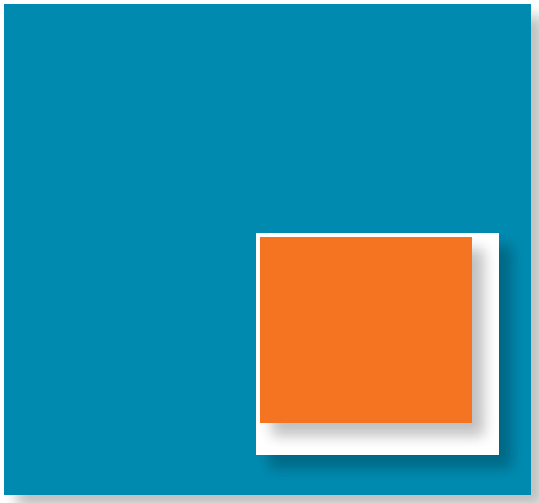
CCSSO
Council of Chief State School Officers

INCORPORATING ENGLISH LEARNER PROGRESS INTO STATE ACCOUNTABILITY SYSTEMS



January 2017

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

Authors:

Pete Goldschmidt, California State University Northridge

Kenji Hakuta, Stanford University

Suggested Citation: Goldschmidt, P. & Hakuta, K. (2017). *Incorporating English Learner Progress into State Accountability Systems*. Washington DC: Council of Chief State School Officers.

One Massachusetts Avenue, NW, Suite 700 • Washington, DC 20001-1431
Phone (202) 336-7000 • Fax (202) 408-8072 • www.ccsso.org

TABLE OF CONTENTS

The Every Student Succeeds Act (ESSA) proposes changes in how states include the nation's growing population of English Learners (ELs) into the accountability system. The purpose of this paper is to identify key issues and questions that might be considered and explored by state decision makers in this area. Our primary audience is anyone in a state agency engaged in making decisions about the state's accountability system and how ELs are included in that system. We provide background information on policy history, basic information on types of accountability models, and several demonstrations of how some options play out on real system data. Throughout this paper we present empirical results from two states: for state 1 we utilize a random sample of the state's students; for state 2 we utilize data from a single district (states 1 and 2 use different ELP assessments.) We encourage state leaders to engage in thoughtful deliberation around how these issues apply to their own state context by simulating similar scenarios using their existing data and applying their analysis to the construction of the state plan.

BACKGROUND

Equal educational opportunity for the nation's English Learners (ELs) is a requirement of the Civil Rights Act of 1964, the Equal Opportunities Act of 1974, and by agreements derived from court decisions including *Lau v. Nichols* (1974) (*Lau*). *Lau* defines opportunity is defined as access to both English language development as well as to core academic content. In their Dear Colleague Letter (DCL), the Office for Civil Rights of the U.S. Department of Education and the Civil Rights Division of the Department of Justice affirm three standards in considering whether programs are in compliance with these laws: (1) whether they are based on sound educational theory; (2) whether they are reasonably implemented; and (3) whether they demonstrate effectiveness within a reasonable period of time. A good accountability system serves the needs of ELs, adheres to specific titles of the Elementary and Secondary Education Act (ESEA), and provides the tools needed to demonstrate program effectiveness.

Assessment of English language proficiency was introduced in the 1970's to identify students limited in their English proficiency (ELs were referred to in legislation prior to ESSA as "Limited English Proficient"). The first English Language Proficiency (ELP) assessment requirements were introduced into ESEA under *No Child Left Behind* (2001) (NCLB). The *Improving America's Schools Act* of 1994, the Act that began the standards-based reform movement in education, included provisions calling for the inclusion of ELs in those assessments.

Under Title III of NCLB, districts were required to report progress (AMAO 1) and status (AMAO 2) on the state ELP assessment. Academic achievement for Title III (AMAO 3) used the same targets as for the EL subgroup under Title I. It is important to recognize that because ELP was part of Title III, consequences for not meeting AMAOs for two or four years were not applicable to students in Title I districts not receiving Title III funding. ESSA has changed this; performance on the ELP assessment is now placed in Title I accountability as a school-level indicator for all EL students, not just for those students in LEAs receiving Title III funding. Under ESSA[1], scrutiny of the ELP assessment will take place through the peer review process for state Title I accountability, under the same umbrella as general academic assessments. This also means that states will need to address how ELP standards—those standards to which the ELP

1 All references to ESSA in this document are to the ESEA as amended by ESSA.

assessments align—corresponds[2] to the academic standards of the state and that the ELP assessment is technically sound. It should also be noted that ELP assessments are not intended to assess content, so no prior content knowledge should be required to successfully answer items on the ELP assessment (Abedi, 2008). The ELP assessment should show ELs' mastery of academic language required for engagement and learning of core academic content through the medium of English.

Under ESSA, states may now include former ELs in the EL subgroup for a period of up to 4 years (formerly up to 2 years). This change will increase the number of former ELs included in the subgroup, and the average academic performance of the subgroup. This move will help to mitigate the inherent instability of the EL subgroup that leads to a selectivity bias, particularly in the higher grades.

Finally, under the new accountability regulations, the ELP indicator may take individual student characteristics into account. The graphic provided by Karen Thompson[3] (Figure 1) clearly shows two important characteristics that drive student reclassification: a student's initial level of English proficiency, and time in the system as an EL. The graph shows differentiated patterns of growth based on these two variables, so depending on the student demographics of a particular school, there may be different expectations for ELP progress. Models that can take individual student characteristics into account in this way are helpful for effectively communicating the importance of these variables to stakeholders and informing accountability system development.

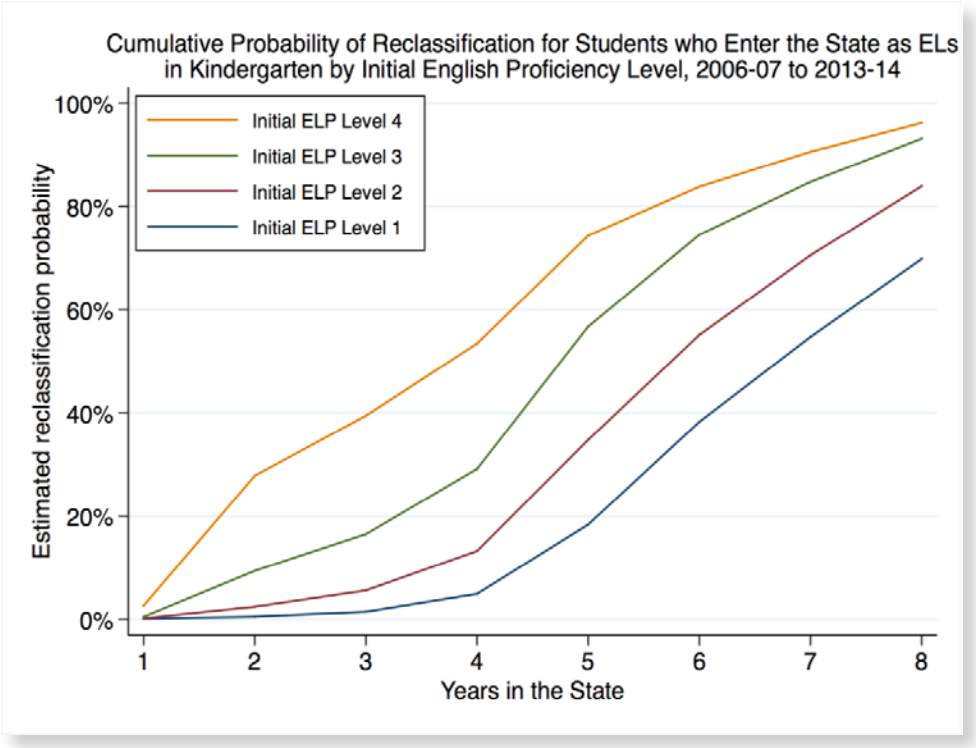


Figure 1: Cumulative Reclassification over Time

2 The law refers to the "alignment" of the ELP standards with the academic standards. However, because language proficiency and academic proficiency are different constructs, the field has come to adopt the term "correspondence". See http://www.ccsso.org/Documents/2012/ELPD%20Framework%20Booklet-Final%20for%20web.pdf p. 92ff

3 Karen Thompson, College of Education, Oregon State University

Another important indicator for states to consider is the English language development (ELD) trajectory over time. Evidence indicates that trajectories are non-linear, with faster growth occurring early and slowing down over time[4]. Second, overall, growth is relatively parallel among the initial ELD levels[5]. Finally, students entering in later grades (in Figure 2 initial grades are 3, 5, and 8) tend to score similarly to students who have been in the system longer and have the same ELD level and have consistent growth trajectories regardless of initial grade of entry. Figure 2 demonstrates these three aspects related to ELs' progress in State 1. This pattern is not absolute, but it is helpful for states to examine this pattern to determine whether their unique state's context interacts with progress model choice. For example, a simple year-to-year gain model would produce larger gains for first year students and subsequently smaller gains (meaning that schools that tend to serve students later in their English language development process, like middle schools, will earn fewer points—not because they are facilitating less growth, but because of the nature of language development growth.) If initial grade of entry is unrelated to subsequent English language growth, then this is beneficial in terms of modeling growth, but problematic because obtaining English proficiency is, as Figure 2 highlights, substantively based on time in a program. This then impacts thinking about a reasonable time to proficiency balanced against time limitations of high school graduation.

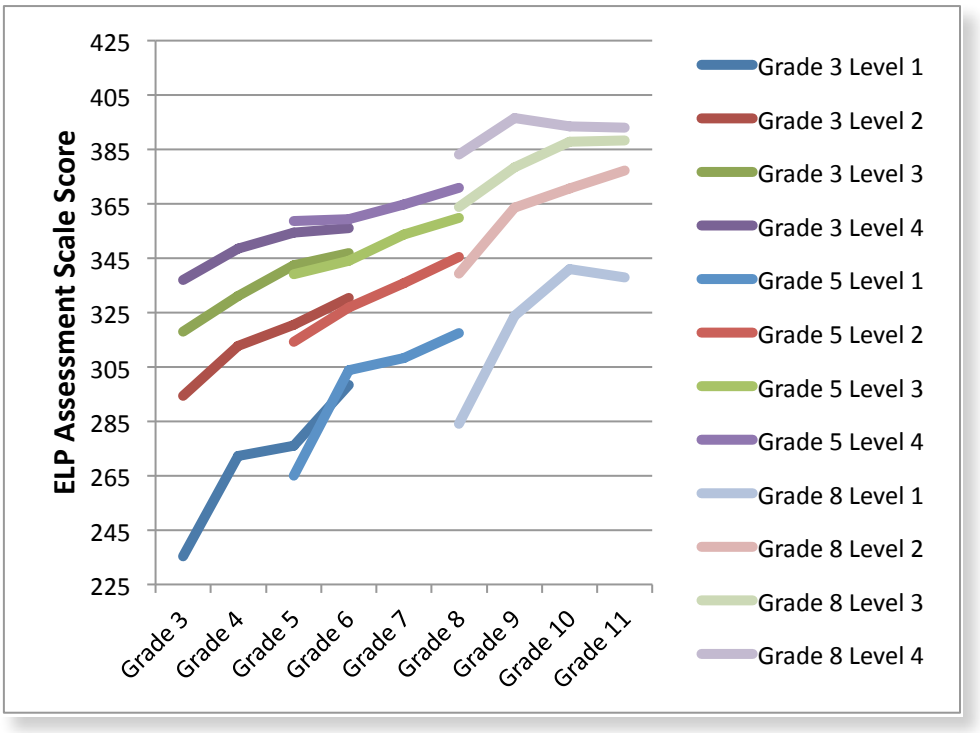


Figure 2: English Language Development by Initial ELD Level and Grade of Entry

4 See Sahakyan, N. & Cook, H. G. (2014). Examining District-Level Growth Using ACCESS for ELLs. WIDA Research Report. Downloaded at file:///C:/Users/Kenji%20Hakuta/Downloads/research-reports-WIDA_report_SAHAKYAN_COOK_district%20level%20growth.pdf

5 The initial ELD levels in figure 2 are 1, 2, 3, and 4. We do not present trajectories for initial level 5 and 6 because too few students contribute to three assessment occasions.

Figure 3 summarizes the average growth trajectory across all grades (Grades 1 to 11). The average trajectory in Figure 3 is consistent with Figure 2 in that growth is steeper early and slows in later grades. As seen in Figure 3, the average trajectory varies considerably; some of this variability can be systematically related to the schools that students attend.

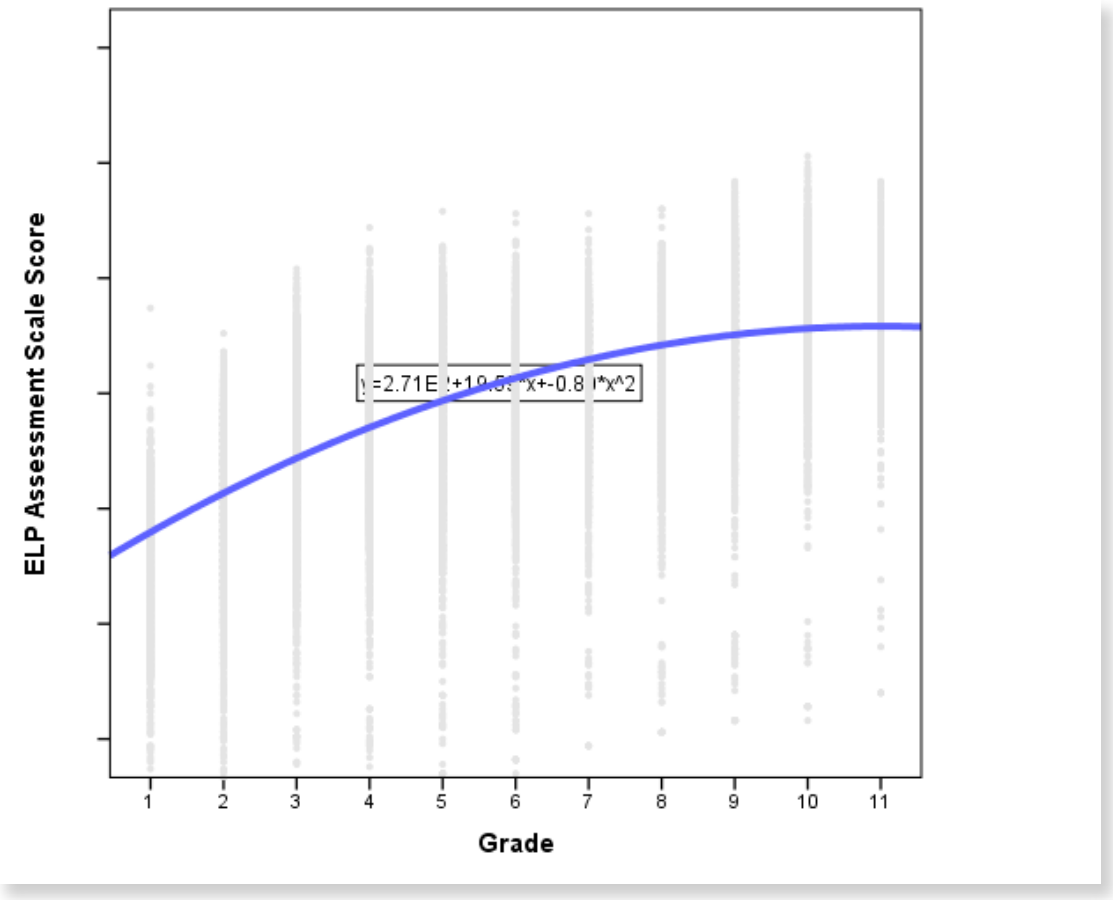


Figure 3: Average Growth Trajectory from Grade 1 to 11

Figures 4a and 4b, show that ELD levels align with content performance administered in English. As EL students obtain higher ELD levels, their content performance begins to approach (and surpass) that of English Only (EO) students. Figures 4a and 4b show that the scores are normalized and EO performance would be equal to 0. In other words, EL students at ELD level 5 would be expected to perform on par with EO students.

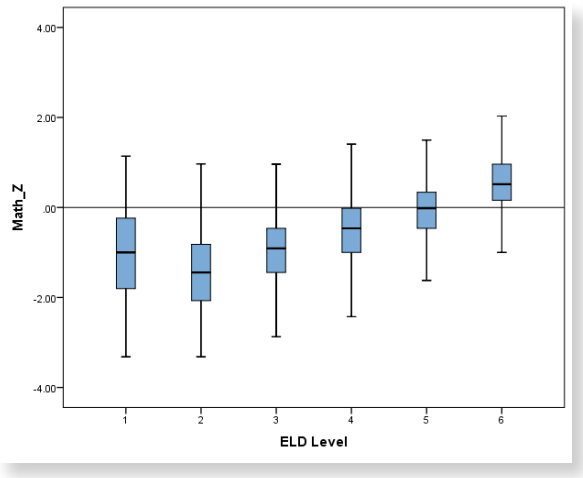


Figure 4a: Performance on Mathematics Assessment by ELD Level

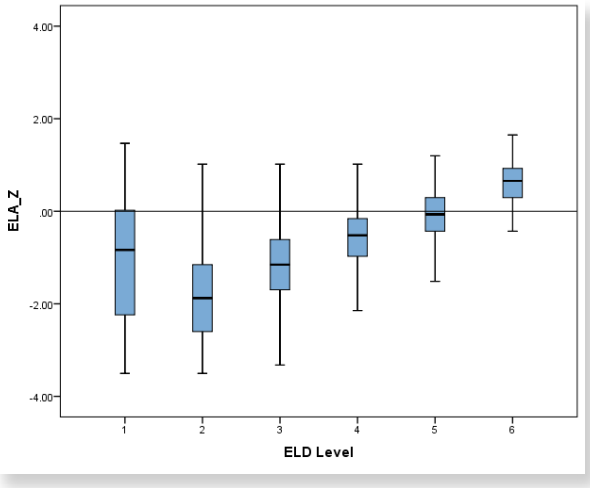


Figure 4b: Performance on English Language Arts Assessment by ELD Level

ESSA indicates that ELD must be considered for ELP progress accountability purposes, but States should consider ELD level in content performance (in English) *and* content growth (in English)[6]. Figures 5a and 5b indicates an apparent relationship between content performance growth and ELD level. In Figures 5a and 5b the vertical axis represents growth in the original assessment metric on State 1's assessment.

6 The law and regulations do not allow academic achievement or growth indicators on the ELA/math test to vary by student characteristics; however, we treat ELD level as a prior assessment result that (potentially) is as informative as ELA/math prior assessment results that are allowed in content growth calculations.

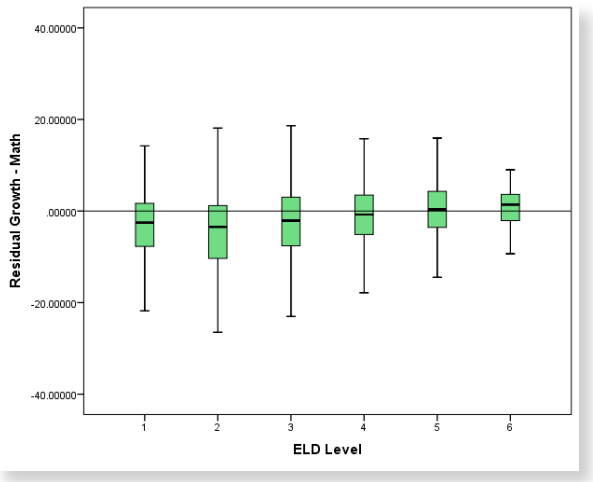


Figure 5a: Growth on Mathematics Assessment by ELD Level

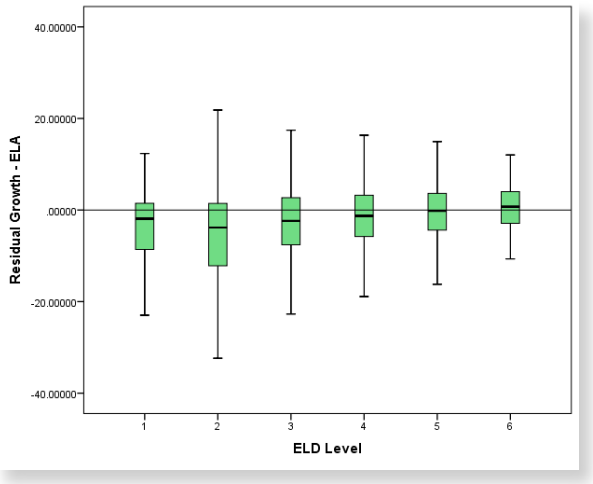


Figure 5b: Growth on English Language Arts Assessment by ELD Level

Assuming a state intends to include content growth in its state accountability plan, including ELD level in a growth model can account for differences in content growth (in English) by ELD level. For example a growth model[7] could be:

$$Y_i = f(X_i) + \text{ELDlevel}_i + \text{EOstatus}_i + \text{RELstatus}_i + e_i \quad (1)$$

Where Y_i is the content for student *i* and $f(X_i)$ is some function[8] of prior performance X for student *i*. ELDlevel is an EL student's ELD level; EOstatus is an indicator variable taking on the value of 1 if a student is an EO and 0 otherwise; and RELstatus is an indicator variable taking on a value of 1 for a REL (Reclassified English Proficient) student and 0 otherwise. The residual term is e for student *i*. For demonstration purposes we run the above growth model in two steps. Step one simply examines

7 We describe various growth models in more detail below.

8 In this instance we use the same subject prior performance, prior performance squared, and prior performance cubed – this model is a conditional status model – which we discuss in more detail, below.

performance differences on the mathematics assessment administered in English by Language status. Consistent with expectations, EO and REL students outperform EL students by a substantively meaningful amount. In Table 1, EL students (represented as constant) are expected to score about 28 points on the state English Language Arts (ELA) assessment, while EOs and RELs are expected to score 12.7 and 9.7 points higher, respectively. In step 2 we utilize the results from Figure 4a and assign an ELD level of 5 to EO students. We then expand the model to include prior achievement and ELD levels.

Table 1: ELA Performance by Language Status

Variable	B	Std. Error	t	*P*
(Constant)	28.423	.185	153.375	0.000
EO	12.722	.196	64.953	0.000
REL	9.682	.481	20.149	.000

Table 2: Conditional ELA Status (growth) by Language Status and ELD Level

Variable	B	Std. Error	t	*p*
(Constant)	1.016	.449	2.264	.024
EO	.158	.307	.513	.608
REL	-.304	.418	-.727	.467
SS_Rprior	.854	.006	152.456	0.000
SS_Rpriord^2	-.003	.000	-11.545	.000
SS_Rpriord^3	.000	.000	-21.602	.000
ELD Level	1.610	.167	9.656	.000

Table 2 indicates that once we take prior performance and ELD level into account, two important things happen. First, ELD level is a significant predictor of performance[9] and second, there is no longer a predicted difference in performace between EO and EL, or between REL and EL. ELD level captures the differences in performance by language classification. ELD thus becomes a reasonble element to include in a student English language content growth model because it is not a student background charactersitic, but a measure of preparedness based on student assessment results.

Figures 6a and 6b display growth results using a growth model that includes ELD level. The results indicate that including ELD level generally eliminates the negative bias in growth for students not already at an ELD level of 5 or 6.

9 ELD level is included to have a linear effect but could be modeled to have a non-linear effect, or as a step function, for example.

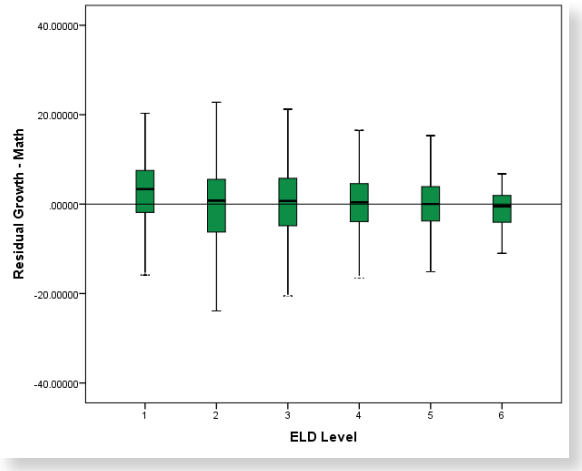


Figure 6a: Math Content Growth Conditioned on ELD Level by ELD Level

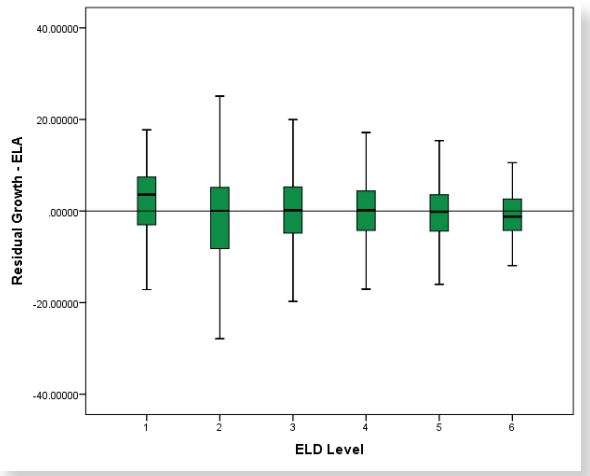


Figure 6b: English Language Arts Content Growth Conditioned on ELD Level by ELD Level

These descriptive analyses and modeling examples demonstrate that understanding the state context is important for developing a meaningful English language progress component of a state's overall accountability system. The remainder of this paper presents some growth modeling options and results based on those options. We then provide some results related to EL subgroup performance on English language content. It is important that we look closely at including ELD level in the content growth model because doing so reduces the impact on schools who serve a larger proportion of ELs. In terms of measuring content growth, this would ensure that schools would not be advantaged nor disadvantaged based solely on its students' distribution of language proficiency.

Before going into the remainder of this paper which will contain many technical details, we suggest that the reader reviews the following questions that your state team might want to consider. These serve as a sort of advance organizer for the information to follow.

A. What are my state's expectations about English Language Proficiency development with respect to:

   1. ELP standards?
   2. Trajectory of development?
   3. Time to proficiency?
   4. Reclassification?
   5. Individual student factors that influence growth?
   6. Instructional program factors that influence time to proficiency?

B. What is my state accountability system trying to accomplish by including ELP as an indicator receiving substantial weight?

C. How do I know if some schools are doing a better job with EL students than other schools? How can the new accountability system help me in determining this?

D. Which models should my state consider for the ELP indicator? What tools do I need to effectively communicate these considerations with LEAs, schools, and stakeholders?

E. What are factors that should be considered in making a selection? Am I concerned with:

   1. Familiarity to stakeholders?
   2. Transparency of the model?
   3. Sensitivity to meaningful variation (not losing meaningful variation between students, between schools, between years)?
   4. Ability to take initial ELP level, time to proficiency, and other variables into account?
   5. Ability to optimize N-size (e.g., address reliability/stability of results while minimizing loss of schools that do not meet minimum N-size)?
   6. "Fairness" across grade bands (elementary, middle, high)?
   7. Year-to-year stability of the model in enabling state's accountability goals?
   8. Model consistency with your state's academic achievement indicator approach?

F. What are my state's considerations in choosing *N*-size? Are we concerned with:

   1. Percent of schools with ELs that are included or excluded from accountability for ELP?
   2. Number of years after reclassification that exited EL students can be included in the academic achievement subgroup (allowable for up to 4 years)?
   3. Discrepancy between ELP and academic achievement *N*-sizes that might come about as a result of decisions about (2)?

G. What kind of data modeling will my state consider in moving forward to include ELs in your plan?

CONSIDERATIONS FOR INCORPORATING EL PROGRESS IN TITLE I ACCOUNTABILITY

Monitoring progress of language development is critical to ensure that students have the opportunities for effective language instruction and timely progress towards academic fluency. Making this monitoring part of a state's accountability system, as the ELP indicator, builds on extensive research and evidence that highlights five important aspects that we briefly noted above: (1) the impact of initial ELD level; (2) the grade of entry; (3) the language development trajectory; (4) the adequate level of proficiency for academic content performance; and (5) time to English proficiency. These elements should underpin a state's plan when considering how to measure progress, which background characteristics to include, minimum *N*, and the impact of specific rules around issues like Recently Arrived English Learners (RAEL).[10]

Another important aspect of improving EL outcomes beyond the scope of this brief is the optimal instructional program for reaching proficiency. We note here, however, that by building a reliable, equitable, and unbiased EL progress monitoring system, states will be able to more efficiently examine this issue.

The first step in designing an accountability system is to develop a Theory of Action (ToA). The ToA should align with the state's conceptualization of which elements and processes are related to student performance; which are measurable, and which are malleable. The ToA should be informed by state context specific to EL progress (i.e., growth trajectories, ELD level change rates, distribution of ELs in schools, etc.).

In order to have a coherent accountability system, it is important to determine *a priori* that for which schools ought to be held accountable. This should be accompanied by a general notion of the relative importance of each aspect of the system.

States must decide whether to use a conjunctive, disjunctive, composite, or hybrid system. A conjunctive system requires specific targets to be met, and affords inferences about schools (e.g., the percentage of EL students who demonstrated one ELD level improvement over the course of the year.) A disjunctive system includes specific targets that cannot be missed. A composite system attempts to incorporate multiple important aspects and often provides more flexibility for the state, but results in high-inference interpretations; it is difficult to precisely determine the success factor for a school with a certain composite score because scores reflect multiple aspects of performance (or performance change). A hybrid would combine elements of all three – for example, a school's final rating might be the average of several components, but in order to avoid target status the academic indicator cannot be in the lowest category.

This discussion relates to the portion of the state's accountability plan that monitors EL progress

10 Ways of including RAELs through models allowable under ESSA is the subject of on-going state collaborative work under auspices of the U.S. Department of Education, led by Robert Linquanti and Gary Cook. These provisions are not the subject of this paper. To the extent that states intend to consider options discussed in that work, it is important to consider the coherence among the indicators impacted by ELs.

towards English language proficiency. As demonstrated below, this progress can be measured in a number of different ways. Before, or in conjunction with, selecting a progress indicator, it is critical to decide how EL progress relates to inferences about schools[11]—that is, if progress equals "X," what does that progress say about the school? How does progress impact expectations for schools earning points for classification purposes? For example, if expected progress is that every EL student advances one ELD level over one year in a particular school, and a school meets this expectation, does this result in a classification of "Meets?" How does "Meets" relate to other classification levels? Is this a realistic expectation based on the empirical evidence? Do minimum N and weighting rules mute or exacerbate inferences about certain schools? These questions will all impact the value placed on EL progress indicator for a school's overall classification.

MODELS[12]

There are many modeling options related to monitoring EL progress and this paper does not provide an exhaustive summary. However, using several recent Council of Chief State School Officers (CCSSO) reports and commissioned papers as a guide, we describe the following models that are commonly applied to monitoring schools for Title I accountability and can be used to monitor English language proficiency progress.

Figure 7 below shows how models might be classified. It is important to note that the term "growth models" is applied to a wide variety of models, some of which do not directly evaluate growth. In Figure 6 we divide models into those that provide a direct measure of growth versus those that provide conditional status from which growth is inferred. Model choice should be driven by a state's ToA; that is, how well it aligns with a state's conception of student progress. However, technical issues (and tradeoffs) need to be considered as well. These include bias, transparency, precision, and stability. The next sections briefly describe the models presented in Figure 7. There is no "best" model," particularly when considering the interplay of state context, accountability system, and ToA.

11 It is important to note that the primary focus of these progress models is the ability to make valid inferences about schools, not necessarily about individual student progress.

12 Based on: Goldschmidt, Pete, Kilchan Choi, and J.P. Beaudoin (2012). *Growth Model Comparison Study: Practical Implications of Alternative Models for Evaluating School Performance*, Council of Chief State School Officers, Washington DC.

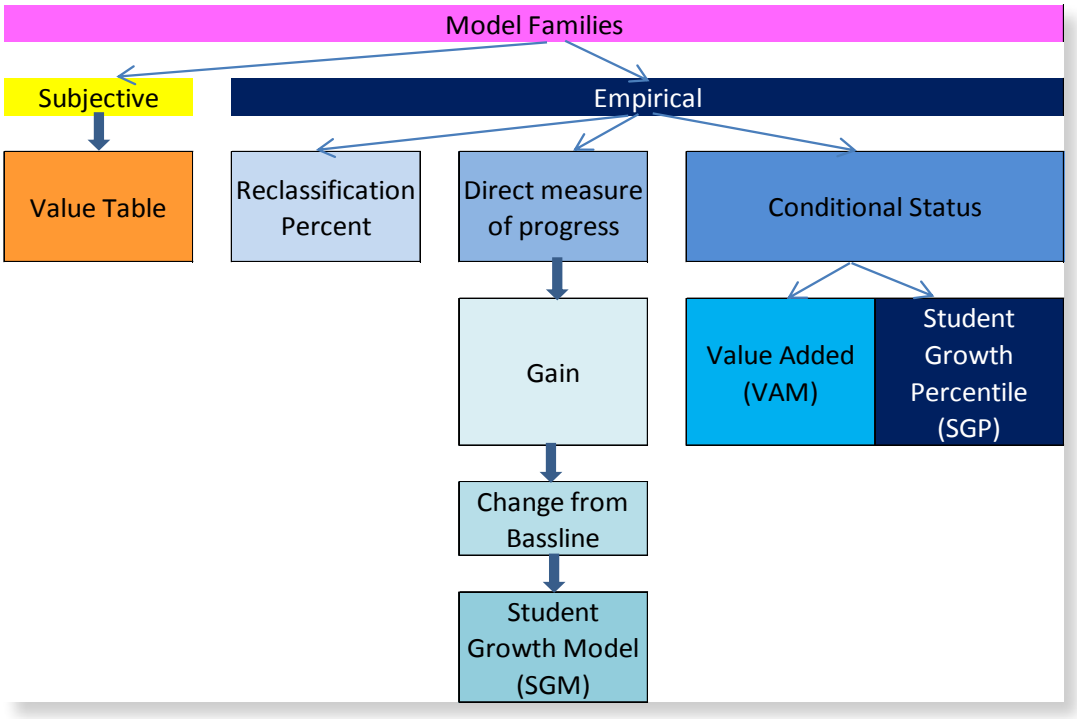


Figure 7: Relationship among Models

Transition Matrix/Value Table

Transition tables or value tables allow states to directly link changes in ELD levels to both school performance and the time it takes to reach English language proficiency.

Table 3: Transition Tables

		Year 1					
	ELD	1	2	3	4	5	6
Year 0	1	0	1	2	3	4	5
	2	-1	0	1	2	3	4
	3	-2	-1	0	1	2	3
	4	-3	-2	-1	0	1	2
	5	-4	-3	-2	-1	0	1
	6	-5	-4	-3	-2	-1	0

Table 3: Transition Tables Cont.

		Year 1					
	ELD	1	2	3	4	5	6
Year 0	1	7	8	9	10	10	10
	2	6	7	8	9	10	10
	3	5	6	7	8	9	10
	4	4	5	6	7	8	9
	5	3	4	5	6	7	8
	6	2	3	4	5	6	7

A Transition table also provides an example of how to consider inferences and alignment with a state's ToA. The top panel of Table 3 presents a simple transition matrix[13] that has a student's ELD level in year 0 on the vertical axis and the same student's subsequent ELD level in year 1 on the horizontal axis. The values in the table show how a student is progressing from year 0 to year 1. The table, based on ELD level increments implies that the state is assuming progression over four years, assuming level 5 representing English language proficiency. States are reluctant to employ accountability systems that take points away (i.e., negative values in the table) and thus appear punitive. The bottom panel of Table 3 eliminates the negative values in the table by adding 7 to each cell and establishes a maximum of 10 points.[14] In this way, a student earns and contributes 8 points each year to the school's performance if she gains one level per year until reaching English language proficiency (at level 5). If this model is in fact based on 10 points and the school adheres to a traditional classification 70, 80, 90 scale of C, B, A, then a school where students on average are gaining a level per year would earn a B.

Additionally, this implies that a school with an average score of 7, or a "C," is facilitated adequate language progress for its students, but students are on average simply maintaining the status quo which is an undesirable outcome. Also, if only F and D schools are candidates for sanctions, then a school in which students on average are not making progress towards proficiency may not be identified. Table 4 presents the same table as the bottom panel of Table 3, but here points for any negative changes in ELD have been eliminated. This may provide scores that better reflect a school's ability to facilitate language progress. However, the transition matrix presented in Table 3 may need further refinement in order to more closely align with a state's ToA.

13 This table reflects linear growth, which is not required under ESSA, and also does not use student characteristics such as initial ELD, for example, which regulations require. Keeping the table simple for the example does not change the major points of this section.

14 We arbitrarily assign 10 points to the EL progress portion of the accountability model. We use 10 because it makes relative comparisons straight-forward.

Table 4: Transformed Transition Table

		Year 1					
	ELD	1	2	3	4	5	6
Year 0	1	7	8	9	10	10	10
	2	0	7	8	9	10	10
	3	0	0	7	8	9	10
	4	0	0	0	7	8	9
	5	0	0	0	0	7	8
	6	0	0	0	0	0	7

Reclassification Rate

Another indicator that can be used to monitor ELs' progress in gaining English language proficiency is the percent of EL students who have been reclassified[15]. This method is highly transparent in that schools with a higher proportion of students who have been reclassified or reached English language proficiency show higher reclassification rates. A state still must consider when students reach proficiency since this method may bias results against elementary schools or middle schools depending on how reclassification takes place in practice.

Other issues may arise with this method. If monitoring is based on the number of ELs at level 4 as the denominator then, a) ELs not at level 4 are not receiving credit for progress towards level 4; and b) a cohort rate using the number of ELs from 4 or 5 years prior as a basis (akin to the 4-year graduate rate calculations) may be limited by higher mobility rates than for EO students, which may make inferences about schools more variable. Graduation rates are based on the 9th grade cohort[16] so a similar approach could be to create grade cohorts (by initial ELD level) and calculate "graduation," or reclassification rates based on these cohorts. For example, for third grade[17] in year 0, a school receives 10 new EL students (5 at ELD level 4, 3 at level 2, and 2 at level 3). For this school, 5 students would be the basis for a reclassification rate calculation in year 1, 3 in year 2, and 2 in year 3[18]. To the extent that students are mobile, this calculation becomes further complicated when a student switches schools/districts before the grade they count towards the reclassification rate. Mobility will have a larger impact when school N sizes are small. To the extent that reclassification is based on multiple factors, these factors may provide incentive to inappropriately reclassify EL students. This method also provides a direct indicator related to the school-level construct of interest, but does not provide much information about the individual student (e.g., exited or not exited).

15 In order for reclassification to be meaningful within a state, it is important to have standardized statewide exit criteria as now required under ESSA, a situation that is not currently in many states. See Linquanti, R. & Cook, H. G. (2013). Toward a "Common Definition of English Learner": Guidance for States and State Assessment Consortia in Defining and Addressing Policy and Technical Issues and Options. Washington, DC: Council of Chief State Schools Officers. Downloaded at http://www.ccsso.org/Resources/Publications/Toward_a_Common_Definition_English_Learner_.html

16 There are some allowable additions and subtractions to the denominator, but in general the 9th grade cohort is the basis for determining a school's graduation rate.

17 This would need to be done for each grade a school serves – which can become cumbersome.

18 This assumes a student progress one ELD level each year.

Simple Gain

A simple gain or change is a direct and transparent measure of student growth, and scores are low inference; the average gain for a school is easily interpreted. A simple gain is calculated for each student and averages are calculated for schools. This approach ignores the clustering of students in schools and thus explicitly ignores school context (Burstein, 1980). A model that ignores the clustering of students within schools and simply aggregates individual student gains up to the school level potentially produces biased estimates of school effects (Willms & Raudenbush, 1989). This occurs because estimates that ignore the fact the students attend specific schools mixes within and between school estimates when the intraclass correlation is greater than zero (Aitkin & Longford, 1986). This applies to any model that is based on individual student scores that are aggregated up to a school. A gain, or change, is simply:

$$\text{Change} = SS_{yr1} - SS_{yr0}$$

SS_{yr1} is the scale score from year 1 and SS_{yr0} is the scale score from year 0. As noted, this provides an unbiased estimate of change, but gains tend to be inversely related to initial performance (due to measurement error in the pre-test); high scorers in year 0 would be related to low gains. Given that growth over time is decelerating, gains from occasion 0 to occasion 1 will be larger than gains from occasion 1 to occasion 2. If linear targets are set, those targets need to consider the natural phenomenon of slower growth over time. Previous research indicates that gains tend to be unstable from one year to the next, resulting in exaggerated variation in school performance from year to year.

However, it is possible to extend simple gains in various ways to improve some of the properties, while retaining the underlying meaning (and transparency). A primary adjustment would be to address measurement error in the pre-test in order to reduce the inverse relationship between the year 0 score and gains by calculating a Kelley True score. A Kelley True score is a reliability adjusted shrinkage towards mean performance.

$$KTSS_{yr0i} = SS_{yr0i}(r) + SS_{yr0}(1-r)$$

Here, the Kelley True score, $KTSS_{yr0i}$, is a function of the observed score in year 0, SS_{yr0i}, and the mean score in year 0, SS_{yr0}. The individual score and the mean score is weighted by r, the reliability of the assessment. More sophisticated options incorporate a conditional standard error of measurement (CSEM), if a classical test theory approach is deemed insufficient. Change is then still calculated as:

$$\text{Change} = SS_{yr1} - KTSS_{yr0}$$

This implies that a school's performance is based on average gains in its students' performance. Score meaning, in its simplest form, means that students have more mastery of given area of content. As noted, in order for gains to be meaningfully interpreted, assessment scores need to be on a vertical scale. However, in some cases researchers have normalized scores (z-scores) within grade levels under the assumption that performance standards are vertically moderated and thus allow for consistent meaning across grades at various anchor points. This approach moves away from an absolute conception of growth to one that considers growth relative to standards. However, if gains are not meaningfully tied to a criterion, then those gains exhibit a "growth to nowhere" phenomenon. Linking gains to a specific endpoint with a fixed time horizon—as in a growth to standard model—also consists of tradeoffs. These are identified below.

As noted above, year over year gains tend to be unstable over time. One potential adjustment that is consistent with the notion that learning is a cumulative process is to treat gains as a change from a baseline measure. Several options exist. Below we consider a simple extension of the gain model presented above and consider a two-year gain.

$$\text{Change} = SS_{yr2} - KTSS_{yr0}$$

In this case a state could use changes for students over two years (and one year change for those with one year of data). Calculating an average annual change using two years remains transparent but improves year to year stability. Thus using a Kelley True score as the baseline and extending gains as change from baseline improves the functionality of gains. One key aspect of using gains is that there can be no missing assessment results on either occasion so gains could be calculated. A Student Growth Model, which provides additional stability, flexibility with handling measurement error, and is robust even if data are missing is another extension of simple gains.

Student Growth Model

A student growth model (SGM) measures growth as a function of time, not as a series of gain scores. This means that gains—or growth—are not calculated by calculating the difference in scores from one year to the next but rather by estimating the relationship between scores and time. For example:

TIME	SCORE
Year 0	100
Year 1	200
Year 2	300

Growth is calculated using regression-based methods. For example, for a single student the link between a growth score and a gain is apparent in that growth is calculated as a slope, $(Score_{yr2} - Score_{yr0}) \ / \ (Year_2 - Year_0)$. Growth = (300-100)/(2-0) = 200/2 = 100 per year.

One big advantage of a SGM is that a student can still be included even with incomplete data. So students who took assessments on 1, 2, or 3 occasions still contribute to a school's average growth. States should determine whether there is a relationship between number of incomplete data points and growth estimates in order to ensure that mobility does not bias results.

An extension of a SGM is to estimate student and school growth using a multilevel—or mixed effects—student growth model. Mixed effects SGMs can incorporate the structure of the data explicitly (considering that students attend specific schools and that students within schools are not independent of one another in terms of the instruction they receive) and mitigate the effects of potential confounding factors (PCF). That is, SGMs tend to reduce the need for additional student background characteristics since prior performance is explicitly considered in the model. States still may choose to include additional student or school information. In our example, we utilize three years of data, but different parametrizations are possible. SGMs theoretically require a vertical scale

for the results to be meaningful, but results can be estimated using normalized scores as long as the focus is comparing schools rather than making inferences about absolute growth (Goldschmidt, Choi, Martinez, and Novak, 2010).

Value Added Model

Fitzmaurice, Laird, &Ware (2004) argue that the choice between analysis of gain scores versus a covariate adjusted model depends on the research question. A covariate adjusted model shows how students differ on the post-test considering their pre-test score. Gain or growth scores show how groups of students (i.e., all students at a school) differ in gains or growth, on average. A covariate adjustment model provides the basis for a value added model. We demonstrate the link between gains and a Value Added Model (VAM) by transforming a gain score into a value added score (model). Starting from the gain score we presented above, we consider the following:

$$\text{Gain} = SS_{yr1} - SS_{yr0}$$

We then replace "Gain" with b:

$$b = SS_{yr1} - SS_{yr0}$$

Without changing any meaning we can add a constant term, γ, to the above equation so that:

$$b = SS_{yr1} - \gamma SS_{yr0}, \text{ where } \gamma = 1.$$

By way of example, if $SS_{yr1} = 30$ and $SS_{yr0} = 20$, then:

$$b = 30 - 1(20) = 10.$$

We can then rewrite the gain score as:

$$b + \gamma SS_{yr0} = SS_{yr1} \text{ (i.e., } 10 + (1)20 = 30).$$

And we rearrange the equation so that:

$$SS_{yr1} = b + \gamma SS_{yr0} \text{ (i.e., } 30 = 10 + 20).$$

Next we take away the restriction that $\gamma = 1$ and the value of γ becomes an estimate based on the data. Finally, because we are estimating this relationship for many students and fitting a line through the data which likely vary, we add an error, or residual, term e:

$$SS_{yr1} = b + \gamma SS_{yr0} + e$$

This is a basic Value Added Model. The advantage of a VAM over a simple gain model is that more variables can be added to the right hand side of the equation (e.g., additional prior test scores, ELD level, student background, etc.). The specifics of the model depend on what the state deems important based on its ToA.

One advantage of a VAM is that it is more robust when used with either vertical or non-vertical scales than a SGM or gain scores. As noted, VAMs do not provide results in terms of growth; rather they address current student performance explicitly accounting for differences in initial performance.

Student Growth Percentiles

Similar to a VAM, Student Growth Percentile models (SGPs) provide results that infer current performance based on past prior performance, or conditional status. A SGP model is similar to a VAM if we consider that a VAM estimates a single line through the data (not necessarily linear), whereas a SGP model estimates the same model 99 times—one for each percentile of the distribution of scores[19]. The specific model in which a student is included depends on which percentile the student's score falls into. The SGP broadens the notion of robustness to scale by focusing on normative position based on student percentile ranks. The SGP model is fully detailed in Betebenner (2009). The SGP model uses quantile regression to measure a student's progress from one year to the next relative to his/her academic peers with similar test score histories[20].

To obtain school level SGP estimates, student-level growth percentile scores are aggregated to higher units of measure. After SGPs are estimated at the student level, it is quite simple to combine them into higher-level aggregates. Usually, the median or mean of the SGP distribution for the school is used to summarize school-level SGP as a single number, the median or mean of the SGP distribution for the school is usually used. The aggregate SGP represents the growth of a "typical" student in a given school.

Advantages of the SGP approach include its robustness to scale requirements and that the normative interpretation of student growth from one year to next is very understandable to a broad array of stakeholders. It is relatively easy to aggregate obtained student growth percentiles to higher units (e.g., teachers and schools). As previously noted, school effects estimated as simple aggregates confound within-school and between-school effects.

One disadvantage to the SGP model is that it requires a substantial amount of data in order to generate sufficient coverage across the percentiles. For English language content at the state level this is not an issue. However, for small populations of ELs in some states this may affect the robustness of estimates. As noted, it is possible to estimate SGP using the percentile rank of the residual (PRR) estimated with an OLS model, which may be a good alternative for states with small EL populations.

Summary of Student Growth, Value Added, and Student Growth Percentile Models

These models allow states flexibility to utilize more data and model progress over time, providing more stability and robustness of results, compared to the simple process of using a current and a

19 The specification of a SGP model generally tries to capture the non-linearities when modeling ordinal data and includes additional components.

20 This estimation is quite distinct from OLS (Ordinary Least Squares) regression, although it is possible to use OLS to estimate SGPs (Castellano and Ho, 2013).

prior score. Generally, including additional data increases the robustness of results as well as stability over time. However, VAM and SGP are both conditional status models, meaning that they estimate where a student's score is expected to be given her prior performance. *Only a Student Growth Model specifically estimates growth over time.* We reiterate that the SGM provides a direct indicator of growth, but in order to make full use of SGM results, the underlying assessment must be on a vertical scale. VAM and SGP are more robust to scale because they are estimating an endpoint conditioned on prior performance. This also implies that VAM and SGP are more robust to assessment system changes as both can readily incorporate prior results based on different (or multiple) prior assessment sources. Another feature that may be important is that SGMs and VAMs are more readily amenable to explicitly accounting for the non-independent clustering of students within schools.

Growth to Standard

We address a Growth to Standard (GTS) model separately to highlight potential limitations of this approach if applied without consideration of the tradeoffs. GTS is relatively transparent, easy to calculate, eliminates the "growth to nowhere" concern, and meets the ESSA requirement of providing a fixed timeline towards English language proficiency. A simple GTS model would compare a student's actual gain to a criterion gain required to meet a specific target. For example, over a five-year period, the criterion would be:

$$C = (\text{Target Score} - \text{Year}_0 \text{ Score})/5$$

We see that a student who started off earning a 200 score, but needs to get to a score of 350 for English language proficiency needs to gain 150/5, or 30 points per year. A student's gain in year 1 would be calculated in the same way as we lay out in the *Gain Score* section, and then compared to C above. We then aggregate this information by counting the gains greater than C (or the proportion of all gains greater than C). This model explicitly affords inferences about whether a student gained "enough". Another common alternative is to count students having demonstrated a one (or partial) ELD level change. School ranking based on this model will likely be correlated to status models as the results depend entirely on whether the current year score is sufficient to pass the set threshold. The difference between this model and the general Annual Yearly Progress (AYP) model is that each student has an individual threshold.

In a simple example that compares a GTS and SGM, we present data on 5th grade students who are projected to meet English language proficiency requirements by 8th grade (originally on a 5 year horizon from grade 3). Table 5 summarizes student growth as well as the growth required to meet the standard (note that there are some students whose last score was above the standard so the required minimum is negative).

Table 5: Descriptive Statistics for Growth to Standard

Descriptive Statistics				
	Minimum	Maximum	Mean	Std. Deviation
Ave Student Growth	-12.4	94.7	23.6	10.2
Ave Growth Req'd	-20.0	83.3	14.5	13.9

For this sample, students are improving an average of 23.6 points per year; the required growth to reach English language proficiency by 8th grade is 14.5 points. Figure 8 summarizes the relationship for students between observed growth and the required growth, C. Students above the diagonal line demonstrate growth greater than C, while students below the diagonal line are not demonstrating sufficient growth.

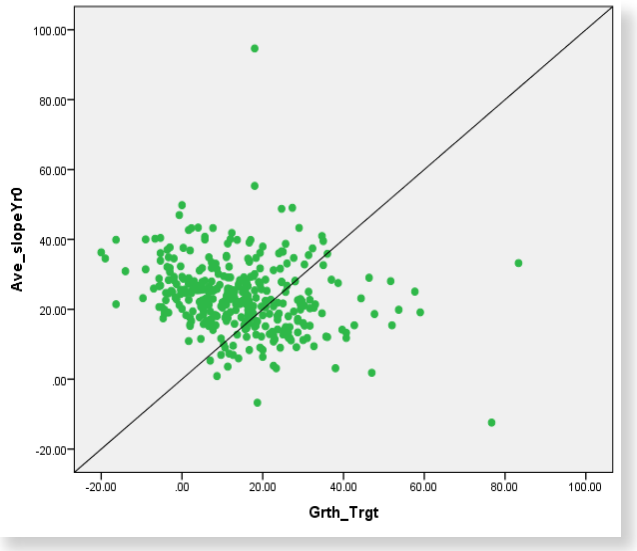


Figure 8: Comparison of actual and required growth to meet GTS standard

The underlying assumption is that the observed growth will continue linearly until 8th grade. But we know that that English language proficiency growth is not linear. So, depending on how the target was set, it may result in students only meeting the target in year 1. If students miss the target in year 1, it is unlikely that they will meet the target in subsequent years. This can result in an unintended incentive for a school to focus on "bubble kids" (students close to meeting specific thresholds).

Figure 9 summarizes the scoring under the GTS model. Each student in the figure is either a 1 or a 0, corresponding to whether they fall above or below the diagonal line in Figure 8.

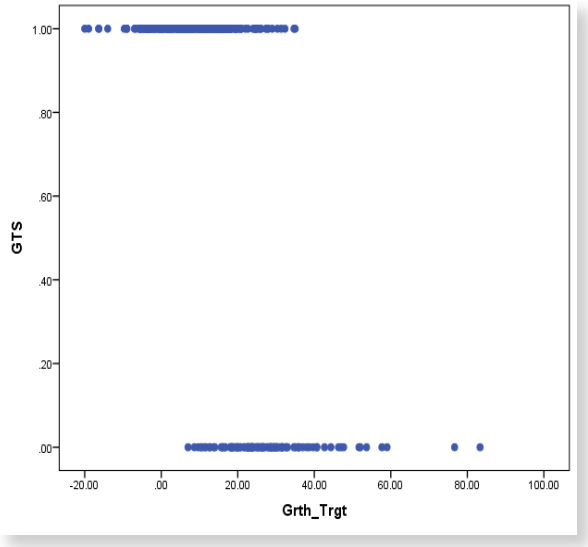


Figure 9: Students contributing 1 or 0 to a school using a GTS model

One difference between the results displayed in Figures 8 and 9 is the amount of information provided. This is highlighted in Figure 10.

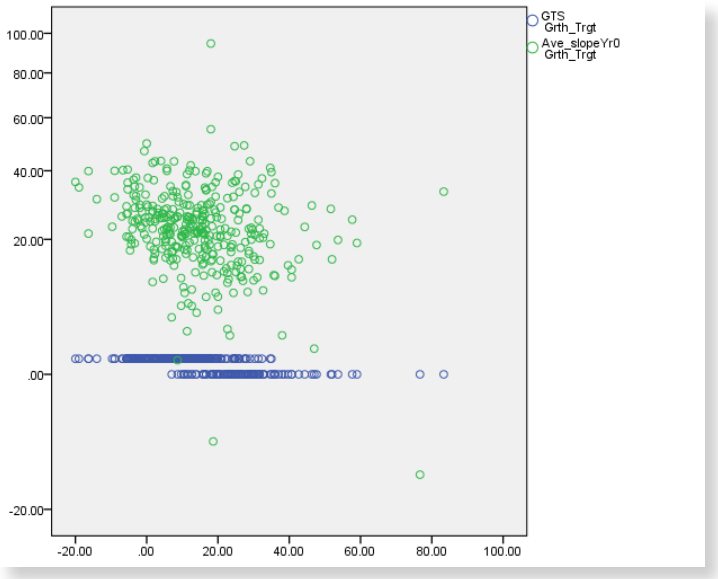


Figure 10: Comparing Growth and Growth to Standard

Figure 10 presents the same students twice—once with their actual growth and once as 0/1 indicating whether that growth was sufficient to meet the target. As noted, counting students contribution as 0 or 1 is a similar construct to how NCLB prescribed data to be presented, creating incentives to focus on "bubble kids."

This binary representation also omits information about actual student growth. For example, the student score at the right-most point at 0 (on the vertical axis) does not count towards meeting the standard, but this student's growth—represented by the data point vertically above the one referenced, indicates that this student had very high growth for the year. Under a GTS model, this students' growth does not count. If we again move along the points at 0 (vertical axis) and look at the second to last point on the right, we see that this student's observed growth was negative (vertically below the referenced point). A substantial amount of information is lost when students are simply counted as meeting or not meeting a target, as opposed to utilizing actual growth, which GTS is generally not designed to monitor.

As noted above, GTS results are primarily driven by the current assessment results, meaning that school-level GTS results are more closely related to status than other growth models. Table 6 presents the *R*2 between ELP status, GTS results, and growth results by school. The results in Table 6 indicate that GTS results are strongly related to ELP status, implying that monitoring progress based on GTS is essentially monitoring current ELP status

Table 6: Proportion of Model Results Attributable to Status

Proportion of	Accounted for by ELP status
Growth to Standard	0.61
Ave. Student Growth	0.21

Figure 11 presents the relationship between growth on the ELP assessment, points earned per student (a school-level average), and the mean growth required for this sample of students to reach English language proficiency by 8th grade (the vertical line). Rather than simply counting students as 0 or 1, points can be assigned to correspond with a "B" if students on average are meeting their GTS growth. Here again, a state's ToA should determine appropriate cut scores. The key to this approach is that each student contributes some amount to a school's rating, points are proportional to success, and information regarding actual growth is not lost by focusing on the yes/no dichotomy.

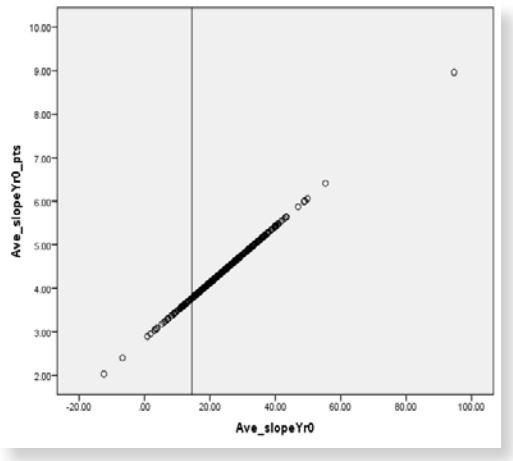


Figure 11: Basing Growth on the GTS Model

MODEL RESULTS

We will now apply data from two states to the growth models described above. We first present results related to monitoring student progress towards English language proficiency, followed by results that include exited students in the EL subgroup to demonstrate the impact on EL subgroup performance on the ELA and mathematics content assessments of a state accountability system.

Empirical Results for Monitoring English Language Progress

We begin by summarizing model results in terms of points earned, using each model to generate scores for each school. These results are presented in Table 7. Comparisons of raw points across models are not warranted since it would be possible to shift points earned by using different scaling methods[21]. The results in Table 7 examine how model results compare across school levels (i.e., elementary, middle, and high school).

Table 7: Model Results

School Level		Value Table	Gain 1	Gain 2	Reclass %	SGP	VAM	SGM
Elem	Mean	3.07	4.05	6.10	1.44	5.24	4.66	4.76
	N school	229	227	224	232	227	226	228
	SD	1.52	1.46	1.38	1.20	1.36	1.06	0.53
Middle	Mean	1.42	1.59	2.13	1.08	4.00	4.05	3.62
	N school	78	78	76	78	78	76	76
	SD	0.90	1.04	0.86	0.78	1.02	0.76	0.16
High	Mean	3.39	3.19	3.18	1.75	5.54	6.70	3.57
	N school	80	80	78	84	80	78	81
	SD	1.85	1.85	1.20	1.12	1.56	1.81	0.27

Given our evidence that time in the system impacts ELD outcomes, that growth is not linear, and that initial ELD level is related to later ELD levels, we would monitor whether the chosen model impacts one school level more than another due to the nature of growth, as opposed to meaningful differences among schools in the quality of English language development programs. Figure 12 presents the proportion of elementary school points earned by middle and high schools using data from both State 1 and State 2. Figure 12 indicates that a value table would result in significantly fewer points earned by middle schools in both states and by high schools in State 2, when compared to points earned by elementary schools. For example, a middle school in State 1 earns about 45% of the points of an elementary school when applying a value table. There could be several explanations for why high schools earn slightly more points than elementary schools in State; high schools earn points depending on the number of new arrivals and students who have not yet reached proficiency.

21 Gain, SGM, VAM, and SGP models were all scaled using HOSS and LOSS of 10 and 0 respectively.

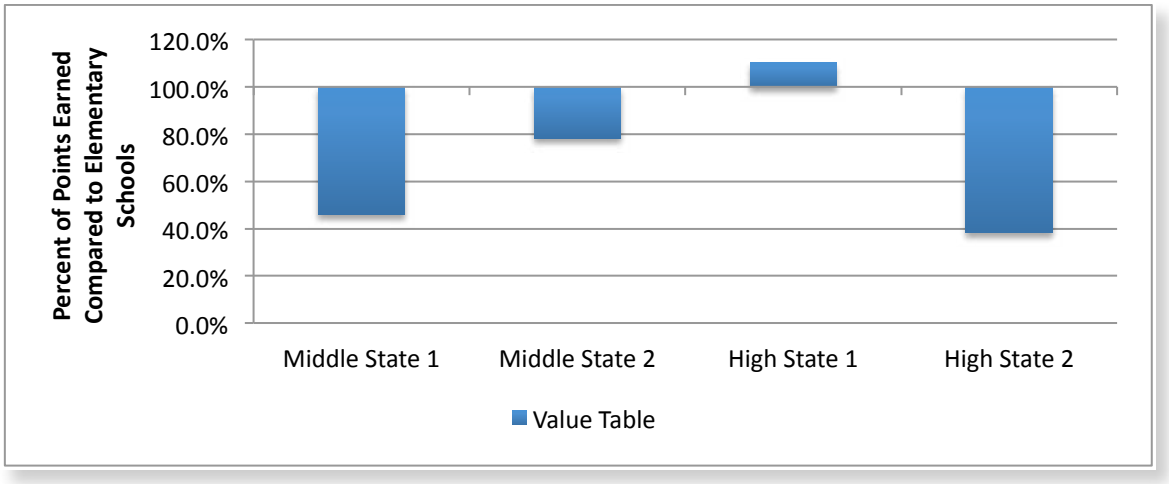


Figure 12: Percent of Elementary School Points Earned Using a Value Table

The number of new arrivals may influence results if the population is large enough and they progress more quickly. Additionally, a larger number of new arrivals likely increase the incentive for high schools to improve language proficiency rates so students can pass ELA and Mathematics assessments for graduation.

The similarities and differences in results among states confirm that each state must examine its own data. The results we present are meant to identify options to consider, provide some guidance around the interpretation of results, and highlight the potential variability of results that may occur across state contexts.

Figure 13 repeats the results from Figure 12, but includes results for all of the models we examine. Generally, middle and high schools tend to earn fewer points than elementary schools. This may in fact be consistent with expectations in that we previously demonstrated that EL growth was greater initially and slows over time. To the extent that more EL students enroll in a state in elementary school, growth results would be better for elementary schools than middle or high schools. However, context and policy play an important role. For example, we see that reclassifications in State 2 seem to occur most frequently in middle school. If this is by design, then the accountability model should reflect this.

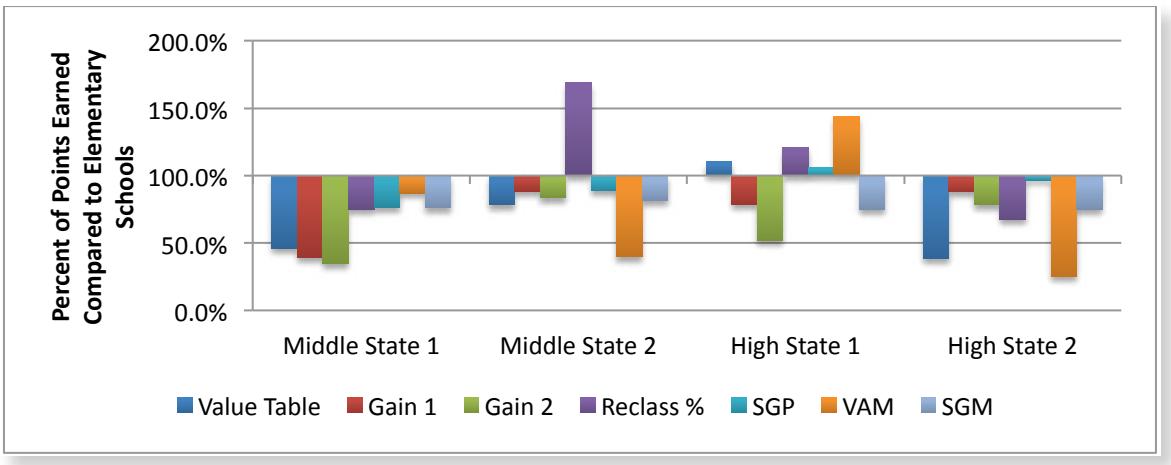


Figure 13: Percent of Elementary School Points Earned

Another aspect to consider is the initial ELD level and enrolled grade ELs have when they enter the system. For example, in State 1, about 78% of ELs enter in elementary school. Initial ELD levels are generally level 3 or higher, as this includes about 75% of students[22]. For State 1 we see students who enter later (in middle or high school) tend to have the same initial ELD levels as those entering in elementary school. For EL students entering in middle and high schools, states must carefully balance their ambitions with outcomes that are attainable. A state's distribution of initial grades and ELD levels will influence how model results vary by school level.

One element that impacts all models is the effect of minimum *N* on the inclusion of schools for monitoring. Given that ESSA shifts EL progress accountability from the district level to the school level, a state's minimum *N* will have a substantively larger impact on whether students (through the schools they are attending) will be included in the system. Figure 14 presents the relationship between minimum *N* and the proportion of schools held accountable for EL progress. The basis is the number of schools with at least one EL student. So, in State 1, applying a minimum *N* of 10, about 75% of possible elementary schools are included. The pattern displayed in Figure 14 is consistent with expectations in that larger minimum *N*s significantly reduce the percent of schools held accountable for EL progress. We note that in Figure 14 for State 2, the elementary and middle school lines are both at 100% because all elementary and middle schools have at least 40 EL students. A minimum *N* of 40 can result in less than 20% of schools participating in EL progress accountability.

Decreasing the minimum *N* unequivocally increases the number of schools held accountable for EL student progress. States should consider reducing the minimum *N* in conjunction with the desired model and the impact *N* has on how precisely their model functions, and make inferences about the school's ability to facilitate EL students' English language development. Minimum *N* won't impact the ability to accurately represent school processes, since inferences are based on the population of EL students in the school. If a school is serving a small number of EL students, inferences show how well a school serves a small number of ELs, and has no bearing on how well a school would serve a larger number of ELs. Minimum N must also be considered in conjunction with weighting – both in terms of the influence a single student can have on overall inferences with respect to EL progress, and the weight assigned to the EL progress portion of the accountability system. These issues are not only relevant in composite systems, but also in conjunctive or disjunctive systems.

One method for increasing EL *N* in a school is to use multiple years of data. However this approach decreases the model's ability to capture a school's current change in performance, and counts students multiple times for the same indicator (some students will count twice while others will count three times). This artificially reduces the standard deviation and conflates the variation between schools with the variation within schools over time.

22 Except those entering in Kindergarten where almost 30% are at ELD level 1.

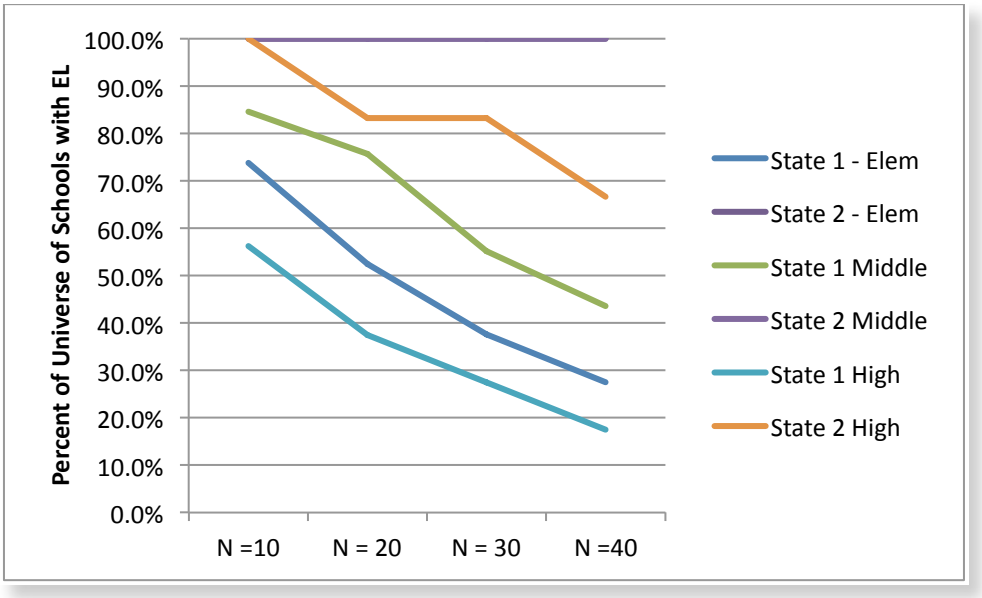


Figure 14: Impact of Minimum N on inclusion in the EL Proggress Component

Table 8 summarizes the impact of minimum *N* by model. Again, the absolute number of points earned is not relevant. What is relevant is the variation across minimum *N*s and the standard deviation of the scores. The results indicate that under a gain model the mean score for schools with fewer than 10 EL students is 2.99, while the mean score for schools with at least 40 EL students is 3.28. We are also able to compare the range of scores. Overall there are not significant differences in points earned by large schools vs. smaller schools (this is also presented as a correlation in Tables 9a and 9b.) but it is important to remember that results vary by model and potentially by state. A key consideration is that the inferences we can make here are based on empirical results and depend on two aspects of research design that address internal and external invalidity. Generally, *N* size does not unequivocally eliminate potential confounding factors, nor does it produce better external validity. A larger *N* helps address some issues, but simply increasing *N* does not resolve the limitations in design. Remember that while increased *N* has benefits in terms of increasing precision, and increasing *N* using multiple years' data may be more representative of school process because it takes a more complete snapshot and is not unduly influenced by an atypical year, it also confounds the school's current year effectiveness by resampling the same students.

Table 8: English Language Proggress Points by Model and Minimum N

Minimum N	Value Table	Gain	Gain2	Reclass %	SGP	SGM	VAM
N < 10	2.72	2.99	3.19	0.17	4.91	5.33	6.90
- 1.5 SD	0.00	0.00	0.00	0.00	1.80	2.19	4.55
+ 1.5 SD	6.14	6.61	6.66	0.43	8.03	8.48	9.24
N =10	2.98	3.64	3.63	0.14	5.13	6.06	6.93
- 1.5 SD	0.97	1.52	1.47	0.00	3.34	2.79	5.38
+ 1.5 SD	4.98	5.76	5.79	0.29	6.93	9.32	8.48
N = 20	2.82	3.31	3.38	0.15	5.08	5.59	6.88
- 1.5 SD	0.62	0.86	1.03	0.00	3.23	1.91	5.47
+ 1.5 SD	5.02	5.77	5.73	0.33	6.92	9.26	8.29
N = 30	2.64	3.31	3.38	0.12	5.09	5.50	6.75
- 1.5 SD	1.28	1.52	1.57	0.00	3.84	2.15	5.49
+ 1.5 SD	3.99	5.09	5.20	0.25	6.34	8.85	8.01
N = 40	2.58	3.28	3.18	0.10	4.94	5.77	6.80
- 1.5 SD	1.01	1.39	1.21	0.01	3.75	1.58	5.84
+ 1.5 SD	4.15	5.17	5.16	0.18	6.12	9.97	7.77

This may be a state's conception of holding schools accountable for EL progress, but while these results are also valid for accountability they provide no ability to generalize beyond the current year and the current population, in which case absolute *N* is less critical[23]. One argument against this notion is that school results may be heavily dependent on the preparedness of the students attending in a particular year. That is, a school that enrolls 5 new ELD level 4 students (out of 7), would have an advantage over a school that happens to enroll 5 new ELD level 1 students (out of a total of 7). While this is certainly true of status measures, it is not true for the EL progress measure that intends to measure progress over time (and can explicitly incorporate initial ELD in the model). The sensitivity of results to minimum *N* across years (i.e., stability) varies by model. This is presented in more detail in Figure 15.

Tables 9a and 9b present the proportion of shared variation in model results and various school level characteristics. A reasonable rule of thumb is that shared variation between results and school factors is less than .05 (i.e., 95% of the variation is not shared). Values above .05 do not necessarily imply that the model does not work, but that additional analysis is required. The results in Table 9a are likely more robust because they are statewide while the results in 9b are based on one district data. As noted, the number of ELs in a school tends to be unrelated to model results. One advantage of using statistical models for accountability is that states can incorporate various adjustments to address potential concerns related to the impact of school level factors[24]. It is beyond the scope of this paper to present the various approaches to addressing potential unwanted relationships identified in results, but it is important to note that there are mechanisms that do not require the use of the variable itself be included

23 This is the same discussions that took place during the initial implementation of NCLB with respect to the use of confidence intervals.

24 This notion applies to addressing school-level factors that are beyond the control of the school such as the number of ELs or the percentage of SWD attending the school.

in the model. For example, if there is a possible relationship between the percent of students who are classified as Students with Disabilities (SWD) and school model results, it is possible to address this without having to include SWD (or percent SWD) in the model[25]. In general, including prior achievement captures much of the variation associated with student background. Models that include several years' (and/or subjects') worth of data generally perform help reduce relationships between model results and student background characteristics. A state should check results carefully to determine whether they are consistent with intentions and whether the model needs to be adjusted.

Table 9a: Shared Variation in Model Results and School Characteristics – State 1

State 1: Proportion of Variation in Model Results Shared with Student Background								
Model		Value Table	Gain 1	Gain 2	Reclass %	SGP	VAM	SGM
State ELA SS		0.03	0.03	0.03	0.12	0.01	0.01	0.00
State Math SS		0.03	0.05	0.08	0.12	0.01	0.00	0.04
Pct Prof ELA		0.02	0.02	0.03	0.12	0.00	0.00	0.01
Pct Prof Math		0.03	0.04	0.09	0.10	0.01	0.00	0.06
Pct SWD		0.00	0.02	0.05	0.01	0.00	0.02	0.08
Pct FRL		0.00	0.00	0.00	0.11	0.00	0.02	0.02
Pct EO		0.01	0.03	0.05	0.04	0.01	0.00	0.08
Number of EL		0.01	0.01	0.01	0.07	0.00	0.02	0.00

Table 9b: Shared Variation in Model Results and School Characteristics – State 2

State 2: Proportion of Variation in Model Results Shared with Student Background								
Model		Value Table	Gain 1	Gain 2	Reclass %	SGP	VAM	SGM
State ELA SS		0.45	0.19	0.30	0.61	0.01	0.19	0.26
State Math SS		0.69	0.63	0.73	0.22	0.02	0.47	0.59
Pct Prof ELA		0.46	0.19	0.31	0.58	0.03	0.20	0.26
Pct Prof Math		0.70	0.69	0.78	0.17	0.03	0.48	0.66
Pct SWD		0.05	0.00	0.03	0.15	0.02	0.10	0.00
Pct FRL		0.01	0.16	0.16	0.02	0.06	0.01	0.09
Pct EO		0.00	0.23	0.19	0.00	0.21	0.01	0.16
Number of EL		0.05	0.00	0.00	0.15	0.00	0.13	0.00

Another important aspect to consider is whether results are stable over time. We would assume that school processes are not so volatile that there are big swings in effectiveness from one year to the next. Previous research indicates that over time status is very stable; this stability depends on the stability of the background characteristics of students (Choi, Goldschmidt, and Yamashiro, 2005). In

25 The ESSA accountability regulations (11/29/16) do not permit the use of disability status as one of the student background characteristics for the ELP goal or indicator in accountability models. As outlines in the text, alternatives to address the relationship between SWD and accountability results should be considered.

examining stability we would not expect scores to have a correlation of 1 from year to year[26]. Table 10 presents results from State 1 and State 2. The potential variability in results is highlighted by looking at the data for State 2, which are from a single district. District results are likely supported by intra-district student, program, and process consistency. We would expect Gain 2 (based on a two year change from baseline) to be more stable than a simple year to year gain (Gain 1). VAM and SGP results are moderately correlated over time—one benefit of these models is that they can be parameterized in different ways to improve stability.

It is worth noting that any translation of model results into points or categories will attenuate the correlations.

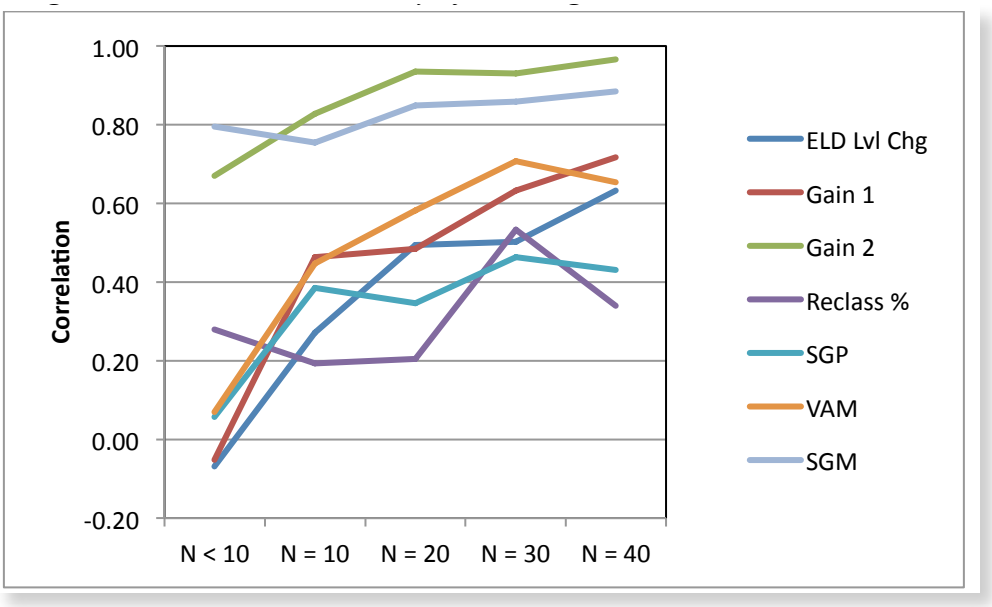
Table 10: Stability of School Level Results over Time

Model	State 1	State 2
Value Table	0.17	0.82
Gain 1	0.25	0.80
Gain 2	0.86	0.95
SGM	0.77	0.95
Reclass %	0.27	0.93
SGP*	0.10	0.27
VAM	0.42	0.49

*Previous multistate studies find this correlation to range from .32 to .46.

Figure 15 presents stability estimates for the EL progress models by minimum *N*. Here we see that model stability varies and increasing *N* generally increases stability, but we also see that models that directly measure growth are less susceptible to variability when minimum *N* changes. Conditional status models (SGP and VAM) are sensitive to *N* sizes, while Gain 2 (two year gains) and the SGM are robust to *N* sizes. This is consistent with previous research on academic content performance (Goldschmidt, et al., 2012), where Gain 2 and SGM appear to be particularly stable, likely due to the nature of the vertical scale on which gains and growth are based.

26 Year to year variability in scores reflect many sources of variability including actual year to year variability effectiveness.



Fiugre 15: Year to Year Stability of EL Progress Model Results

Table 11 presents the correlations among model results. This allows states to examine whether inferences about schools are sensitive to the selected model. Even if a chosen model's results are not highly correlated with other model results, that model may align well with a state's ToA and thus is an appropriate choice. However, a state should have a strong rationale to support that choice. Table 11 shows that models that intend to capture the same phenomenon based on the same data are moderately to highly correlated. This is consistent with expectations. For example, the Gain 2 model and the SGM are both capturing changes in scores over multiple years, and we would expect that correlation to be high. Previous research indicates that conditional status models will be highly correlated (Castellano and Ho, 2015). Our results are influenced by the preponderance of small *N*, which impacts model performance. Some models show results that are not well aligned with the other models because the first model is not directly related to the magnitude in the change of performance from one year to the next.

Table 11:Correlations among Model Results

State 1						
	Gain 1	Gain 2	Reclass %	SGP	VAM	SGM
Value Table	0.62	0.72	0.25	0.02	0.70	0.72
Gain 1		0.97	-0.06	0.53	0.65	0.96
Gain 2			0.07	0.42	0.68	1.00
Reclass %				-0.35	0.01	0.08
SGP					0.27	0.41
VAM						0.67
State 2 (District)						
	Gain 1	Gain 2	Reclass %	SGP	VAM	SGM
Value Table	0.60	0.37	0.26	0.70	0.12	0.09
Gain 1		0.60	0.39	0.82	0.06	0.26
Gain 2			0.16	0.27	0.04	0.86
Reclass %				0.45	0.08	-0.13
SGP					0.22	-0.06
VAM						-0.05

Monitoring EL Performance on Reading/Language Arts and Mathematics Assessments Administered in English[27]

ESSA allows states to include reclassified EL (REL) students in the EL subgroup for up to four years after reclassification. Figures 16–18 summarize the effect of including REL students in the EL subgroup on the percent of schools held accountable for EL students by minimum *N*.

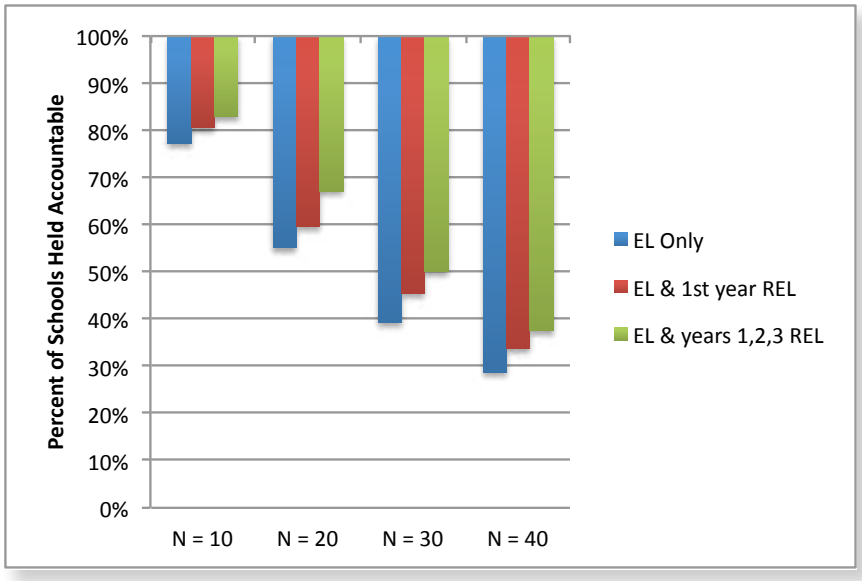


Figure 16: Effect of Including REL in EL Subgroup - Elementary

27 The results presented are based on the state mathematics assessment administered in English. Results based on ELA are substantially similar.

The blue column in Figure 16 indicates what proportion of schools would be included in the accountability system for the EL subgroup on the state mathematics assessment administered in English. A state using an *N* of 10 would include a little less than 80% of all elementary schools with any EL students. Including reclassified students in the EL subgroup for ELA and mathematic assessments administered in English (status or growth) increases the *N* count of the EL subgroup and thus may increase the number of schools included for the EL subgroup. The red column shows the percent of schools included if only students reclassified the previous year are added back into the EL subgroup. The green column represents the impact of adding in students who exited 1, 2, and 3 years prior. Including these students increases the percentage of included schools to about 82% of the total elementary schools with any ELs when a state applies a minimum *N* of 10.

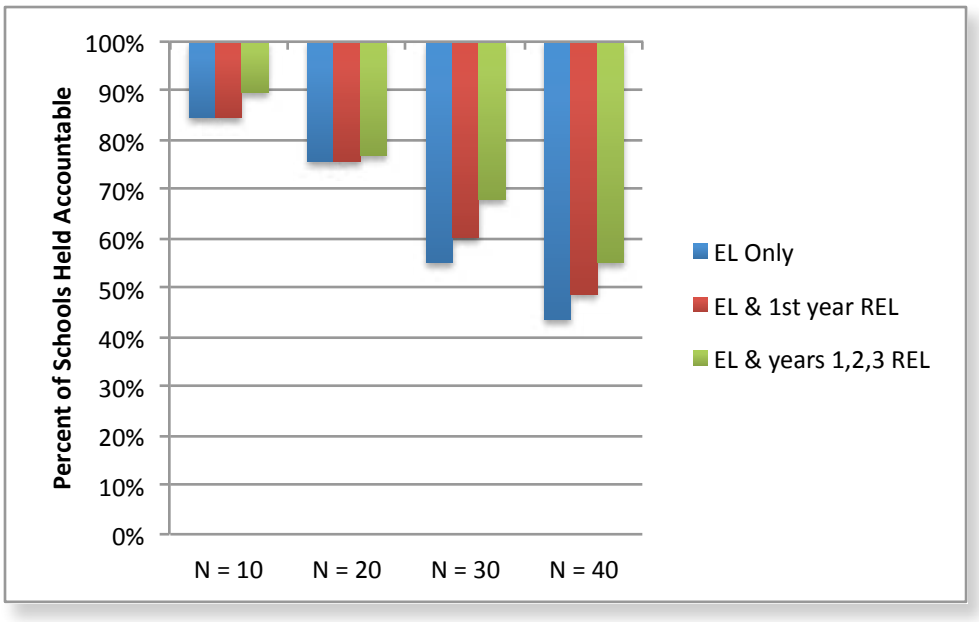


Figure 17: Effect of Including REL in EL Subgroup - Middle School

Figures 16–18 show what happens when we include students who were reclassified one, two, and three years prior. The results indicate that the overall effect of including REL in the EL subgroup is less impactful than the effect of minimum *N*. The impact does vary by school level, with the largest effect in high school—particularly for a minimum *N* of 40. Including as many RELs as possible increases the percentage of high schools in a manner similar to reducing the minimum *N* by 10 students. In high school this is also true for *N* = 30.

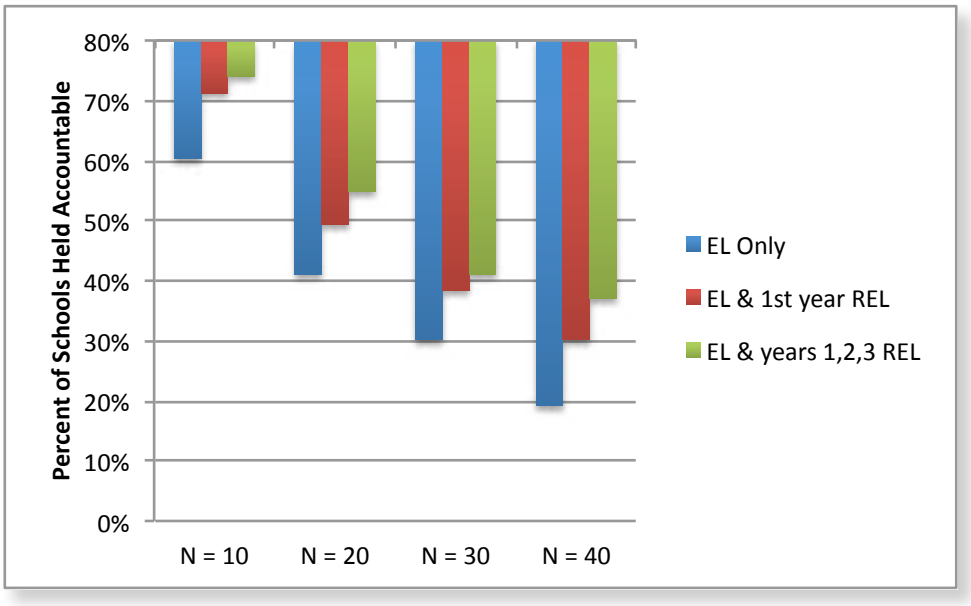


Figure 18: Effect of Including REL in EL Subgroup - High School

Figures 19–21 present the impact of including REL students in the EL subgroup on status (percent proficient) for elementary, middle, and high schools, respectively. The results indicate that there is a consistent pattern across all school levels. Although the absolute level of performance remains low, the average percent proficient increases.

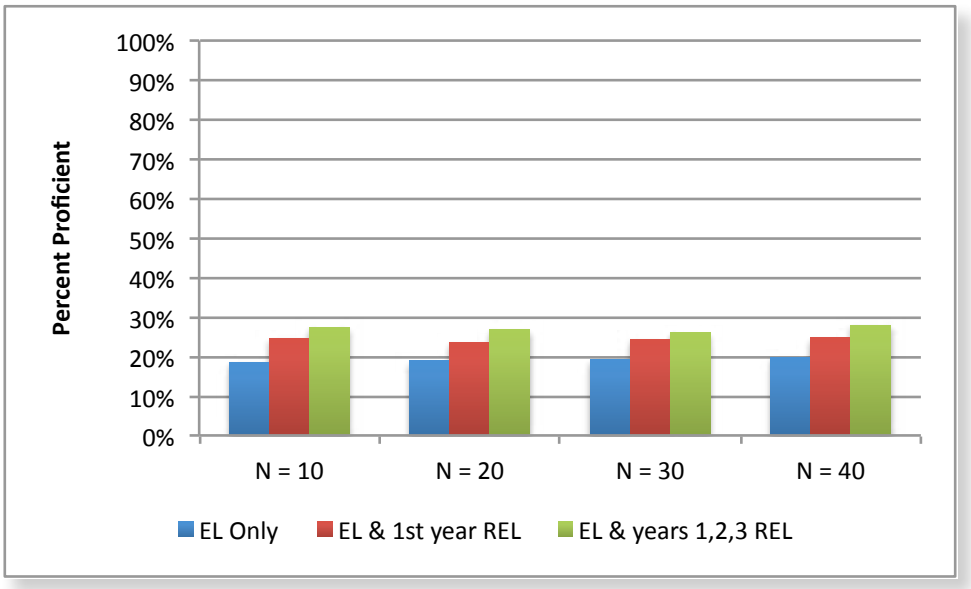


Figure 19: Impact of Including REL in the EL Subgroup on Percent Proficient - Elementary

Figures 19–21 are interpreted similarly to Figures 16–18. The first (blue) column represents the percent of the EL subgroup that are proficient while the second column (red) indicates the percent proficient when the sample includes students reclassified the previous year. The third (green) column indicates the percent proficient when the EL subgroup includes students reclassified 1, 2, and 3 years prior.

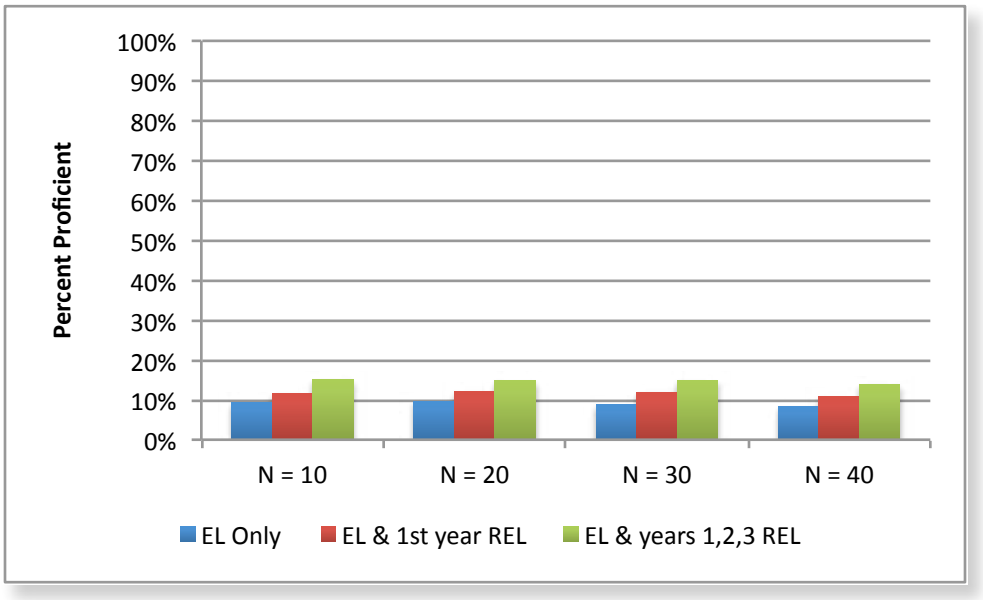


Figure 20: Impact of Including REL in the EL Subgroup on Percent Proficient - Middle

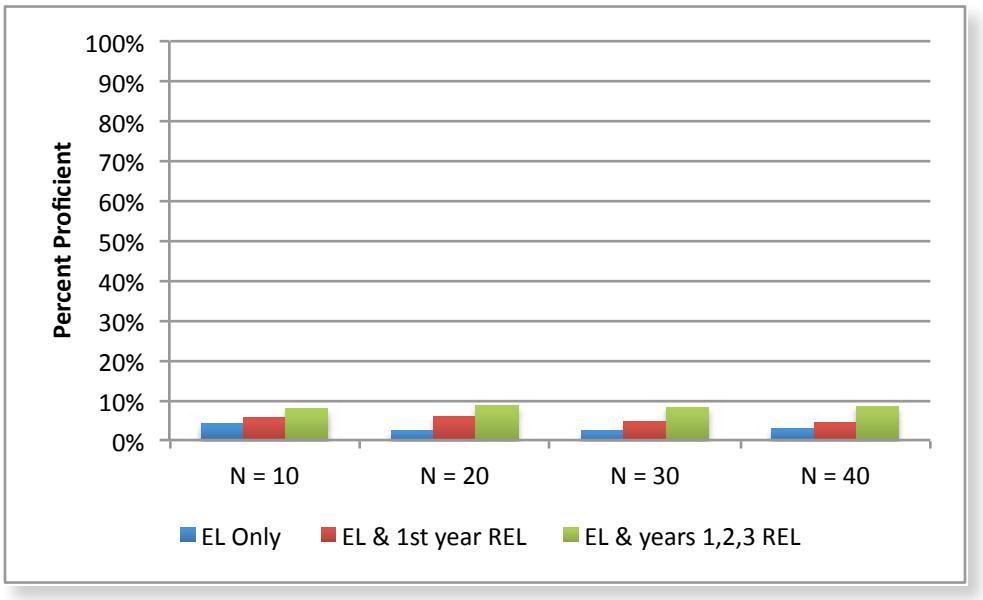


Figure 21: Impact of Including REL in the EL Subgroup on Percent Proficient - High

When considering including REL students in the EL subgroup for accountability, it is important to consider whether performance of EL students is subject to a "masking effect" (or similarly whether REL student performance is masked by keeping REL students in the EO subgroup[28]). The potential impact of "masking" depends on three factors: (1) the gap between EL and REL performance; (2) the ratio of REL to EL students; and (3) the performance of the EL subgroup[29]. The impact specifically on RELs within the

28 Results should be reported separately for each group regardless of how accountability is calculated.

29 Performance can be presented either as scale scores or proficiency.

EO subgroup depends on ratio of EOs to RELs. Assuming that 100% of RELs are English proficient, the reported EO/REL proficiency rates will depend on the EO proficiency rate. The equation (N_{REL}/(N_{REL} + N_{EO})) equals the ratio of actual REL performance difference from 100% proficient to the group reported percent proficient. For example, if in year 0 there are 90 EO students and 10 REL students, then 10/(10 +90) = .10. This value indicates that if the REL proficiency rate decreases by 10 percentage points, the EO/REL group reported rate will only decrease by one percentage point[30]. The actual decrease depends on the proficiency rate of the EO students. This means that the relative impact of monitoring REL students in the EO subgroup grows as performance of the EO students decreases.

The other aspect of masking is "hiding" the performance of EL students by including successful REL students. A state's ToA should drive how RELs are included. In general, there are two conceptions to the decision of excluding or including REL students in the EL subgroup. One is to give credit to program success by not attenuating results by eliminating the most successful students from proficiency calculations. The other is to be concerned about "masking" EL performance by keeping REL students in the group calculations. Two factors impact the effect of including RELs in the EL subgroup: the gap in performance between RELs and ELs and the ratio of RELs to ELs. So, we see that (the REL-EL performance gap)*(N_{REL}/(N_{EL}+ N_{REL})) = amount the EL/REL subgroup will increase due the inclusion of RELs in the EO subgroup. But, the actual reported amount depends on the proficiency rate of ELs. For example, if the REL-EL performance gap is 20 percentage points (in terms of percent proficient), there are 20 REL students and 80 EL students, then 20*(20/(80 + 20)) = 4. The reported EL/REL subgroup proficiency rate will be 4 percentage points higher than if the state reported ELs alone. All else being equal, as performance of ELs decreases, the relative impact of including RELs in the subgroup grows.

Combining groups of students will result in some level of a "masking effect", which is why it is important to disaggregate reporting by subgroup. However, the potential for this effect to change inferences about schools (and trigger subsequent action) is relatively small. A key component of reporting for either group combination (REL/EO or REL/EL) is driven primarily by the performance of the "base" group. Significant impacts on schools will arise when the EL proficiency rate is similar to the gap between ELs and RELs, and the ratio of RELs to ELs is close to 1.

Figures 22 – 24 present the impact of including REL students in the EL subgroup on ELA growth. In every instance, growth is higher for EL and REL students than for EO students. The impact of including REL students in the EL subgroup is marginal and consistent with expectations. Some research suggests that growth on the mathematics assessments administered in English for REL students tends to regress to state mean performance the longer the student has been reclassified. This is generally consistent with results presented in Figures 22–24[31].

30 This is true assuming the EO proficiency rate remains constant.

31 The growth results in figures 22–24 for EL and REL students vary quite substantially (these figures only present mean growth). The effect size difference between EO and EL/REL is close to 1, even though the absolute magnitude of the difference in growth is much larger. An effect size of 1 would still be considered large.

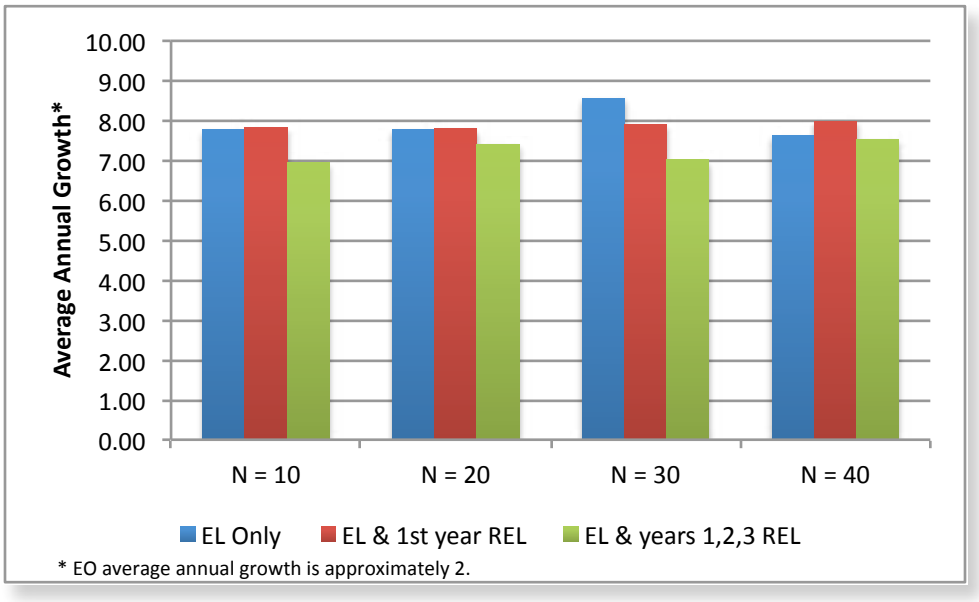


Figure 22: Impact of REL on EL Subgroup Growth - Elementary

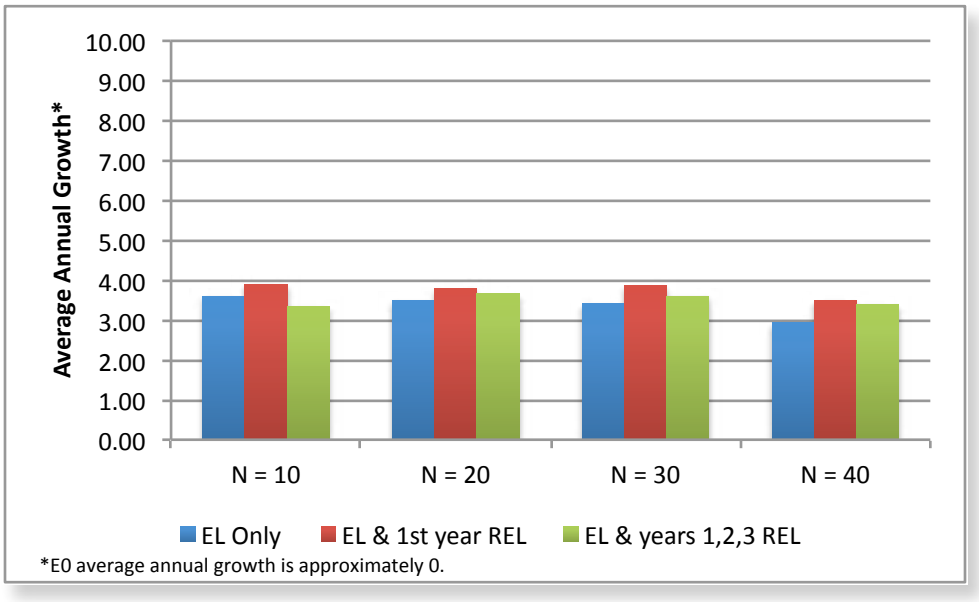


Figure 23: Impact of REL on EL Subgroup Growth - Middle

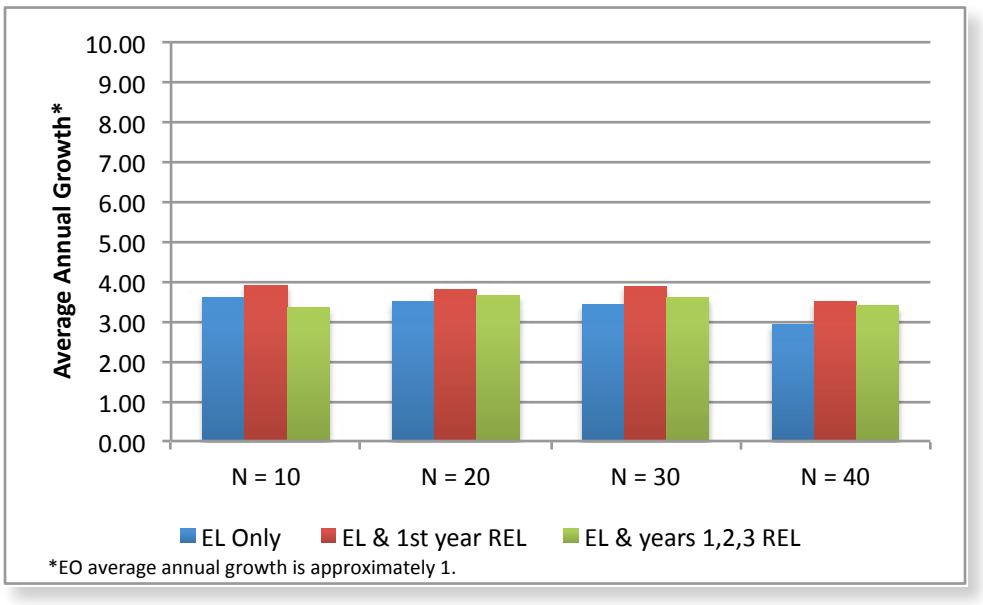


Figure 24: Impact of REL on EL Subgroup Growth - High

Including REL into the EL subgroup has fairly consistent, positive effects and helps states and schools monitor language development particularly for ELs on academic language arts and mathematics assessments administered in English.

Another important consideration for states is incorporating Recently Arrived EL (RAEL) students into the accountability system. States have several options under the ESSA amendments and the new accountability regulations (11/29/16) for doing so. Under the first exception, they must assess academic mathematics content mastery[32], but can choose not to assess RAEL students in ELA in year one, and use year two results as part of the academic content portion of the accountability system. Under the second exception, states must assess RAEL students in ELA, in year one, excluding those ELA results altogether in year one, while calculating growth (gain) in year two for academic content accountability purposes. In deciding how to include RAELs as they transition from RAEL to EL in the system, it is again important to examine state context and English language performance trajectories. One important aspect is whether after accounting for a student's initial ELD level, RAEL status is related to performance. Evidence suggests that a student's ELD level is related to performance but that RAEL status does not provide any additional information, meaning that if a state system intends to include ELD status for all facets of EL calculations, then RAEL status could be ignored[33]. Interpretations are thus based on student performance and not arbitrarily on labels. Also, it is important to consider the growth trajectory of EL students: initial content scores will likely be low and growth will be high. Again, if ELD level is included, status and growth regardless of ELD level will be unrelated to initial performance.

32 States can assess students in their native language for three-to-five years on the state reading/language arts assessments. There is no limit to assess mathematics and science in a student's native language.

33 There certainly may be other reasons to explicitly highlight RAEL status.

If the state is considering a growth model for content, it must consider how that growth model aligns with the approach they take for considering RAEL status. Option 2 anticipates a gain from year 1 to year 2, while the state model may be based on different expected outcomes; the important piece here is that inferences about students should be consistent.

It is also important to recognize that reporting (on school report cards) and calculations may differ. States including REL in the EL subgroup for example, should report on REL performance as its own subgroup. Also, it may be informative to report on RAEL students independently, as well.

CONCLUSION

Each time ESEA is reauthorized, a new set of opportunities are created for each state to review and assess its systems and processes. There are clear issues of compliance to the new law, but much more important are ways in which changes can be used by the state to engage in leadership to improve programs and student outcomes. A state's chosen accountability system and how it captures status and growth of the English Learners is an important piece of this leadership. Under ESSA, states have a new set of opportunities and challenges, particularly when it comes to including ELs in Title I accountability, and accounting for English Language Proficiency.

In addition to ESEA, the standards for compliance with the Civil Rights Act (the so-called Castañeda standards described in the introduction) present an additional lens to determine the effectiveness of an accountability system for EL students. A responsible accountability system for Title I should be based on sound educational theory and demonstrate effectiveness within a reasonable period of time. Accountability systems should provide information that can triangulate with state and local EL plans and visions that have been developed to align with a state's ToA.

The considerations described in this paper for developing a state accountability model for ELs that incorporates English language proficiency will, we hope, inform decisions that go into the new state plan. We repeat the questions that might be considered in arriving at decisions:

A. What are my state's expectations about English Language Proficiency development with respect to:
    1. ELP standards?
    2. Trajectory of development?
    3. Time to proficiency?
    4. Reclassification?
    5. Individual student factors that influence growth?
    6. Instructional program factors that influence time to proficiency?

B. What is my state accountability system trying to accomplish by including ELP as an indicator receiving substantial weight?

C. How do I know if some schools are doing a better job with EL students than other schools? How can the new accountability system help me in determining this?

D. Which models should my state consider for the ELP indicator? What tools do I need to effectively communicate these considerations with LEAs, schools, and stakeholders?

E. What are factors that should be considered in making a selection? Am I concerned with:

   1. Familiarity to stakeholders?
   2. Transparency of the model?
   3. Sensitivity to meaningful variation (not losing meaningful variation between students, between schools, between years)?
   4. Ability to take initial ELP level, time to proficiency, and other variables into account?
   5. Ability to optimize N-size (e.g., address reliability/stability of results while minimizing loss of schools that do not meet minimum N-size)?
   6. "Fairness" across grade bands (elementary, middle, high)?
   7. Year-to-year stability of the model in enabling state's accountability goals?
   8. Model consistency with your state's academic achievement indicator approach?

F. What are my state's considerations in choosing *N*-size? Are we concerned with:

   1. Percent of schools with ELs that are included or excluded from accountability for ELP?
   2. Number of years after reclassification that exited EL students can be included in the academic achievement subgroup (allowable for up to 4 years)?
   3. Discrepancy between ELP and academic achievement *N*-sizes that might come about as a result of decisions about (2)?

G. What kind of data modeling will my state consider in moving forward to include ELs in your plan?

States should first consider what is meant by progress and how results will be used (develop a ToA). There tend to be tradeoffs between technical soundness and transparency, although this is not all-or-nothing. We presented different aspects that can be weighted differently in deciding which model to pursue.

The results we developed indicate that monitoring student gains or growth using multiple years (based on a vertically scaled assessment) provides results that have good technical properties, are relatively transparent, maintain meaning over aggregation, and succeed at differentiating between schools. One method bases school scores on average growth, while another method uses average growth toward a certain standard in a set period of time. This approach eliminates the "growth to nowhere" effect, but does not reduce progress to a status measure, thus avoiding creating a bubble student.

A major concern for many states is monitoring EL progress given that some schools have many ELs, some have very few. States can compensate for this variation in a number of ways. They can choose to simply exclude the schools with the lowest number of ELs; use multiple years of data; or set very low minimum *N*s. We demonstrated that some progress models are stronger with smaller *N*s than others. There is no single correct approach, and each solution comes with tradeoffs. If a state's ToA is to include as many students and schools as possible then a small *N* may lead to some loss in stability, but this can be mitigated by choosing a model that is robust to EL counts in a school.

In general, no single method is best, nor is the most appropriate choice the same for each state. Choices should be made based in the broadest context possible for state policies that affect English Learners and their learning situation.

REFERENCES

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, 149*(1), 1-43.

Betebenner, D. (2009). Norm- and Criterion-Referenced Student Growth, *Educational Measurement: Issues and Practice,* Winter, 28(4), pp. 42–51.

Burstein, L. (1980). The analysis of multi-level data in educational research and evaluation. *Review of Research in Education*, *4,* 158-233.

Castellano, K., and A. Ho (2013). Contrasting OLS and Quantile Regression Approaches to Student "Growth" Percentiles, Journal of Educational and Behavioral Statistics, 38(2), 190-215.

Castellano, K., and A. Ho (2015). Practical Differences Among Aggregate-Level Conditional Status Metrics From Median Student Growth Percentiles to Value-Added Models, *Journal of Education and Behavioral, 40(1), 1 35-68.*

Choi, K., P. Goldschmidt, and K. Yamashiro (2005) Exploring Models of School Performance: from Theory to Practice in National Society for the Study of Education v. 104, Joan Herman and Ed Haertel Eds. Blackwell.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley.

Goldschmidt, Pete, Kilchan Choi, and J.P. Beaudoin (2012). *Growth Model Comparison Study: Practical Implications of Alternative Models for Evaluating School Performance*, Council of Chief State School Officers, Washington DC.

Goldschmidt, P., K.C. Choi, and F. Martinez, and J. Novak (2010). Using growth models to monitor school performance: comparing the effect of the metric and the assessment, *School Effectiveness and School Improvement,* 21(3), 337-357.

Willms, D. & Raudenbush, S. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability, *Journal of Educational Measurement*, 26(3), 209-232.

Council of Chief State School Officers

One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
voice: 202.336.7000 | fax: 202.408.8072