AMERICAN INSTITUTES FOR RESEARCH®

Value Added Model for Measuring Student Learning Gains:

General Approach and Framework for Evaluating Candidate Models

General Approach and Framework for Evaluating Candidate Models

AIR proposes a team that integrates expertise in statistical modeling of value-added, proven capabilities in stakeholder outreach and deep knowledge of, and historical perspective on, Florida's accountability initiatives. Working in collaboration with the Florida Department of Education, Florida's Value-added Technical Advisory Committee (VTAC), and Florida's stakeholders, we will support the evaluation and development of a value-added system that will support Florida's talent management initiative.

While the RFP refers to value-added models (VAMs), in fact, each VAM can be decomposed into at least four components:

- The data and metrics on which the model relies
- A statistical model that yields estimates of systematic student growth among students taught by a common teacher
- A method of, or approach to, classification of teachers as highly effective, effective, or ineffective
- An approach to reporting the results of the model

As the initial phase of this project proceeds, the Department may find that each "model" is assembled from these separable parts. It is entirely possible that none of the popular comprehensive models ideally suits Florida's purposes as it is currently assembled. Instead, the Department may find the optimal model by splitting apart the components and reassembling them into a new, comprehensive whole. Below, we discuss each of the components and some of the key considerations to which they give rise.

Data and Metrics

Every value-added model begins with assessment data—measures of student proficiency taken over time. All VAMs require a linkage between students and the teachers responsible for them. Some models require ancillary data, such as student demographics, to obtain estimates. Not only do some models require different data, different models often place differential demands on the data. For example, many VAMs (McCaffrey, et al, 2004) require that the assessment data use a common metric across grades (a vertical scale). Others (Betebenner, 2008) eliminate this requirement by focusing purely on normative relationships, studying percentile ranks each year.

In light of this reliance on specific data elements, reflecting on the strengths and limitations of the current Florida data is worthwhile.

Assessment scores

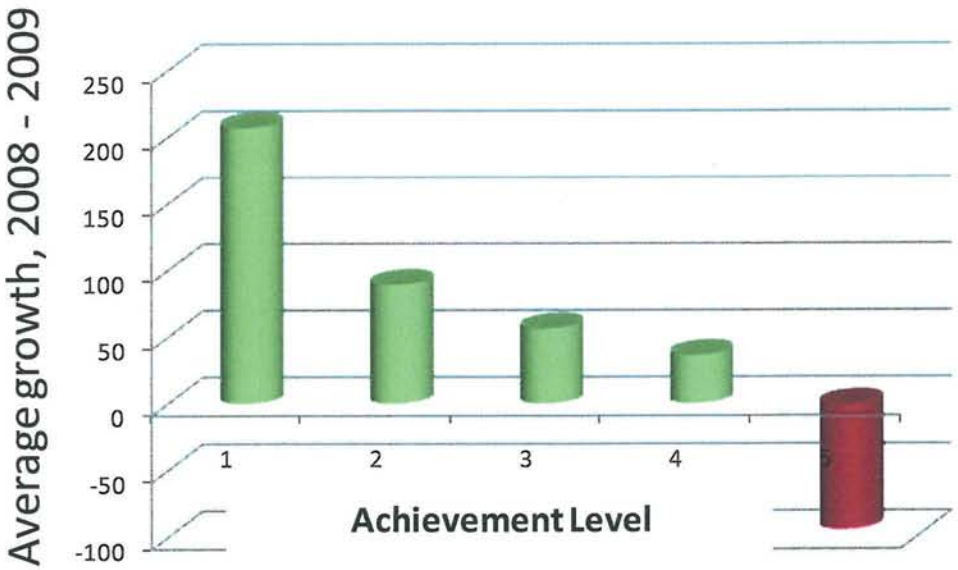
Our last two projects working with FCAT data to estimate value-added models introduced us to at least three salient characteristics of the current FCAT data:

- Measured growth varies dramatically, and perhaps not entirely accurately, across different achievement ranges within grades.

- Average measured growth varies dramatically across grades.
- Average measured scores within a grade fluctuate across years, with predictable impact on average growth scores.

The variation described in the first point suggests limitations in measurement rather than true differences in the rate at which students learn. Exhibit 2 graphs the average growth from grade 4 to grade 5 on the FCAT for grade 5 students in 2009. This pattern is typical—the greatest measured growth is observed among the lowest-achieving students, and growth among the top-achieving students is almost always negative. This, of course, does not reflect real growth. If it did, there would be no durable inequalities in achievement—the best would drop back to the middle, and stragglers would quickly catch up.

Exhibit 2: FCAT Grade 5 Reading—Typical Growth Pattern
(original analysis conducted for the Foundation for Excellence in Education)

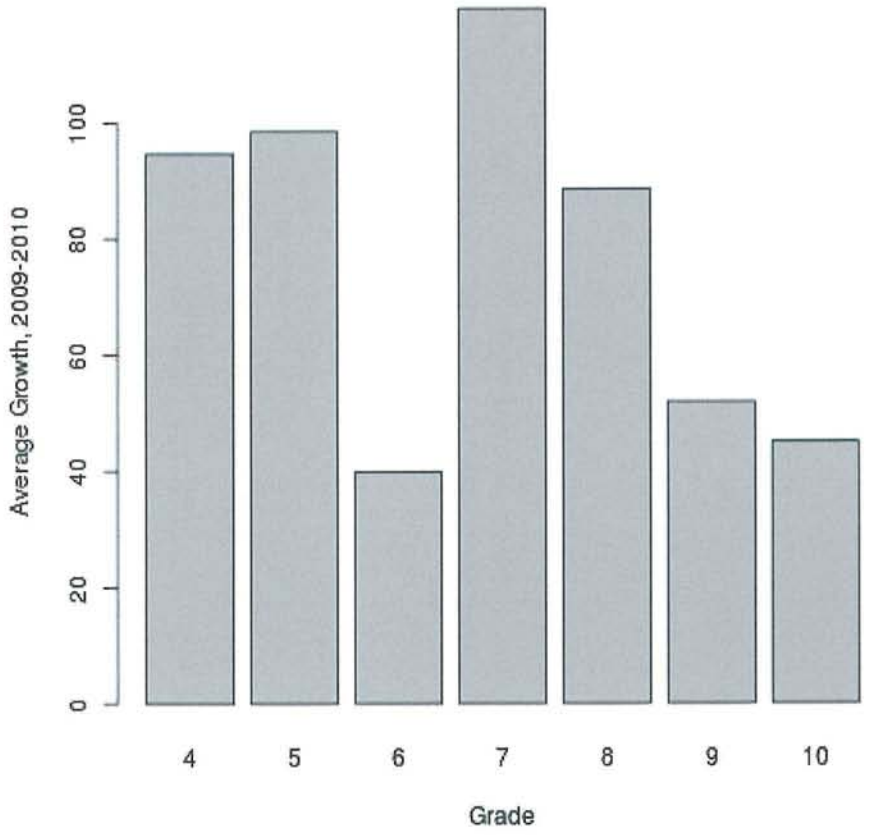


The pattern apparent in Exhibit 2 is a classic example of "ceiling and floor effects." The FCAT does not adequately measure the range of student performance, so the lowest-achieving students hit a floor. Because they have hit the floor, measurement error is almost guaranteed to be positive because the scores can go no lower. Further, these measurement error effects are likely to be large because test scores are typically stable at the proficiency cut points and far less stable near the ceiling and the floor. A growth model that does not account for this will unfairly penalize teachers of high-achieving students and reward teachers of low-achieving students, regardless of the actual effectiveness of the teachers.

The second point, illustrated in Exhibit 3, may also be a methodological artifact. Whether this phenomenon reflects real differences in learning or an idiosyncrasy of the current FCAT developmental scale, it has implications for the evaluation of teachers. If it is an artifact, ignoring

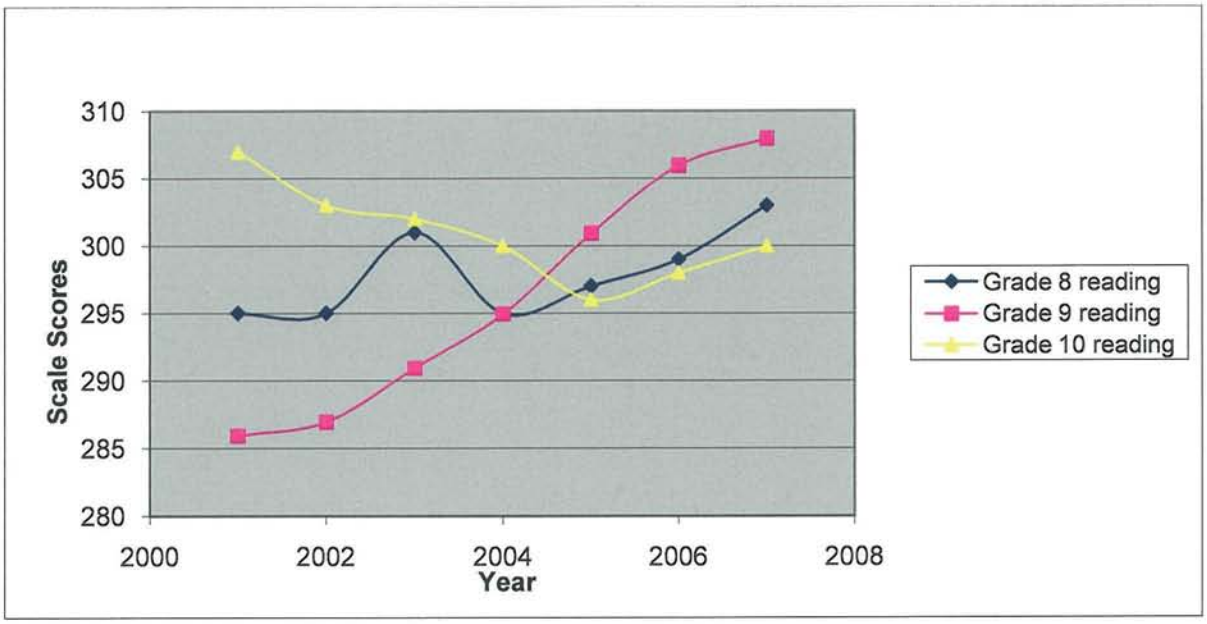
it would advantage teachers teaching certain grades. If it reflects real differences in learning, it might signal greater concentrations of good teachers in some grades but might also reflect the differential predilections of students at various stages in their development. In that case, teacher evaluation policy could inadvertently establish incentives for teachers to teach in some grades rather than others.

Exhibit 3: FCAT Average Growth by Grade, 2008–2009



Finally, we see sometimes substantial deviations in the average performance of students across years. Exhibit 4 shows average FCAT score trends from 2000 to 2008. Those data illustrate the sorts of fluctuations that sometimes appear in assessment data. For example, grade 8 reading scores spike inexplicably in 2003, only to return to their prior level in 2004, and begin ascending the next year. Grade 10 scores plummet at the same time the adjacent grade's scores climb, only to reverse direction in 2006. Although it is tempting to explain these fluctuations as reflecting the work of exceptional cohorts or some other one-time factor, in fact the swings are often within the expected range of random fluctuation due simply to linking one year's test form to the next. This unfortunate effect of equating error can have a profound impact on value-added estimates if it is not recognized and appropriately managed.

Exhibit 4: Florida Comprehensive Assessment Test



We recognize that Florida is adopting Next Generation of Sunshine State Standards and implementing FCAT 2.0 in 2011 to measure student mastery of these standards. We also recognize the desire and intent to link FCAT 2.0 to the existing scale to preserve the longitudinal trend. In the longer term, these characteristics of the test may change as the next generation test is introduced. However, the initial years will necessarily draw on the existing data, and Florida will be well served to take the match between the model and the limitations of the FCAT scale into account in the choice of model.

Teacher/Student Match

The second component involves accurately attributing the students taught by each teacher. Florida is a national leader in accurately attributing students to teachers via their course enrollments; however, because the data have not yet been used for high-stakes decisions, they are not perfect. As with school grades and adequate yearly progress (AYP) designations, it is essential to have data accuracy reviewed and verified beyond initial reporting.

In our work on the Foundation for Excellence in Education's Excellence in Teaching awards program, we found the data attributing students to teachers to be quite accurate when we interviewed principals about potential awardees. However, on occasion we learned that the wrong teacher had been associated with a class, that a second teacher also had responsibility for the class, or that a student was misassigned. On occasion, data entry errors, special circumstances, or a lack of care in identifying teachers responsible for particular classes led to inaccurate data.

This experience suggests that the Department and districts should collaborate to facilitate an attribution "clean up" process. The process could mirror the process used for school accountability to verify the Survey 2/3 match, the assessment records match, and the accuracy of

the student demographic information, but be conducted at a teacher level rather than the school level. The Department is well positioned to give districts, schools, and teachers an opportunity to review the data on which their value-added estimates are based to ensure accuracy and provide transparency.

Demographics and Other Data

The Invitation to Negotiate (ITN) indicates that the Department is interested in exploring models that control for demographic and other contextual data. Doing so effectively requires that the data be recorded accurately. For many data elements, the Department already has effective processes in place to encourage and support accuracy. For example, the groups used for AYP reporting undergo a data verification process in which districts have an opportunity to update data for accuracy.

Other data elements in the Education Data Warehouse (EDW), particularly those data elements not used for accountability or funding decisions, tend to be less accurate. For example, we found that the teacher experience data, which are not used operationally and not used in an accountability system, were largely inaccurate.

Models that rely on contextual variables will need to reflect the realities of which data are currently collected accurately, or operational data collection procedures will have to change.

Statistical Models

Florida seeks value-added models that are fair, accurate, reliable, and stable. VAMs run the full gamut from simple and transparent to quite complex and nuanced. Whichever model is selected must be estimable by the Department staff, preferably using SAS software without undue computational burden or proprietary licenses. Therefore, we will begin by evaluating statistical models according to this criterion.

All VAMs attempt to estimate the systematic component of growth associated with a school, teacher, or other input. To measure growth, the models must control for prior achievement in some way.

The most transparent models simply subtract the prior score of a student from his or her current score. This model gives an estimate of individual student growth that may then be aggregated through a variety of mechanisms to estimate the teacher's impact.

A slightly more flexible version of the simple approach controls for prior achievement in a regression model. Under this approach, it is straightforward to introduce covariates that effectively compare a teacher with other teachers teaching students with similar measured characteristics. This model is similar to that used by Sass and colleagues (Mihaly, McCaffrey, Lockwood & Sass, 2010) and by Kane and Staiger (2008) in their now well-known experiment in Los Angeles.

More-complex models model student proficiency as a deviation from an average—possibly the average within a grade level or the average within a grade level, district, and school. Teachers

are then associated with the average of these deviations, either directly, in what are called *fixed-effect* models, or indirectly through correlations in the stochastic effects associated with students taught by the same teacher. The latter are called *random-effects* models. Once the model parameters are estimated in a random-effects model, a second step is necessary to estimate teacher effects. This step is usually accomplished through an *empirical Bayes* calculation, which may yield estimates with a small bias but can actually give better estimates, on average, than unbiased estimators.

The more-complex models may be estimated with or without covariates and admit a wide variety of specification details. For example Sanders' model Ballou, D., Sanders, W., and Wright, P. (2004) explicitly assumes that the impact of a teacher never decays and that the impact has a lasting effect on the student. McCaffrey and colleagues (2004) relax this assumption and allow the model to estimate the decay over time. We discuss some of these models in a subsequent section.

A comparative analysis of different estimating models will require a framework for evaluating the extent to which each is fair, accurate, reliable, and stable.

Assessing fairness requires using a variety of policy judgments beyond the technical evaluation. It is possible that some students are simply more difficult to teach than others. Is it fair to teachers to require them to achieve similar growth, thereby requiring some teachers to work harder than others to gain the same evaluation? Is it fair to students to expect less of some than of others?

Policy judgments must be made by Florida's policymakers and stakeholders. AIR can support this decision-making process by offering clear, transparent descriptions of what the model is estimating. A technical summary of disparate student expectations can summarize differences among models. Specifically, we propose to calculate, for each model, the lowest target growth for any demographic group to the highest target growth for any demographic group. A model with a fixed, criterion-referenced expectation would have a ratio of 1. A model that controls for the demographic mix of students taught by a teacher would likely have a ratio lower than 1. A similar statistic can be calculated across performance groups. This feedback will enable policymakers to make informed judgments about fairness of expectations. AIR can support this decision-making process by offering clear, transparent descriptions of what the model is estimating. Other technical aspects of fairness depend on the estimator's accuracy, reliability, and stability.

We take accuracy to refer to the extent to which estimates from the model match the values of the true, underlying trait that they are designed to measure. Statisticians typically look at two criteria in this regard: unbiasedness and consistency. Any statistical estimate fluctuates from its real values. An unbiased estimate fluctuates around the true value. Consistency refers to the extent to which the estimates get closer to their true value as the sample size increases.

As mentioned above, random-effects models often yield estimates that may be biased. But in an important sense, they may be better than available unbiased estimates because neither unbiasedness nor consistency is particularly helpful without reliability—the precision of the estimates. The standard error of an estimate provides an estimate of its precision (σ_e). The total

variance of a statistic across a population can be decomposed into the sum of its random variance (σ_e^2) and its true, systematic variation (σ_t^2). The reliability coefficient,

$$\frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2},$$

offers a useful measure of the reliability of the estimates.

Stability is related to reliability. It refers to the consistency of repeated measures over time. An unreliable measure will not be stable. A perfectly reliable measure will be unstable if the underlying trait being measured varies over time. When the trait, in this case teacher effectiveness, is perfectly stable over time, reliability and stability will be the same. Of course, we know that teacher effectiveness is not a stable trait. Inexperienced teachers learn and become more effective. Sometimes teachers have a bad year or lose interest in their vocation and become less effective.

To some extent, evaluating the stability of a measure will rely on judgment. It is probably not plausible that true teacher effectiveness in one year is uncorrelated with teacher effectiveness in the next. Some teachers are better than others. However, it is probably also implausible that true teacher effectiveness is perfectly correlated across years.

An evaluation of stability will depend on an accurate, critical examination of the reliability measures. If the reliability is overestimated, results can seem unstable even when their variation reflects only the variability of measurement. If the estimate of the standard error is wrong, or the assumptions on which it is based mischaracterize the real world, the reliability will be overestimated. Therefore, it is important to test the assumptions on which the standard errors are based and to evaluate the precision of the estimates if those assumptions are violated. This will require simulation studies drawing on both real and artificial data to evaluate the volatility of the model under assumptions thought plausible in the real world.

Finally, applying the various models to the FCAT data, we can examine whether they appear to be estimating the trait that they set out to estimate, teacher effectiveness. Although this cannot be done directly, we can certainly look for the shadows on the wall. For example, Feng, Figlio and Sass (2010) set out a model that estimates average student deviations from their demographic group–level means. If estimated accurately, estimates from this model ought to be uncorrelated with the demographic mix in the classroom. Therefore, we can evaluate whether this goal is achieved, and even whether it generalizes to demographic groups not explicitly included in the model.

Classification of Teachers

With estimates of teacher performance in hand, value-added providers typically offer some standard guidance on using these estimates to classify teachers. This is one important place where the stages of the model can be usefully split, and Florida may find a mix-and-match approach to be the most beneficial.

In theory, growth criteria can be either normative or criterion referenced. In practice, normative

targets have been widely favored over fixed criteria. For example, the increasingly popular Colorado model (Betebenner, 2008) is purely normative, often suggesting that teachers be divided into those whose students learn more than the median and those whose students learn less. Sanders' model (Ballou et al, 2004) seeks a *de facto* ordering of teachers.

In an educational climate where norm-referenced student tests have been all but abandoned, and schools and districts have been judged on fixed criteria, this logic has not found its way into teacher evaluation. In some ways this is surprising given that a normative system establishes "moving targets," guaranteeing that not all teachers can reach the goal. In any event, whether targets are expressed in terms of fixed criteria or normative relationships is important as Florida selects a VAM.

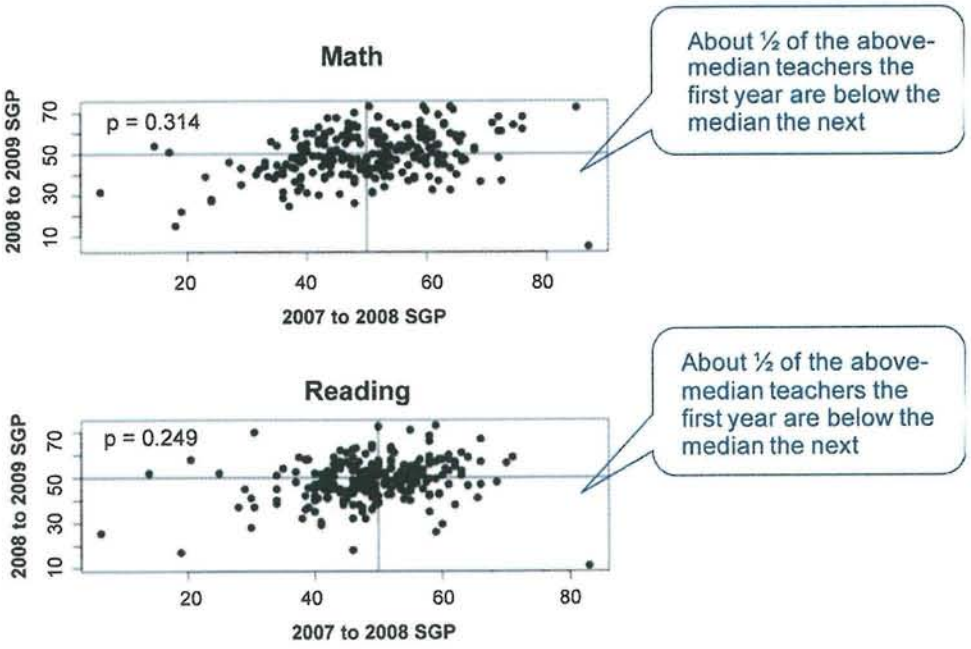
A second set of criteria involves the classification consistency of the method used to make judgments about teachers. For example, McCaffrey, Sass, and Lockwood (2008) found that teacher classifications were quite unstable over time. These findings are summarized in Exhibit 5.

Exhibit 5: Classification Stability of Four Value-Added Models (McCaffrey, Sass, & Lockwood, 2008)

Study	Best teachers not identified as a best teacher the next year	Best teachers identified as below average the next year
Ballou (2005) Tennessee	50%	30%
Aaronson et al. (2007) Chicago	57%	20%
Koedel and Betts (2007)	35%	30%
McCaffrey et al. (2008) Hillsborough County Florida	40%	22%

A recent application of the Colorado model in Hawaii showed similarly inconsistent classifications. Exhibit 6 plots teachers' percentile ranks for mathematics (top panel) and reading (bottom panel). Using above/below the median as a classification strategy leads to inconsistent classifications of about half of all teachers between 2008 and 2009 in both subjects—approximately what we would expect from random classifications.

Exhibit 6: School-Level Growth Estimates for 2008 vs. 2009



Most value-added models do not explicitly attempt to describe their classification consistency. Doran and Cohen (2009) present a broadly applicable approach to evaluating the classification consistency of VAM-based teacher classification. This approach enables analysts to calculate the expected number (and percentage) of false positive and false negative classifications.

We propose to evaluate models on the basis of their classification consistency (false positives and false negatives) and the extent to which policy choices are available to influence these values.

Reporting Results

We recognize that the Department wants the reporting to be transparent and to maximize the instructional value of the reports. We agree and will work with the Department toward these objectives. We do note, however, that some purposes are better served through the FCAT reporting system. Specifically, value-added models do not require or use item-level or strand-level scores. Integrating fine-grained content reporting would be a distraction from the primary responsibilities for the VAM, which are monitoring, tracking, and evaluating teacher effectiveness.

The VAM reporting system does have an important instructional role to play. Students in Florida will benefit if the system can establish clear targets and expectation for each teacher in terms of what he or she has to accomplish over the next year. Most VAMs use multiple years of data to establish an estimate, and those estimates change as new data flow into the system. An ideal VAM reporting system will tell each teacher what he or she individually must accomplish with students during the new school year to be considered effective or highly effective. Some models are more amenable to this sort of reporting than others.

A second, related feature of reporting is the transparency of the model and the relationship between student outcomes and teacher classifications. The Department recognizes that teachers and principals will need a mechanism to provide feedback about specific data elements (e.g., a teacher associated with the wrong class, students misidentified by classroom). The more transparent a model is, the more quickly and easily educators will be able to spot and report anomalous data.

Summary evaluation of models

In reviewing the components that make up most value-added models, we noted that different value-added estimates estimate different underlying things. It is not simply that they get different estimates of teacher effectiveness—they implicitly *define* teacher effectiveness *differently*. Therefore, in summarizing models, we propose to begin with a clear, concise description of how each model defines teacher effectiveness.

After that, we have identified 18 indicators to describe and compare models, grouped into four categories: data, model, classification, and reporting. We propose to report each indicator, working with Department staff, advisors, and the VTAC to summarize the overall quality and fit to the Florida context within each of the four categories using a 5-point scale.

Exhibit 7 presents a sample chart that might be used to begin the comparison of models. Of course, the detail behind each measure and additional factors can be taken into account. However the project will require a mechanism to summarize this large amount of information in a way that can be used to structure consideration, discussion, and debate. We reiterate that the Department may choose to mix and match aspects of different models to define one that best meets Florida's needs.

Exhibit 7: Sample Chart That May Be Used for Initial Summary and Comparison of Models

INDICATOR	MODEL 1	MODEL 2	[...]	MODEL K
Concise description of what the model estimates				
Data: Suitable for Florida FCAT Data (Overall Rating)				
Is or can be made robust to linking error across years				
Is or can be made robust to imperfections in vertical scale				
Model: Accurate and Reliable (Overall Rating)				
Can be estimated using SAS with reasonable computation burden and without specialized software				
Ratio of lowest-to-highest expectation across demographic groups				
Ratio of lowest-to-highest expectations across performance groups				

INDICATOR	MODEL 1	MODEL 2	[...]	MODEL K
Unbiased estimates				
Consistent estimates				
Reliability coefficient				
Available, accurate standard error estimator				
Standard error estimator that accurately describes real-world stability				
Uncorrelated with presumed independent factors				
Correlated with presumed related factors				
Classification Consistency (Overall Rating)				
Normative or criterion-referenced growth targets?				
Classification accuracy amenable to by policy decisions?				
False positive rate at recommended configuration				
False negative rate at recommended configuration				
Reporting (Overall Rating)				
Suitable for actionable feedback for teachers				
Sufficiently transparent to support appeals/verification process				