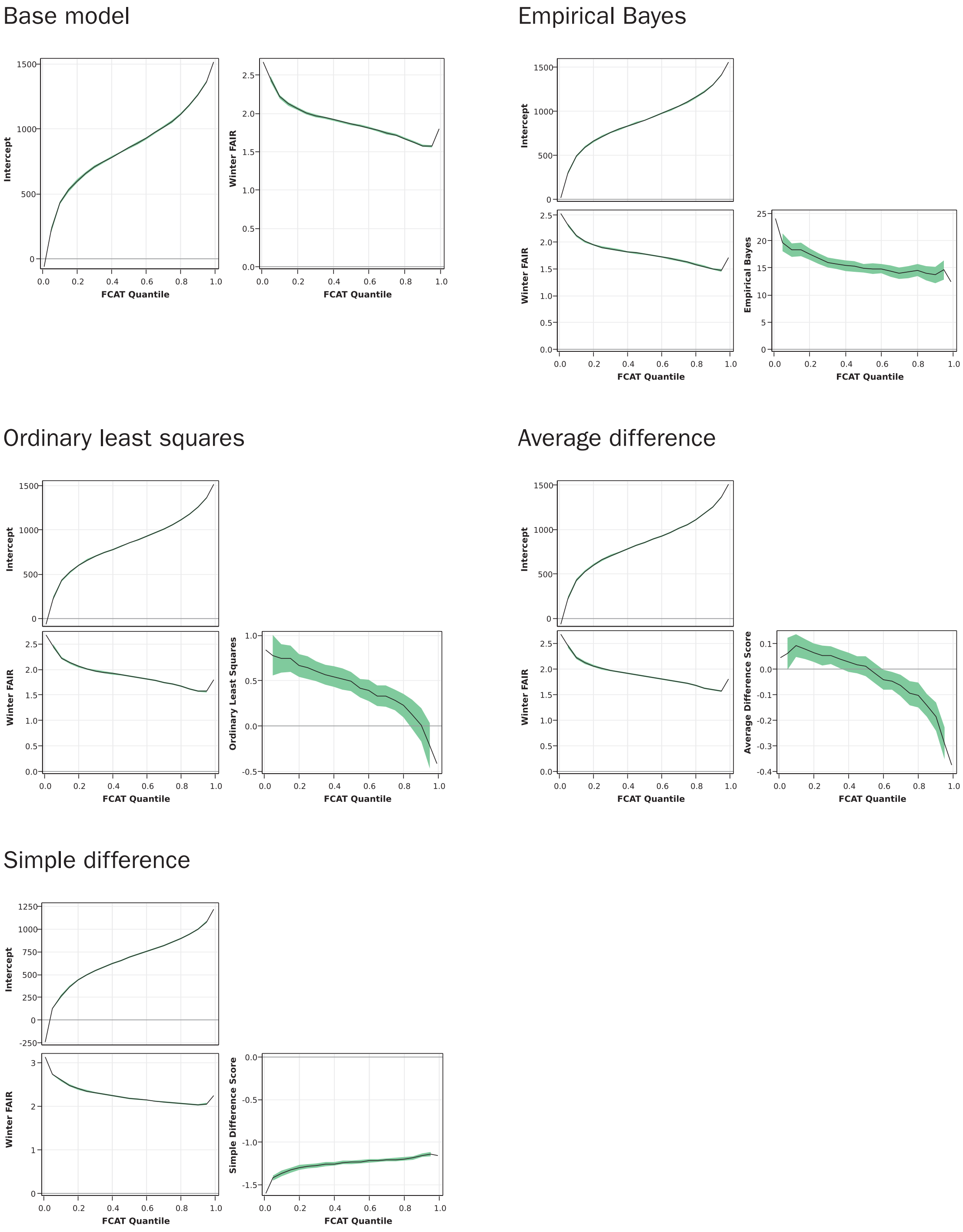
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors’ analysis based on data from Florida Department of Education (2010).

**Figure C5. Grade 7: unstandardized multiple quantile regression process plots centering time at the winter FAIR: estimated parameter by quantile for the FCAT, 2009/10**


Base model

Empirical Bayes

Ordinary least squares

Average difference

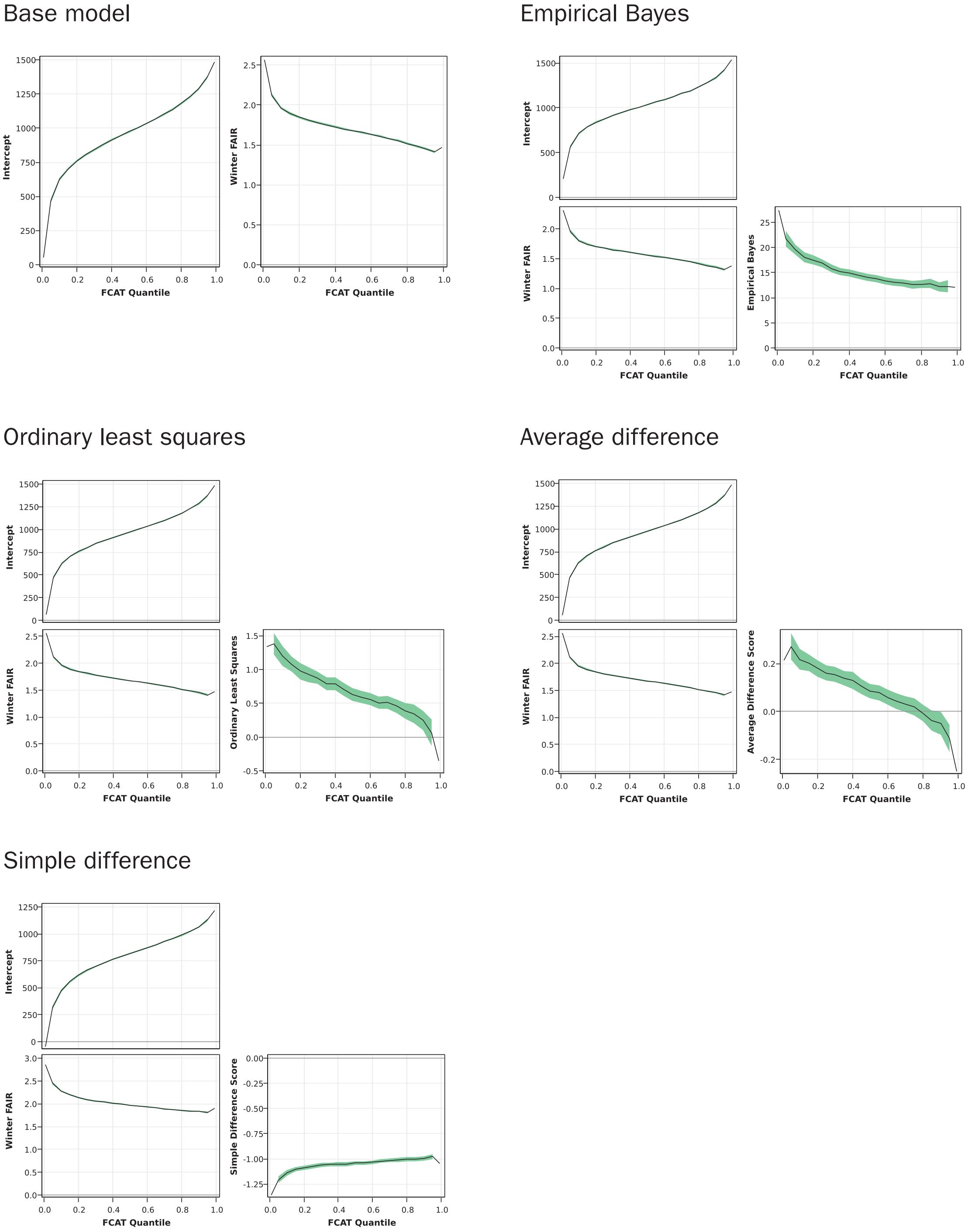
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

## Figure C6. Grade 8: unstandardized multiple quantile regression process plots centering time at the winter FAIR: estimated parameter by quantile for the FCAT, 2009/10



Base model

Empirical Bayes

Ordinary least squares

Average difference

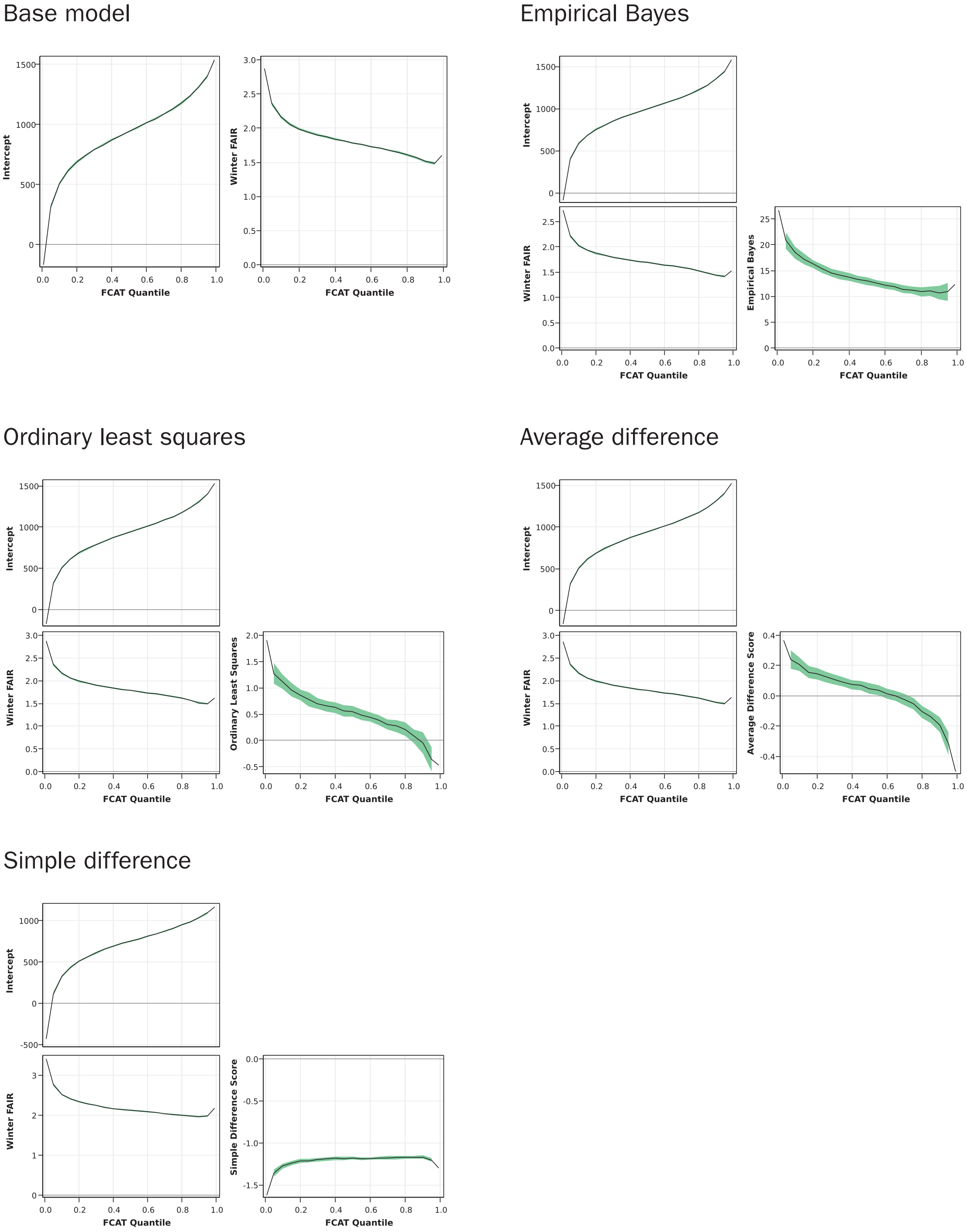
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Figure C7. Grade 9: unstandardized multiple quantile regression process plots centering time at the winter FAIR: estimated parameter by quantile for the FCAT, 2009/10**



Base model

Empirical Bayes

Ordinary least squares

Average difference

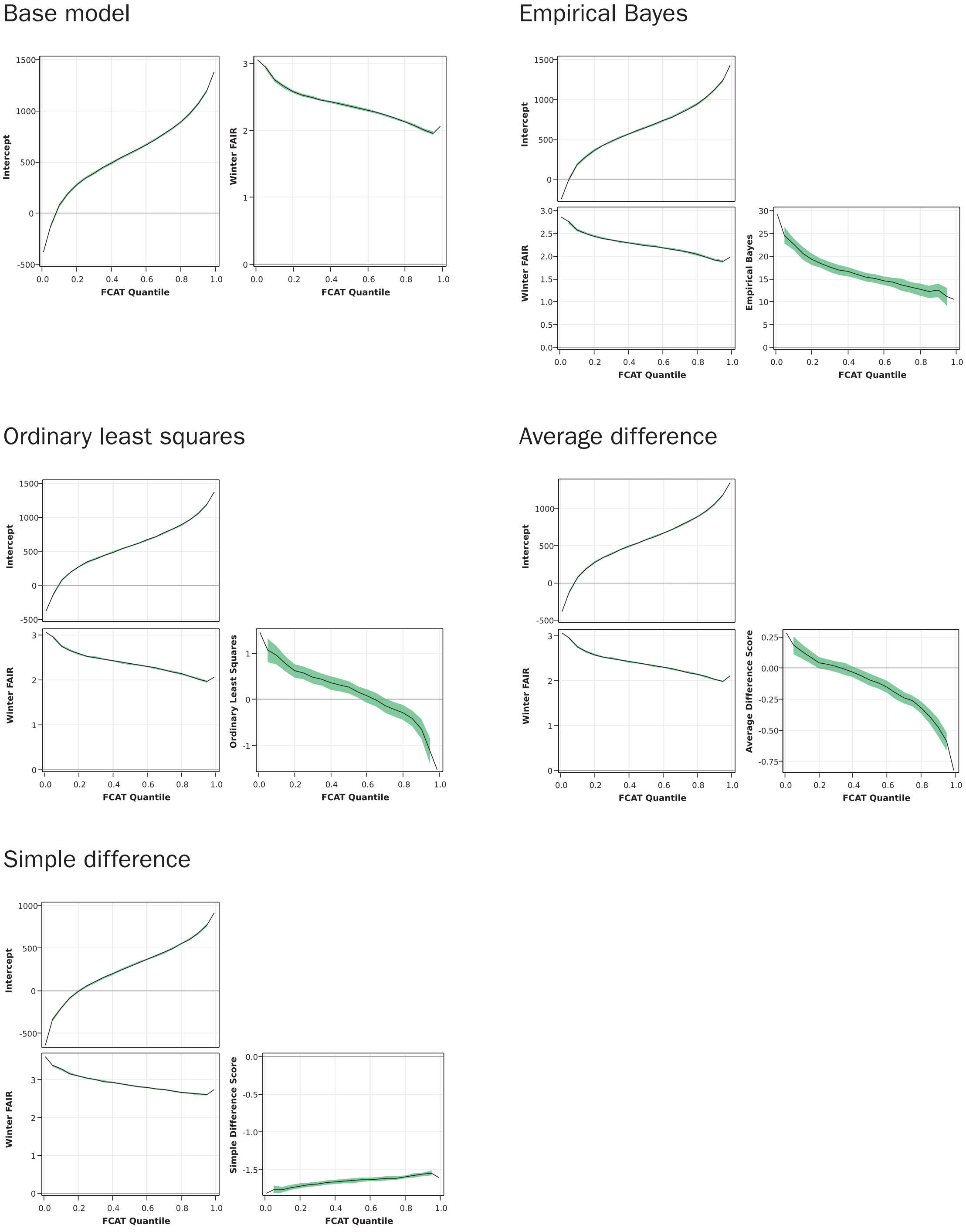
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Figure C8. Grade 10: unstandardized multiple quantile regression process plots centering time at the winter FAIR: estimated parameter by quantile for the FCAT, 2009/10**



Base model

Empirical Bayes

Ordinary least squares

Average difference

Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

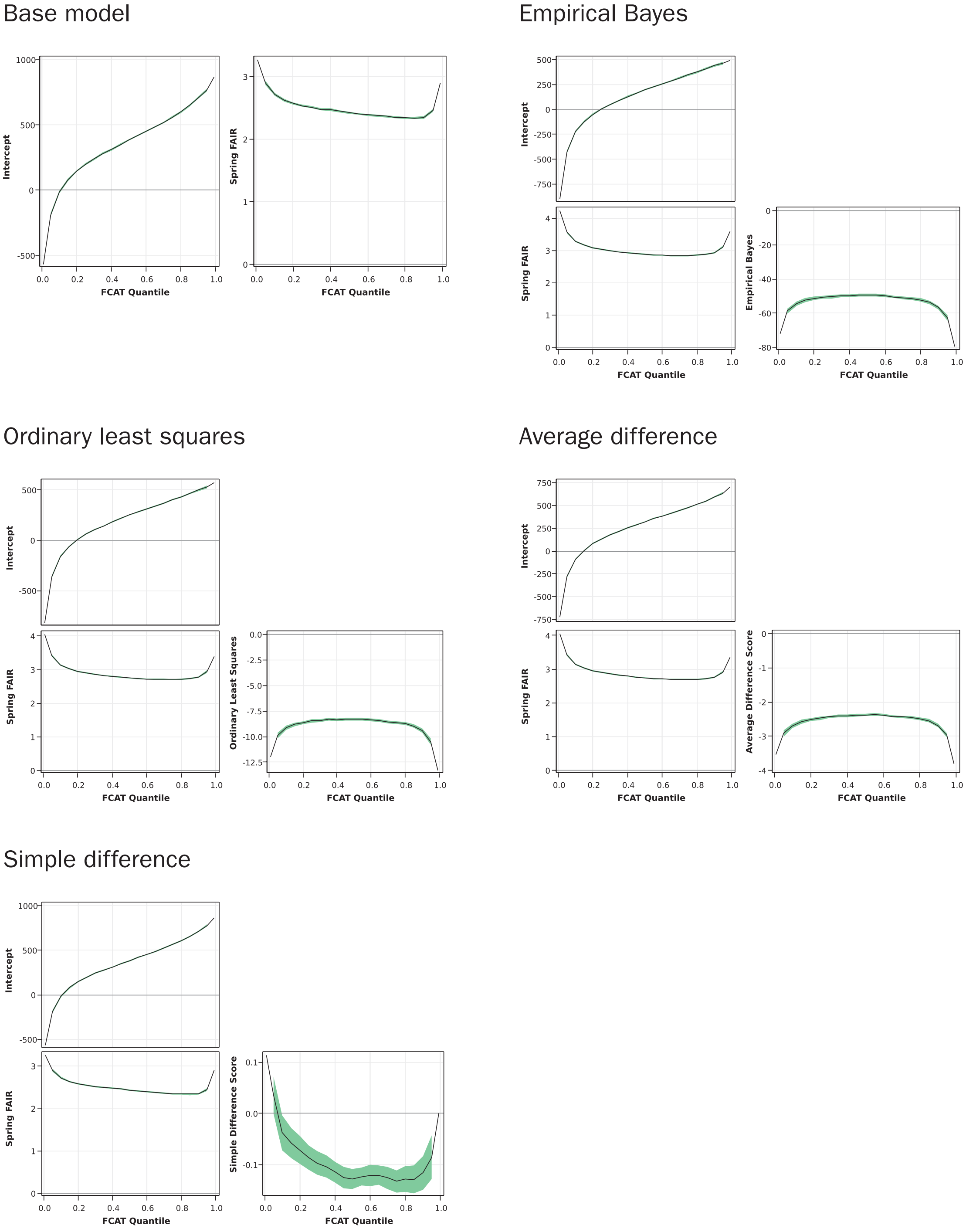
**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

# Appendix D. Unstandardized multiple quantile regression process plots centering time at the final (spring) status on the Florida Assessments for Instruction in Reading

This appendix contains quantile process plots that depict the relation between growth and reading comprehension performance conditional on student reading comprehension performance after controlling for initial (fall) status. Three graphs are included in each reported process plot because the models have two predictors as well as an intercept; the base model plot, which includes only one predictor (status) plus an intercept, has just two graphs. The intercept portion of the process plot displays the predicted reading comprehension score (y-axis) across the distribution of Florida Comprehensive Assessment Test (FCAT) scores when spring status and growth measure are 0 (x-axis). This plot shows that students at lower quantiles of FCAT performance have lower reading comprehension scores. The remaining plots reflect the slope coefficients for status and growth conditional on reading comprehension performance. Most pertinent is the growth plot, which shows that the coefficient decreases as FCAT performance increases.

## Figure D1. Grade 3: unstandardized multiple quantile regression process plots centering time at the spring FAIR: estimated parameter by quantile for the FCAT, 2009/10



Base model

Empirical Bayes

Ordinary least squares

Average difference

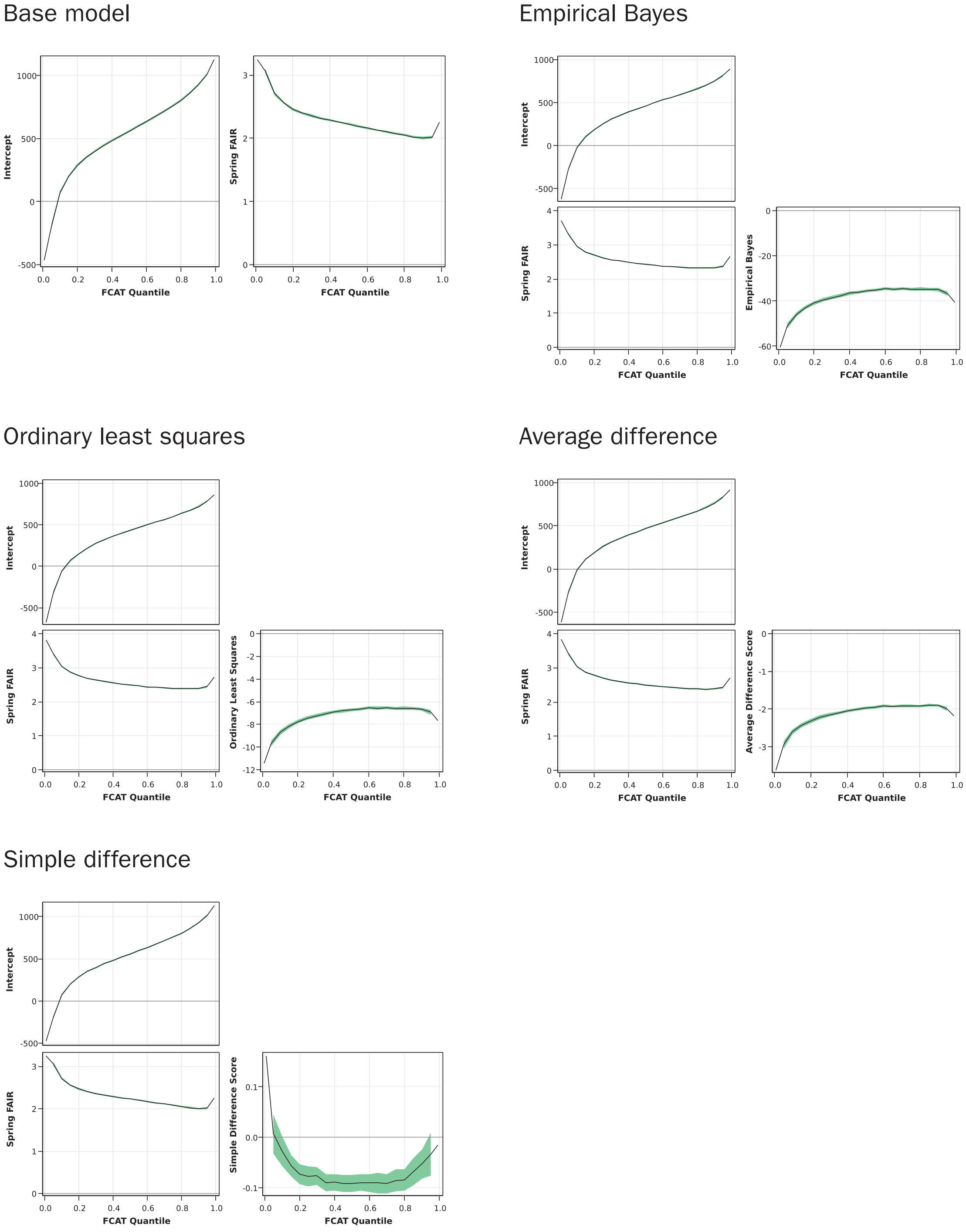
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Figure D2. Grade 4: unstandardized multiple quantile regression process plots centering time at the spring FAIR: estimated parameter by quantile for the FCAT, 2009/10**


Base model

Empirical Bayes

Ordinary least squares

Average difference

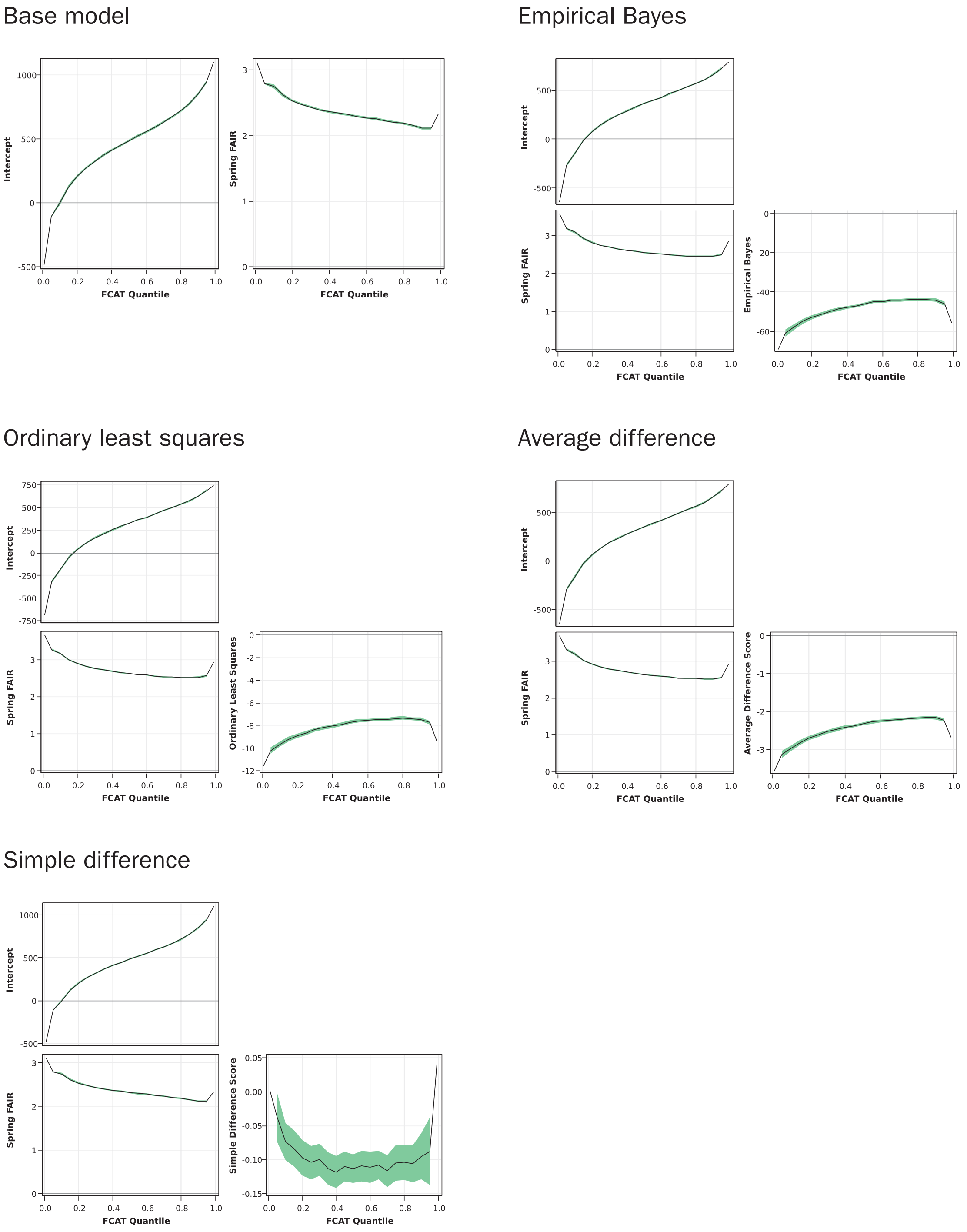
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors’ analysis based on data from Florida Department of Education (2010).

## Figure D3. Grade 5: unstandardized multiple quantile regression process plots centering time at the spring FAIR: estimated parameter by quantile for the FCAT, 2009/10



Base model

Empirical Bayes

Ordinary least squares

Average difference

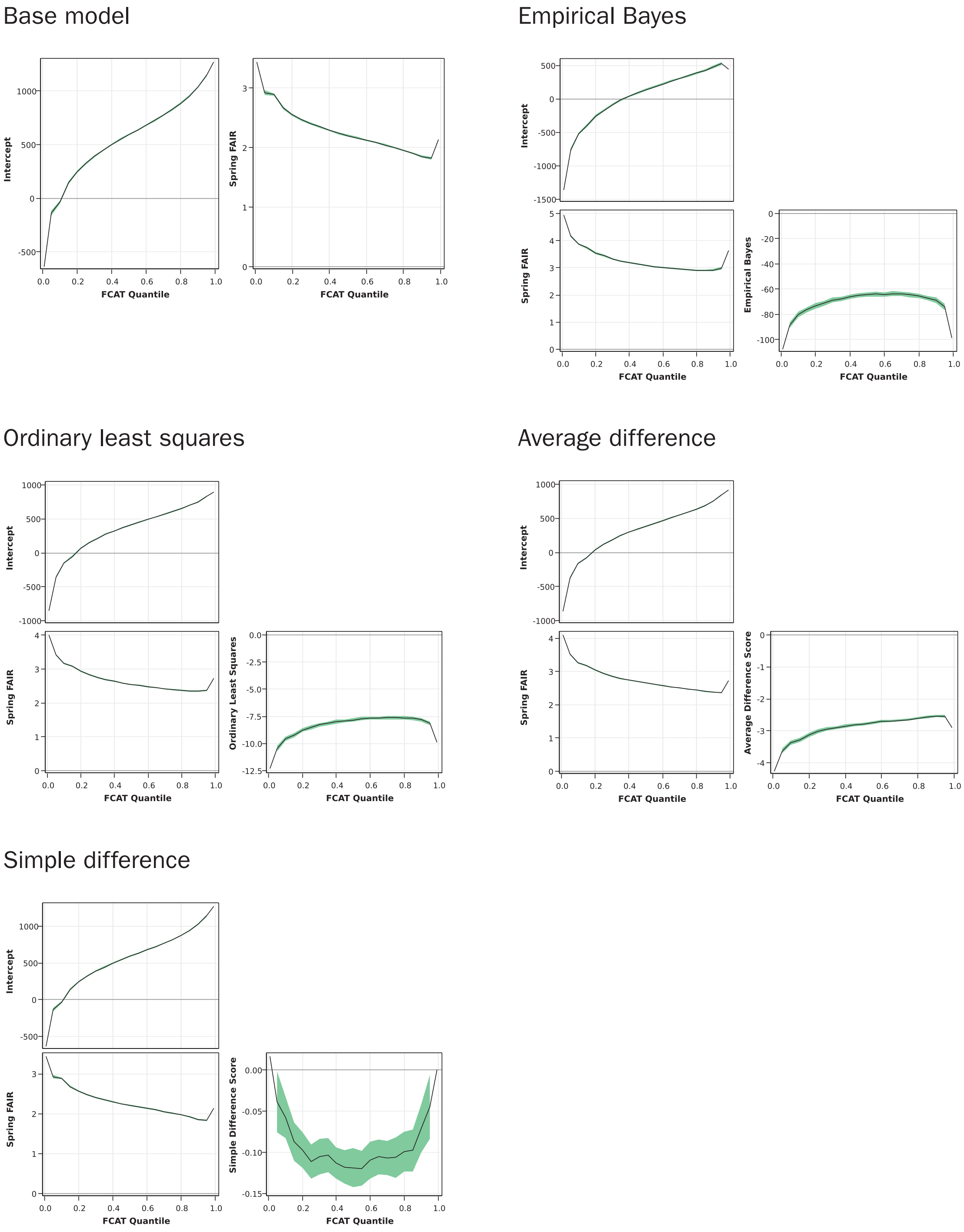
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

**Figure D4. Grade 6: unstandardized multiple quantile regression process plots centering time at the spring FAIR: estimated parameter by quantile for the FCAT, 2009/10**



Base model

Empirical Bayes

Ordinary least squares

Average difference

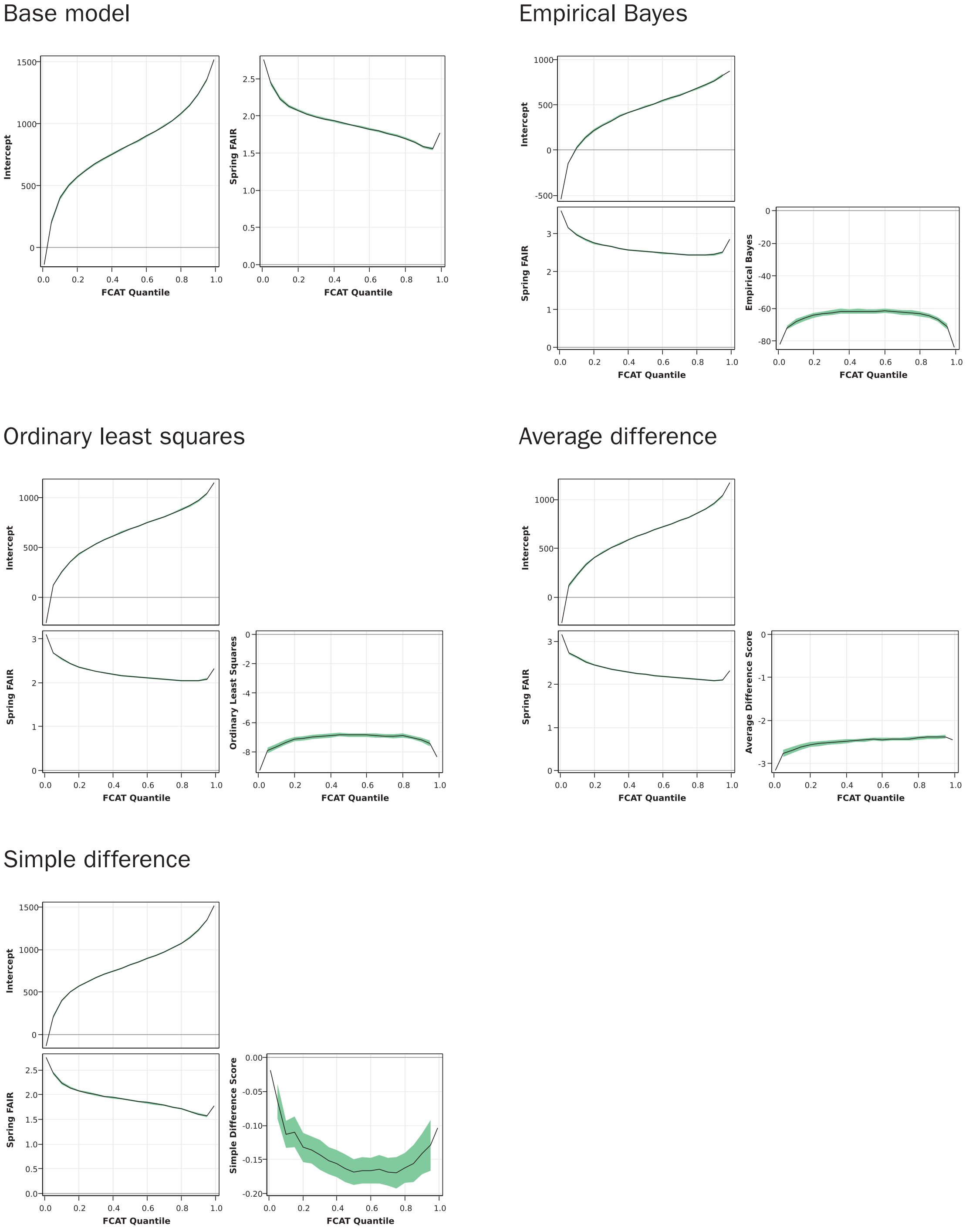
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

## Figure D5. Grade 7: unstandardized multiple quantile regression process plots centering time at the spring FAIR: estimated parameter by quantile for the FCAT, 2009/10



Base model

Empirical Bayes

Ordinary least squares

Average difference

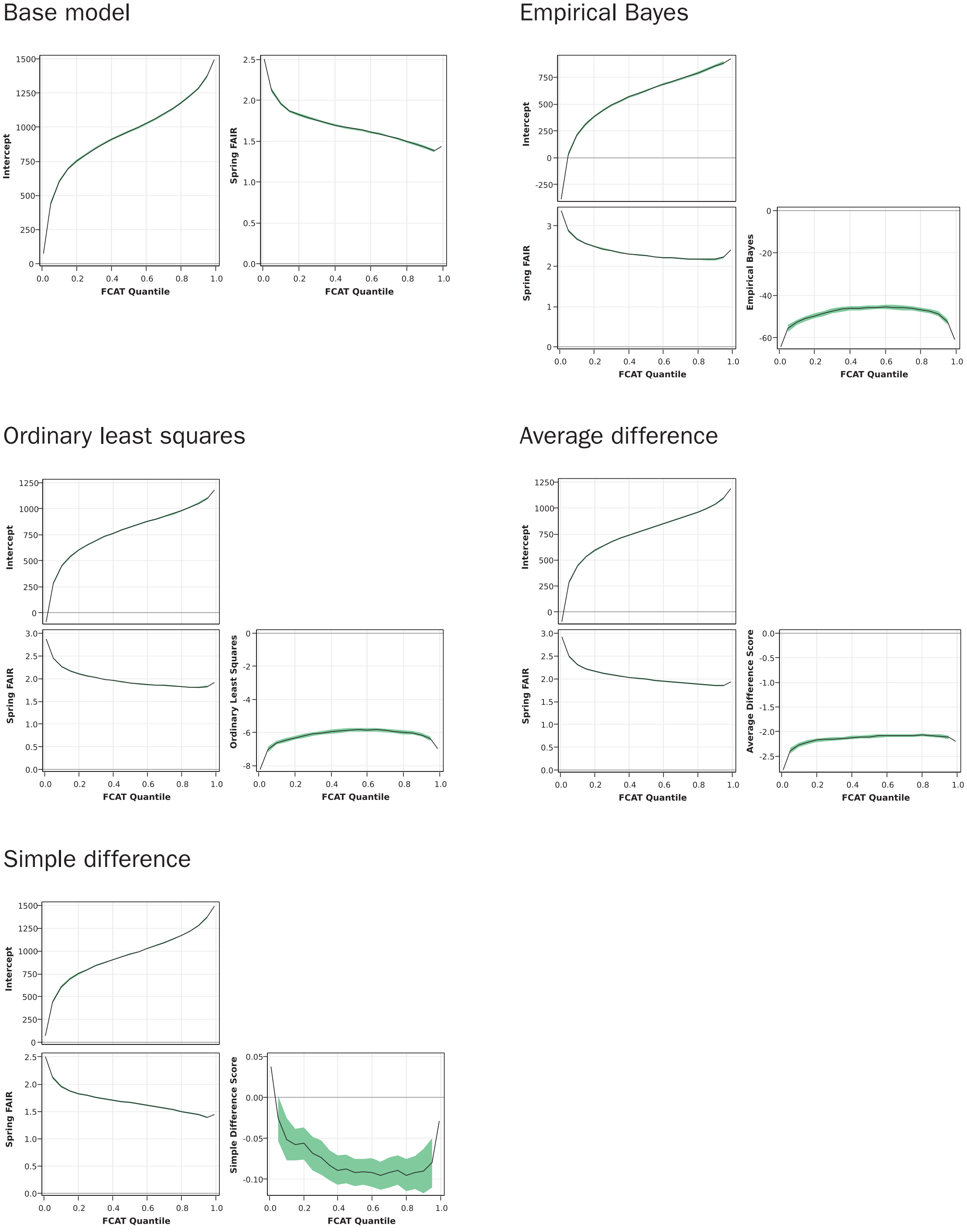
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

## Figure D6. Grade 8: unstandardized multiple quantile regression process plots centering time at the spring FAIR: estimated parameter by quantile for the FCAT, 2009/10



Base model

Empirical Bayes

Ordinary least squares

Average difference

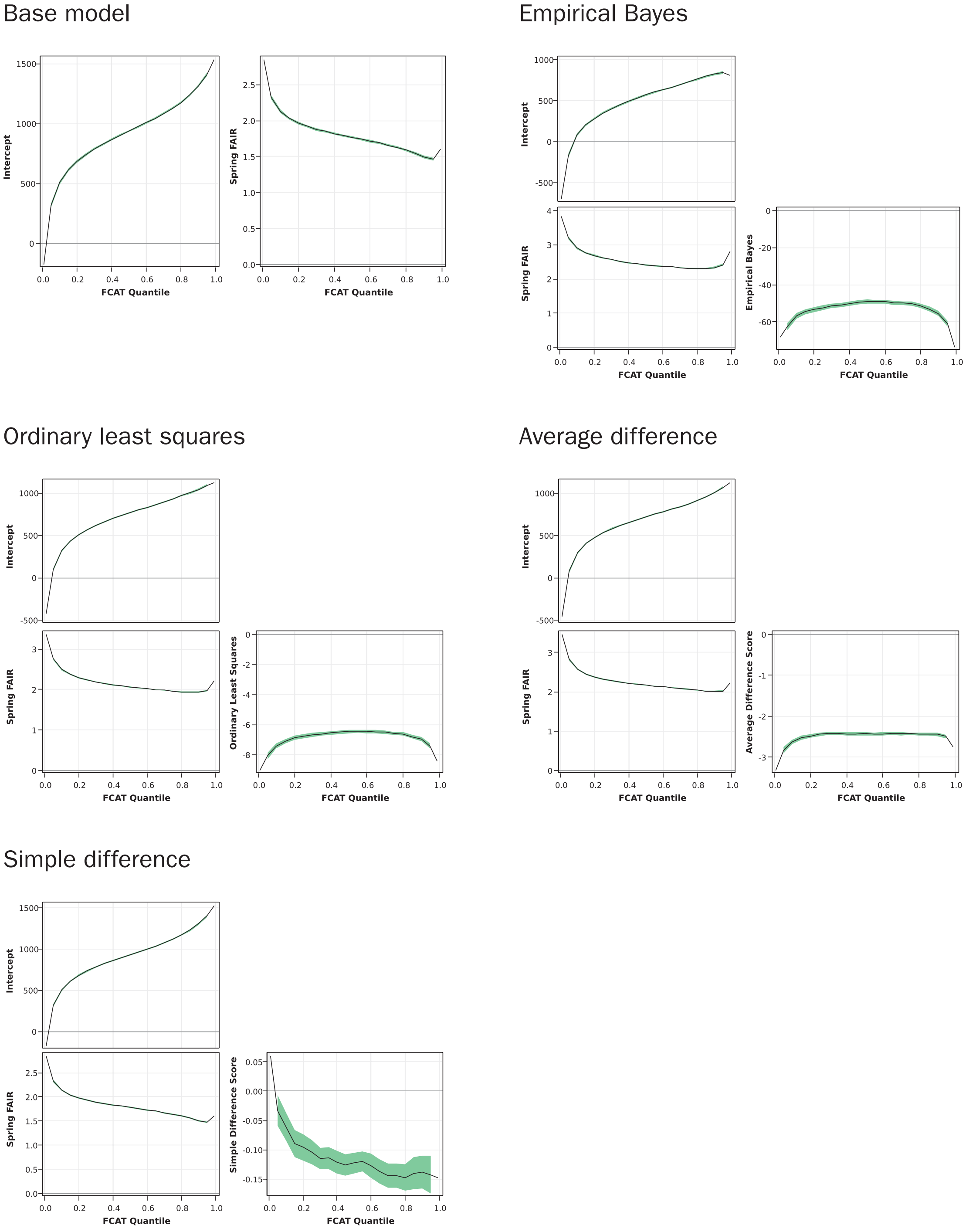
Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

## Figure D7. Grade 9: unstandardized multiple quantile regression process plots centering time at the spring FAIR: estimated parameter by quantile for the FCAT, 2009/10

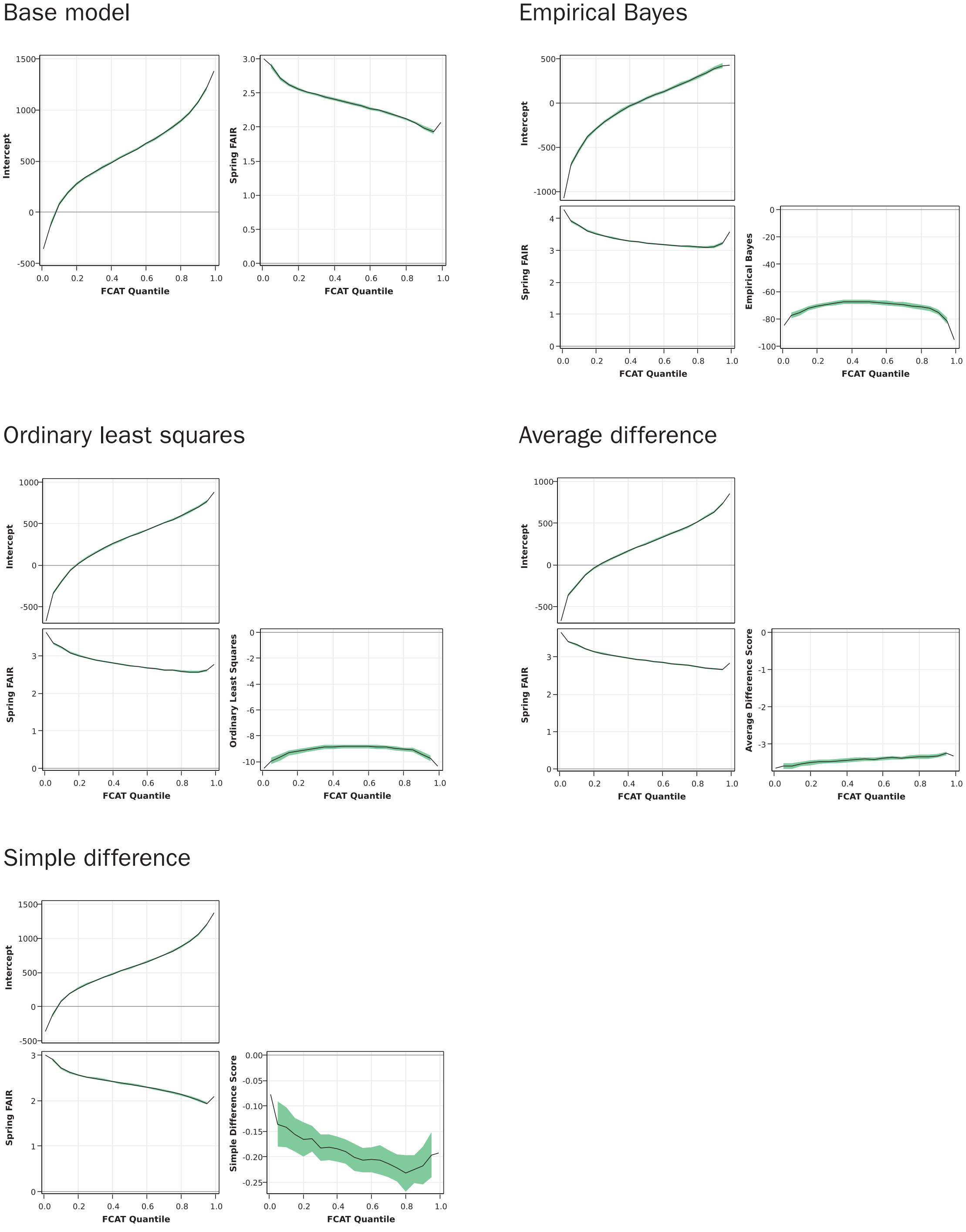


FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors’ analysis based on data from Florida Department of Education (2010).

## Figure D8. Grade 10: unstandardized multiple quantile regression process plots centering time at the spring FAIR: estimated parameter by quantile for the FCAT, 2009/10



Base model

Empirical Bayes

Ordinary least squares

Average difference

Simple difference

FAIR is Florida Assessments for Instruction in Reading. FCAT is Florida Comprehensive Assessment Test.

**Note:** At 95 percent confidence intervals.

**Source:** Authors' analysis based on data from Florida Department of Education (2010).

# Notes

1. In a cloze task students are presented with a portion of text with certain words removed (cloze text) and asked to replace the missing words. Cloze tasks require that students understand context and vocabulary.
2. For more technical information on multilevel models, see Raudenbush and Bryk (2002), Hox (2010), and O'Connell and McCoach (2008).
3. Little's (1985) "missing completely at random" was assessed for all variables by grade, and in all instances the null hypothesis (that data were missing completely at random) was rejected ($p < .001$).
4. The pattern of results was the same regardless of whether data used were the original scores or the imputed data. Cohen's *d* for the difference between the original and imputed data ranged from –0.05 to 0.00 and averaged –0.02, –0.03, and –0.02 for fall, winter, and spring FAIR across grades 3–10. Results are available from the first author on request.
5. In each model the model-adjusted $R^2$ was equal to the model-estimated $R^2$ due to the large samples at each grade level ($n$ = 100,000) and the small number of predictors (two) in each model.
6. Model diagnostics for the multilevel analysis included an evaluation of the residuals by time-point. All models indicated that the residuals at each time-point were centered on 0. Results are available from the first author on request.
7. Unstandardized regression coefficients for each model by grade are reported in appendix A, tables A1–A8. Table A10 reports the $\Delta R^2$ between growth measures controlling for winter status by grade.
8. Unstandardized regression coefficients for each model by grade are reported in appendix A, tables A1–A8. Table A9 reports the $\Delta R^2$ between growth measures controlling for fall status by grade.
9. Unstandardized regression coefficients for each model by grade are reported in appendix A, tables A1–A8. Table A11 reports the $\Delta R^2$ between growth measures controlling for spring status by grade.

# References

Ardoin, S. P., & Christ, T. J. (2008). Evaluating curriculum-based measurement slope estimates using triannual universal screening. *School Psychology Review, 37,* 109–125. http://eric.ed.gov/?id=EJ817291

Ardoin, S. P., & Christ, T. J. (2009). Curriculum based measurement of oral reading: Estimates of standard error when monitoring progress using alternate passage sets. *School Psychology Review, 38,* 266–283. http://eric.ed.gov/?id=EJ842725

Ardoin, S. P., Christ, T. J., Morena, L. S., Cormier, D. C., & Klingbeil, D. A. (2013). A systematic review and summarization of the recommendations and research surrounding Curriculum-Based Measurement of Oral Reading Fluency (CBM-R) decision rules. *Journal of School Psychology, 51,* 1–18. http://eric.ed.gov/?id=EJ1001681

Azen, R. (2013). Using dominance analysis to estimate predictor importance in multiple regression. In Y. Petscher, C. Schatschneider, & D. L. Compton (Eds.), *Applied quantitative analysis in education and social sciences* (pp. 34–64). New York: Routledge. http://eric.ed.gov/?id=EJ879631

Barth, A. E., Stuebing, K. K., Fletcher, J. M., Cirino, P. T., Romain, M., Francis, D., et al. (2012). Reliability and validity of oral reading fluency median and mean scores among middle grade readers when using equated texts. *Reading Psychology, 33,* 133–161. http://eric.ed.gov/?id=EJ969748

Calhoon, M. B. (2005). Effects of a peer-mediated phonological skill and reading comprehension program on reading skill acquisition of middle school students with reading disabilities. *Journal of Learning Disabilities*, 38, 424–433. http://eric.ed.gov/?id=EJ722272

Calhoon, M. B., & Petscher, Y. (2013). Individual sensitivity to instruction: Examining reading gains across three middle-school reading projects. *Reading and Writing, 26,* 565–592. http://eric.ed.gov/?id=EJ999056

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Furby, L. (1970). How we should measure “change”—or should we? *Psychological Bulletin, 74,* 68–80.

Cummings, K. D., Park, Y., & Schaper, H. A. B. (2013). Form effects on DIBELS next oral reading fluency progress-monitoring passages. *Assessment for Effective Intervention, 38,* 91–104. http://eric.ed.gov/?id=EJ995832

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch,C., Cable, A., Tackett, K. K., & Schnakenberg, J. W. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes on older struggling readers. *Review of Educational Research*, 79, 262–287. http://eric.ed.gov/?id=EJ879149

Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J. D. (2010). Creating a progress-monitoring system in reading for middle-school students: Tracking progress toward meeting high-stakes standards. *Learning Disabilities Research & Practice, 25,* 60–75. http://eric.ed.gov/?id=EJ881446

Florida Department of Education. (2001). *FCAT handbook—A resource for educators*. Tallahassee, FL: Author.

Florida Department of Education. (2005). *FCAT briefing book*. Tallahassee, FL: Author.

Florida Department of Education. (2009). FAIR 3–12 Technical Manual. Tallahassee, FL: Author. Retrieved August 9, 2010, from http://www.fcrr.org/FAIR/3–12_Technical_Manual_FINAL.pdf

Florida Department of Education. (2009–2011). *Florida Assessments for Instruction in Reading (FAIR)*. Tallahassee, FL: Author.

Foorman, B. R., & Petscher, Y. (2010a). *Summary of the predictive relationship between the FAIR and the FCAT in grades 3–10.* Technical Report. Tallahassee, FL: Florida Center for Reading Research.

Foorman, B. R., & Petscher, Y. (2010b). *The unique role of the FAIR Broad Screen in predicting FCAT reading comprehension*. Technical Report. Tallahassee, FL: Florida Center for Reading Research.

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' ORF using DIBELS. *Journal of School Psychology, 46,* 315–342.

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Education Achievement. Retrieved January 14, 2008, from http://dibels.uoregon.edu

Harcourt Brace. (2003). *Stanford Achievement Test, Tenth Edition: Technical data report*. San Antonio, TX: Author.

Hintze, J. M., & Christ, T. J. (2004). An examination of variability as a function of passage variance in CBM progress monitoring. *School Psychology Review, 33,* 204–217. http://eric.ed.gov/?id=EJ683511

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Jenkins, J. R., Graff, J. J., & Miglioretti, D. L. (2009). Estimating reading growth using intermittent CBM progress monitoring. *Exceptional Children, 75,* 151–163. http://eric.ed.gov/?id=EJ842530

Kim, Y. S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology, 102,* 652–667. http://eric.ed.gov/?id=EJ892640

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

Little, R. J. A. (1985). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83,* 1198–1202.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16,* 421–437.

No Child Left Behind Act of 2001. (2002). Pub. L. No. 107–110, 115 Stat. 1425.

O'Connell, A. A., & McCoach, D. B. (Eds.). (2008). *Multilevel modeling of educational data.* Charlotte, NC: Information Age Publishing.

Petscher, Y., Cummings, K. D., Biancarosa, G., & Fien, H. (2013). Advanced (measurement) applications of curriculum-based measurement in reading. *Assessment for Effective Intervention, 38,* 71–75. http://eric.ed.gov/?id=EJ995833

Petscher, Y., & Kim, Y. S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology, 49,* 107–129. http://eric.ed.gov/?id=EJ911351

Petscher, Y., & Logan, J. A. R. (in press). Quantile regression in the developmental and behavioral sciences: An introduction. *Child Development.*

Petscher, Y., Logan, J. A. R., & Zhou, C. (2013). Extending conditional means modeling: An introduction to quantile regression. In Y. Petscher, C. Schatschneider, and D. L. Compton (Eds.), *Applied quantitative analysis in the education and social sciences* (pp. 1–34). New York: Routledge.

Pyle, N., & Vaughn, S. (2012). Remediating reading difficulties in a response to intervention model with secondary students. *Psychology in the Schools, 49,* 273–284. http://eric.ed.gov/?id=EJ989957

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2004). *HLM6: Hierarchical linear and nonlinear modeling.* Chicago: Scientific Software International.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Quantitative Methods in Psychology, 92,* 726–748.

SAS Institute Inc. (2012). *Base SAS 9.3 Utilities: Reference.* Cary, NC: SAS Institute Inc.

Schatschneider, C., Buck, J., Torgesen, J. K., Wagner, R. K., Hassler, L., & Hecht, S., et al. (2004). *A multivariate study of factors that contribute to individual differences in performance on the Florida Comprehensive Reading Assessment Test* (Technical Report No. 5). Tallahassee, FL: Florida Center for Reading Research. Retrieved May 4, 2009, from http://www.fcrr.org/TechnicalReports/Multi_variate_study_december2004.pdf

Schatschneider, C., Wagner, R. K., & Crawford, E. C. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning and Individual Differences, 18,* 308–315. http://eric.ed.gov/?id=EJ807603

Singer, J. D., & Willett, J. B. (Eds.). (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* London: Oxford University Press.

Tichá, R., Espin, C. A., & Wayman, M. M. (2009). Reading progress monitoring for secondary-school students: Reliability, validity, and sensitivity to growth of reading-aloud and maze-selection measures. *Learning Disabilities Research & Practice, 24,* 132–142. http://eric.ed.gov/?id=EJ850229

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency.* Austin, TX: Pro-Ed.

Vaughn, S., Cirino, P.T., Wanzek, J., Wexler, J., Fletcher, J. M., Denton, C. D., et al. (2010). Response to intervention for middle school struggling readers: Effects of a primary and secondary intervention. *School Psychology Review, 39,* 3–21. http://eric.ed.gov/?id=EJ886407

Vaughn, S., Wexler, J., Leroux, A., Roberts, G., Denton, C., Barth, A., & Fletcher, J. (2012). Effects of intensive reading intervention for eighth-grade students with persistently inadequate response to intervention. *Journal of Learning Disabilities, 45,* 515–525. http://eric.ed.gov/?id=EJ982001

Vaughn, S., Wexler, J., Roberts, G., Barth, A., Cirino, P. T., Romain, M. A., Francia, D., Fletcher, J., & Denton, C. A. (2011). Effect of individualized and standardized interventions on middle school students with reading disabilities. *Exceptional Children, 77,* 391–407. http://eric.ed.gov/?id=EJ931144

Woodcock, R. W. (1998). *Woodcock Reading Mastery Test-Revised.* Circle Pines, MN: American Guidance Service.

Yeo, S., Fearrington, J. Y., & Christ, T. J. (2012). Relation between CBM-R and CBM-mR slopes: An application of latent growth modeling. *Assessment for Effective Intervention, 37,* 147–158. http://eric.ed.gov/?id=EJ964499

Zumeta, R. O., Compton, D. L., & Fuchs, L. S. (2012). Using word identification fluency to monitor first-grade reading development. *Exceptional Children, 78,* 201–220. http://eric.ed.gov/?id=EJ970677