Considering English Language Proficiency within Systems of Accountability under the Every Student Succeeds Act

Susan Lyons, Ph.D. and Nathan Dadey, Ph.D.
Center for Assessment

March 27, 2017

This paper was written in collaboration with the Latino Policy Forum with significant financial support from the High Quality Assessment Project.

For more information or support please reach out to:

Susan Lyons
Center for Assessment
slyons@nciea.org

Karen Garibay-Mulattieri
Latino Policy Forum
karengaribay-mulattieri@latinopolicyforum.org

Table of Contents

Introduction

The requirement for an indicator of "progress in achieving English language proficiency" (English language proficiency) for English learners (ELs) must now be included in state systems of educational accountability under the Every Student Succeeds Act (ESSA, §1111(c)(4)). Specifically, the statute requires that English language proficiency be addressed in two[1] specific ways within systems of accountability—as part of the state's **long-term and interim goals,** and as part of an annual system that **meaningfully differentiates schools**. ESSA's inclusion of English language proficiency within Title I accountability systems represents a key juncture in accountability policy that provides states the opportunity to define, or redefine, progress in achieving English language proficiency in a system of accountability that considers all EL students.[2] The goal of this brief is to first provide an overview the ESSA requirements around English language proficiency within systems of accountability, and then to offer guidance on the ways in which (a) progress in achieving English language proficiency can be defined, (b) these various definitions can be incorporated into ESSA-compliant state accountability systems, and (c) a state can evaluate the validity of a state ESSA accountability system for meeting EL policy goals.

States must first establish a vision for English learners and English language acquisition embedded in a coherent theory of action before engaging in accountability system design. There are a variety of design decisions that must be made in order to create a new school accountability system under ESSA. The new federal law permits a wide latitude in the specifics of state accountability systems – allowing for variety of types of indicators reported, the stated goals and targets, and the rewards or consequences for schools. Therefore, state leaders need to base complex design decisions on a clear state vision and a theory about how the accountability system will function to support that vision. By providing clearly articulated educational goals for all students, and for English learners in particular, the state vision provides the basis for the evaluation of any particular aspect of the accountability system, as well as the role the accountability system plays within the state educational system. That is, a clearly outlined vision and accompanying theory of action is necessary to facilitate the design of a coherent accountability system.

ESSA Requirements

ESSA includes a number of major provisions regarding ELs and English language proficiency, many of which are similar to provisions in the No Child Left Behind Act of 2001. Outside of accountability, these provisions include requirements that states have adopted English language

[1] These two uses are mandated by the statute. However, these are not the only two uses for ELP indicators – states may wish to develop additional uses with their systems of accountability, not for federal compliance, but in order to better meet specific policy needs.

[2] Under the No Child Left Behind Act of 2001, the achievement of EL students was covered under Title III and thus accountability for EL student only applied to local educational agencies receiving Title III funds.

proficiency standards aligned with state academic standards, annual administration of an assessment of English language proficiency for all ELs and statewide entrance and exit requirements for ELs (cf., CCCSO, 2016). In terms of accountability, the law has two specific requirements around English language proficiency[3]:

1. **Long-Term Goals and Interim Progress**. The statewide accountability system must include "State-designed long-term goals, which shall include measures of interim progress towards meeting such goals… for increases in the percentage of such students making progress in achieving English language proficiency, as defined by the State" (ESSA, §1111 (b)(4)(A)).
2. **Annual Indicator**. The statewide accountability system must also include an annual measure of "progress in achieving English language proficiency, as defined by the state… within a State-determined timeline for all English learners" for all public schools in the state, which is to be used as part of a "system of meaningful differentiation" to identify schools for intervention[4] (ESSA, §1111 (b)(4)(B to D)).

These two requirements can be tightly or loosely coupled. For example, the annual measure of progress towards English language proficiency used to differentiate schools could be defined by working backwards from the state's long term goals (i.e., tightly coupled), or a state could define their progress towards English language proficiency indicator and long term goals separately (i.e., loosely coupled).

The final regulations on ESSA accountability were overturned by Congress under the Congressional Review Act as of March 10, 2017. Since the regulations have been pulled, Department of Education will not be allowed to release new regulations that are substantially similar to the revoked regulations, meaning that the law will likely need to be implemented by states without regulatory clarification (Ujifusa, 2017). Appendices A and B provide tables that provide and separate the language of the statute from the language of the regulations. While it is the regulations will not be legally enforceable, they may still be useful to states in providing additional specificity about statutory intent and are thus referenced throughout this paper, when relevant.

Though the regulations did provide further detail regarding the ESSA requirements, both the statute leaves a number of decisions regarding the progress towards English language proficiency indicator in the hands of states. For example, what constitutes "progress in achieving English language proficiency?" Should the long-term goals and measures of interim progress define progress in achieving English language proficiency in the same way as it is defined for the annual indicator? What timeline is defensible? The sections below consider these types of questions and detail the requirements of the law.

[3] Note, ELs are also included as a federally accountability subgroup for all of the other indicators within the accountability system, which means EL performance on each of the indicators must be reported separately for every school.

[4] i.e., either Comprehensive Support and Improvement or Targeted Support and Improvement.

Long-Term Goal and Measures of Interim Progress

The statute and regulations, in particular, require that a state develop ambitious "long-term goals and measures of interim progress for increases in the percentage of all English learners in the State making annual progress toward attaining English language proficiency" (34 C.F.R. §200.13(c)). There is considerable flexibility in how the state defines "making annual progress toward attaining English language proficiency for each student"—i.e., student-level progress for the required indicator.

Student-level progress. The regulations clarified that states should develop a procedure for calculating research-based, student-level targets for English learners to reach English language proficiency. This does not mean that all English Learners need to have the same targets, but instead, that their growth targets must be calculated using a consistent methodology for all students. The regulations stated that the procedure may take into account any of the following student-level characteristics: student's initial level of English language proficiency, time in language instruction, grade level, age, native language proficiency, and limited or interrupted formal education, if any. The regulations also required that the targets be based on research and data. For example, it would be reasonable for a state to expect larger gains for younger students than for older students; by setting targets to reflect the known differences in language acquisition, the state can reasonably expect all English Learners to show progress. These targets must also be based on a state determined maximum number of years by which a student should reach proficiency. As with the targets, this maximum number of years must be based on research and can vary by student demographic factors. Importantly, if an English learner does not attain English language proficiency within the state-determined maximum, that student must be provided English learner services until attainment. There is a body of research that can be leveraged to help states make informed decisions about target setting (see Hakuta, Goto Butler & Witt, 2000; MacSwan & Pray, 2005; Motamed, 2015; Slavin, Madden, Calderón, Chamberlain, & Hennessy, 2011). Additionally, states should be using existing prior data to model language proficiency trends within the state to better understand the likely implications of the target-setting decisions. Using state-specific data helps ensure that the targets and maximum number of years to reach proficiency are reasonable and achievable.

Long-term goal and interim progress. The student-level targets describe what is expected from individual students. The statue and regulations also require that the State set a long-term goal for the population of EL students. As the state must set a long-term goal which defines the percentage of English learners in the state making progress toward English language proficiency at a given point in the future. States have flexibility in what specific percentage the goal is, the timeframe for achieving the goal, and how measures of interim progress—intermediary goals that define increases in the percentage of English learners in the state making progress toward ELP—are defined. States must also provide descriptions for how each of these elements is established. As with setting student-level targets, states will want to spend time

examining their trend data to understand the historical progress of English learners toward English proficiency within the state.

The long-term goals, and supporting measures of interim progress, should also be aligned with the state's vision for ELs and the theory of action for making progress toward the vision. If the student-level targets are set thoughtfully and appropriately, it may not be unreasonable for a state to set the long-term goal of 100% of ELs making progress toward proficiency annually. However, with such high expectations, careful consideration should be given to what interventions and supports will be provided by both local and state actors. Unless EL students are offered substantially more support, a long-term goal that assumes levels of improvement well beyond those shown in historical trends may be suspect. Ultimately, the long-term goal should be challenging but achievable given the level of support and the time necessary to implement program improvement. Once the goal and timeframe are established, the measures of interim progress may be defined by backwards mapping—with an ambitious yet reasonable trajectory of attainment towards the goal—to set intermediate benchmarks that would indicate progress to success on the long-term goal. It is worth noting that like EL language acquisition, program improvement and progress toward the long-term goal may not follow a linear pattern.

Annual Indicator

It is a common misconception that accountability systems under ESSA represent a U-turn from those under No Child Left Behind (NCLB). The assessment provisions under ESSA are highly similar to NCLB's and annual, statewide content assessments in math and English language arts (ELA) remain a large part of accountability. However, accountability systems under ESSA require multiple additional indicators, including an indicator of progress toward English proficiency for English learners. In all, there are at least five categories of indicators that comprise accountability systems under ESSA:

1. Academic achievement as measured by annual, statewide assessments in math and ELA in grades 3-8 and high school;
2. Academic progress such as growth or achievement gap for elementary and middle schools (this is optional for high schools);
3. Graduation rate for high schools. This indicator category must include the 4-year cohort-adjusted graduation rate and may also include extended-year graduation rates;
4. Progress in achieving English language proficiency, the topic of the current paper; and,
5. Additional indicator(s) of school quality or student success.

Importantly, consequences for schools are attached to the summative annual determination based on all of the indicators listed above. Identification for targeted and comprehensive support must be informed by all of the accountability indicators. This is distinct from the long-term goals in that federal accountability does not *require* school-level consequences or action related to performance on those goals.

The English language proficiency indicator must be reported for at least all English learners in grades 3-8 and those who are assessed in grades 9-12. States may choose to include the assessment results of English learners in earlier grades and may have good reason to do so given that younger students tend to show the most growth in English language proficiency. This decision is discussed in more detail in the section entitled "Incorporating English language proficiency into Systems of Accountability." The final regulations further outline three requirements related to the indicator of progress towards English language proficiency: 1) it must use objective, valid measures of student progress on the proficiency assessment, comparing results across years, 2) the indicator of progress must be aligned with the applicable timelines for a student to attain English proficiency within the State-determined maximum number of years, and 3) the indicator may also comprise a measure of proficiency, for example, the percentage increase of English learners attaining proficiency on the English language proficiency assessment as compared with prior years. Lastly, all indicators in Title I accountability must be reported individually using at least three levels of performance. This means that the ELP indicator must differentiate among schools by reporting at least three categories of performance. The following section of the paper provides a deep dive into the different options for defining the measure of progress for this accountability indicator.

Defining and Evaluating an English Language Proficiency Indictor

We start with a heuristic to help show all of the major pieces that influence a school indicator of progress in achieving English language proficiency. Figure 1 illustrates that the state context, the specific model used to define the English language proficiency indicator, and the business rules around the implementation of the model all play a role in determining school performance classification on the English language proficiency indicator. State context deals with the on-the-ground reality of EL students within the state (e.g., Are ELs concentrated in a small number of schools or spread out across many schools? Are ELs concentrated in particular grades?). The statistical model refers to the methodology used to produce scores based on the English language proficiency assessment, which can then be aggregated to the school level. This area encompasses both the class of model used, as well as the way the model is specified and estimated. For example, does the model control for student characteristics and, if so, which ones? Finally, the business rules specify how the results of the statistical model are aggregated (e.g., How many students are needed before a school receives a score? Will the results be pooled over years? Will reclassified ELs be included in the aggregation?). In addition, the information in any one box can inform decisions in another box. For example, if there are few EL students per school, the state might want to choose a smaller n-size in to in order to provide ratings to as many schools serving ELs as possible, despite issues with precision caused by small sample sizes. These types of tradeoffs are common and a state will need to weigh the positives and negatives of any particular approach. We present these categories here as a structure that can be useful for guiding state discussions about this indicator.

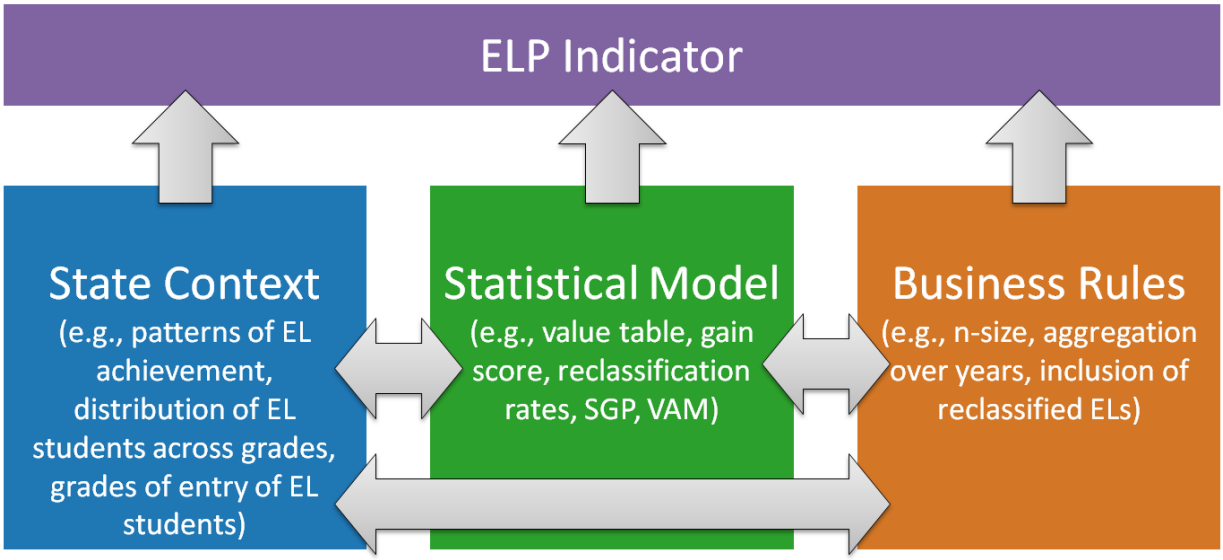

Figure 1. *Heuristic of Areas of Concern for English language proficiency Indicator*.

Finally, it is worth re-emphasizing that the English language proficiency indicator is one of multiple indicators that will ultimately decide the classification of a school under the full accountability system. Thus, the ultimate impact of the English language proficiency indicator needs to be considered in relation to the other indicators. For example, what role will the English language proficiency indicator play? What weight will the English language proficiency indicator have? Again, such questions need to be considered in light of a state's vision and theory of action. These questions are considered more deeply in the section entitled "Incorporating English language proficiency into Systems of Accountability."

Defining Progress in Achieving English Language Proficiency

The law requires than an indicator of *progress* in achieving English language proficiency be used. This requirement has generally been understood as requiring the quantification of across year changes in individual student performance on the English language proficiency assessment. However, the regulations do allow for the English language proficiency indicators to be a combination of growth and status. Given this understanding, prior work examining growth models for general student populations is applicable (e.g., Castellano & Ho, 2013; Goldschmidt, Choi, & Beaudoin, 2013) but should be re-evaluated in light of the unique characteristics of ELs.

In their recent paper, "Incorporating English learners Progress into State Accountability Systems," Goldschmidt and Hakuta (2017) evaluate options for growth indicators from a predominantly technical perspective. In this paper, we build on their work by integrating their perspective with additional considerations related to the implementation and evaluation of the English language proficiency indicator within an accountability system.

Some common approaches for characterizing change across years follow (Goldschmidt & Hakuta, 2017)[5]:

- **Transition (or Value) tables**: Transition tables describe growth as a student's change in performance level from one year to the next dependent on a student's prior status. Transition tables often use performance levels that are divided into sub-performance levels to illustrate growth within a performance level.
- **Proficiency rates:** Hakuta and Goldschmidt (2017) offer that the percentage of students reaching English language proficiency is a relevant indicator for monitoring ELs' progress. They argue this method is transparent, but note some challenges in that it will be sensitive to policies regarding reclassification and does not award credit for progress toward proficiency, only counting those students who reach proficiency.
- **Gain scores**: Gain scores describe a student's growth based on the difference between test scores- calculated by subtracting an earlier score from a later score. Gain scores require the use of a vertical scale (i.e., scale scores that range across grade levels). Gain scores can be in the raw metric of the scale scores or they can be normalized in order to provide a norm-referenced interpretation of relative growth.
- **Growth rates**: Growth rates characterize the rate at which student scores change over time. This is determined by calculating a best fit line, or a trend line, across a series of data points to estimate a student's growth rate. This estimate can be linear or non-linear.
- **Student Growth Percentiles (SGP)**: SGPs are based on the percent of academic peers a student outscores (i.e., growing faster than 35% of my peers). Academic peers are those students who have similar prior test scores. SGPs are reported on a 1-99 scale, with lower numbers indicating lower relative growth and higher numbers indicating higher relative growth. For example, if a student has an SGP of 65, it means the student has demonstrated more growth than 65% of his or her academic peers.
- **Value-Added Models**: Value-added models describe growth as the impact educators or institutions have on student achievement. While not all VAMs have the same model structure, many are residual models, calculated by comparing how much the performance in a given unit (e.g., class, school, or district) deviates from the average expected change in performance for that unit.
- **Growth-to-Target**: Each of the above models characterize growth in terms of magnitude, but do not explicitly account for whether a student has achieved English language proficiency. Each of the approaches can be modified to account for the required growth to meet a particular target or standard (e.g., English language proficiency). For example, adequate growth for SGPs is often defined in terms of the growth necessary to for non-proficient student to achieve proficiency within a given number of years ("catch-up" growth; Betebenner, 2011).

[5] These models are not all mutually exclusive. For example, SGP can be combined with growth-to-target (i.e., adequate growth percentiles).

This section of the paper provides a more in-depth discussion of three of the possible growth measures to highlight examples of how states could consider and weigh the various merits of an indicator relative to their state context and the policy goals. The three measures considered are: value tables, value-added models, and growth-to-target methods. These measures were chosen because they may be particularly promising for the English language proficiency indicator and each provide for a different inference related to student growth.

Value Tables. Value or transition tables allow policy makers to explicitly value growth across the performance categories in a way that aligns with the state's policy goals (Hill et al., 2005). Value tables are simple and transparent, in that they assign numerical values to changes in achievement. Movements across the achievement levels that are considered more desirable (e.g., from non-proficient to proficient) are given higher values, and thus, schools are awarded more credit. The school score resulting from a value table would be the average points for all of the English learners within the school and therefore, the student growth inference resulting from the value table is: *How valued is the observed student growth as measured by progress on the performance levels?* The values would be deliberated and decided upon at the state level, with the involvement of key stakeholder groups to ensure that the numerical values in each cell accurately reflect the state's theory of action. An example value table is provided in Figure 2 using the performance levels from the WIDA ACCESS fir ELLs 2.0 exam, an English language proficiency consortia assessment currently used in 38 U.S. states and territories. In the example provided, more points are awarded for moving into the higher levels of attainment than the lower levels, since growth at the high end of the scale is generally more difficult to achieve. Additionally, schools are awarded no points for students who lose English language skills across years. States may want to consider awarding some points, or even negative points, to these cells, depending on the state's theory of action.

		Year 2					
		1: Entering	2: Beginning	3: Developing	4: Expanding	5: Bridging	6: Reaching
Year 1	1: Entering	25	50	75	100	150	200
	2: Beginning	0	25	50	75	125	200
	3: Developing	0	0	25	50	100	200
	4: Expanding	0	0	0	25	75	200
	5: Bridging	0	0	0	0	25	200

Figure 2. *Example Value Table with WIDA Performance Levels*

One of the primary benefits of value or transition tables is their transparency for schools and other stakeholders. Once schools know how their students have scored on the English language proficiency assessment, they should be able to calculate their score on the English language proficiency indicator easily. Additionally, the values are set in a way that reflects the state's

theory of action, for example, schools can be incentivized and rewarded to improve English language proficiency for those students who typically have the most difficulty showing growth. One of the drawbacks of using a value table to measure growth for the English language proficiency indicator is that this methodology will be only loosely related to the state's long-term goals and measures of interim progress in that the typical use case for value tables does not include the creation of individual student targets aligned to the state's defined timeline for reaching proficiency. This would mean that the state would have to create a separate methodology for calculating student targets in order to track progress on the long-term goals. This could be done relatively easily, for example, by expecting that students' progress by one achievement level per year. However, the simplicity of this model for setting student targets may not be reasonable and, as with any target-setting scheme, should be modeled to better understand whether this kind of progress is reasonable to expect for students who have historically reached proficiency. Alternatively, more complex versions of value tables that take into account the student characteristics—including time in EL's programming—could be created. However, this would require the design and use of multiple value tables and may remove some of the transparency associated with this method. Some states are getting around the need to set student-level targets all together by changing the nature of the long-term goal itself (e.g., By 2025, 95% of ELs will reach English proficiency within five years).[6]

Value-Added Models. Value-added models are a diverse collection of statistical techniques that are better defined by their use rather than their structure. Often, value-added models are regression-based and are used to compare students' predicted growth to actual growth. The difference—the residual—is often attributed to programmatic effectiveness, or the value-added by the program to the student growth. Most value-added models are covariate-adjusted which means they can control for student and school contextual effects that may contribute to explaining student growth trajectories. In this way, value-added models are said to "isolate" the effects of the program on student achievement, regardless of school and student characteristics (a very strong assumption that has rarely been validated). The student growth inference related for value-added models is: *How effective is the EL program at eliciting student growth compared to other programs in the state?* It is important to note here that this inference is inherently norm-referenced in that EL progress is not measured relative to a criterion, but in comparison to progress made by other ELs in the state. Value-added models will identify those programs making better than average progress with their EL students, average progress, and below average progress.

One of the benefits of the value-added modelling framework for the English language proficiency indicator is that states can easily take into account of the student characteristics that research has shown to be relevant for explaining EL language acquisition (e.g., age, proficiency

[6] Depending on how strictly the language of the statute is enforced, this type of long-term goal for the ELP indicator may not be compliant.

in native language, limited or interrupted formal education etc.). This has the benefit of potentially leveling the playing field for schools in that the amount of actual growth necessary to show positive value-added scores is conditioned on these factors. Additionally, for states that administer the WIDA's ACCESS for ELLs 2.0 as their English language proficiency assessment, the significant increase in the performance standards on that assessment may make a normative model—such as value-added models or student growth percentiles—particularly desirable as the state EL programing adjusts to the new expectations for proficiency. If states are not substantially modifying their exit criteria for EL proficiency, then the increase in WIDA's performance standards will likely lead to a larger percentage of EL students failing to exit EL programming (i.e., not reaching the new proficiency standard). If this is the case, normative growth models such as VAMs or SGPs can still provide for meaningful differentiation among schools, while value tables or growth-to-target models will likely lead to low performance for all schools in a way that decreases the ability of the indicator to effectively discriminate among programs.

The drawbacks of using value-added models for the English language proficiency indicator include the lack of transparency for schools and stakeholders and a sometimes high cost in terms of time, resources, and capacity to develop and implement. Additionally, this type of modeling typically requires large student samples for estimation. If there are few ELs per grade level, states may run into issues due to insufficient sample size to run value-added estimates. Value-added models may also be difficult to explicitly link to the state's long-term goals and measures of interim progress. Because the estimate generated from a value-added model will often be a "program-effect," it is likely that the state would have to develop a separate methodology for estimating and tracking student-level targets in order to report on progress toward the long-term goal.

Growth-to-Target. As with value-added models, growth-to-target growth frameworks are diverse in their mechanics but similar in conceptualization. In general, these models begin with the definition of a target for each student over a specified timeframe. In the case of the English language proficiency indicator, the "target score" would most likely be the scale score(s) corresponding to proficiency on the English language proficiency assessment. The trajectory of student growth necessary to reach proficiency within the state-defined timeframe is mapped from the student's starting status. The trajectory for the student can be different depending on student characteristics that might interact with that student's language acquisition. Based on this trajectory, annual targets are set for each student to track their progress toward proficiency. Figure 3 provides examples of different possible growth trajectories with annual benchmarks. The school's accountability metric could then be as simple as the percentage of students meeting their annual growth targets (e.g., the scale score necessary staying on or above the growth trajectory). The growth inference is therefore: *What percentage of students are on-track to achieve English language proficiency within the state-defined timeline?*

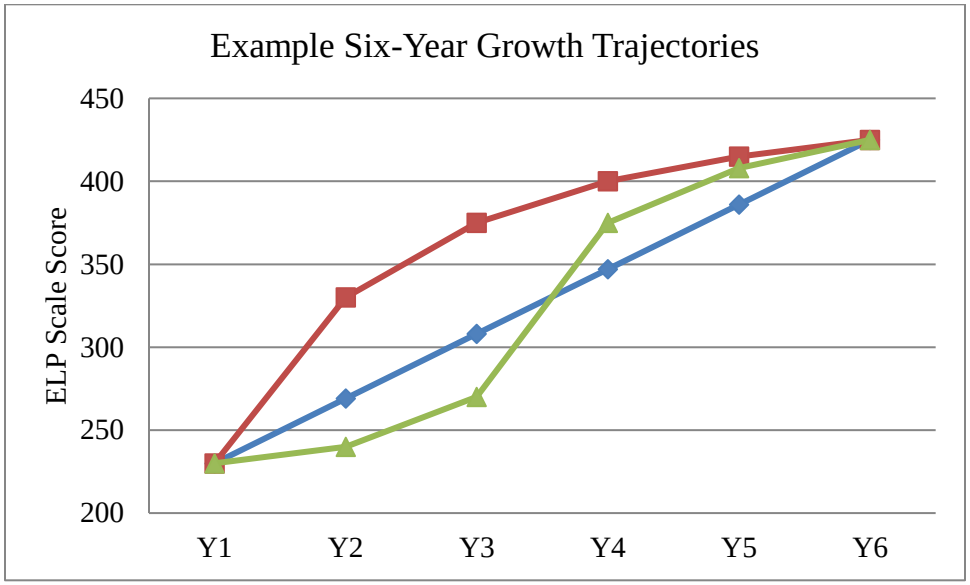

Figure 3. *Example Growth to Target Trajectories for Grade 3*

As illustrated in Figure 3, the care with which the projected growth trajectories are set will greatly influence the degree to which schools are able to achieve success on this indicator. This is both a strength and a limitation of the growth-to-target model, in that the growth trajectories will only be as valid as the theoretical framework supporting the hypothesized trajectory of language acquisition for ELs. Additionally, it is recommended that states use historical performance within the state on the ELP assessment to help inform the trajectory and targets.

Another way to operationalize the growth-to-target model is with adequate growth percentiles. Adequate growth percentiles are based on the Student Growth Percentile (SGP) model (Betebenner, 2008). A student's SGP is the percentile rank of the magnitude of student growth, as compared to the growth of the student's academic peers (i.e., students with the same or similar histories of English proficiency attainment). Student growth percentiles have desirable properties for evaluating growth normatively in that all students, no matter their prior level of performance, have an equal probability of scoring a high or low SGP. Adequate growth percentiles may be particularly well-suited for the English language proficiency indicator in that they combine the normative growth interpretation with a criterion-based target. A student's adequate growth percentile is the SGP that a student needs to achieve in order to be on-track for meeting proficiency within the state-defined timeline. Adequate growth percentiles directly correspond to the student score needed to meet the target in a typical growth-to-target model, but provide additional normative information to the educators and parent about just how difficult—or likely—that target is to achieve.

One of the clear strengths of using a growth-to-target model as the English language proficiency indicator is the tight connection with the state's long-term goals and measures of interim

progress. Growth-to-target models set annual student-level targets aligned to a timeline for reaching proficiency. This feature allows for a seamless connection between scores on the English language proficiency indicator and progress toward state-level goals on English proficiency. As noted earlier, one of the potential drawbacks of growth-to-proficiency models is our limited understanding of ambitious yet reasonable growth trajectories for acquiring English language proficiency. The meaningfulness of the indicator for differentiating among schools and English language program quality depends on the care that is taken to appropriately define student growth trajectories. This metric is particularly susceptible to losing the ability to capture true differences among schools if the trajectories are unrealistic or not based on research about student learning. This might be especially problematic if the proficiency standard itself is unattainable for most students. For example, if the proficiency standard is set at the 90th percentile of performance on the ELP assessment (which is not unseen given the raised performance standards on ACCESS for ELLs 2.0), then after a few years almost all students will likely fail to meet their targets and the growth-to-standard model will have limited utility for meaningfully differentiating schools.

The benefits and potential drawbacks for each of the growth models discussed in this paper are summarized in Table 1.

Table 1.
Summary of Discussion of Growth Models

	Growth Inference	Benefits	Potential Drawbacks
Value Tables	*How valued is the observed student growth as measured by progress on the performance levels?*	• Transparency for schools and stakeholders. • Points award directly affect policy values for movement across achievement levels. • Can be directly tied to theory of action for incentivizing particularly difficult growth.	• Likely only loosely related to progress toward state's long-term goals on ELP indicators. • May become overly complex if student-level characteristics are taken into account by creating multiple value tables.
Value-Added Models	*How effective is the EL program at eliciting student growth compared to other programs in the state?*	• Modeling can easily accommodate for including student characteristics related to English language acquisition. • Norm-referenced inferences may be preferable in potential scenario where the exit criterion is highly ambitious for most EL students.	• Potential for lack of transparency, and high cost on time, resources, and capacity for developing and implementing the model. • High sample size required for estimation. • May be difficult to explicitly link to state's long-term goals.
Growth-to-Target	*What percentage of students are on-track to achieve English language proficiency within the state-defined timeline?*	• Growth trajectories can be based on research and historical performance in the state. • Tight connection with the state's long-term goals and measures of interim progress.	• Model will only be as strong as the theoretical framework dictating the growth trajectories. If these are not set correctly, the metric will lose its ability to capture true differences among schools. • Potential for losing meaningfulness in presence of high proficiency standard.

Evaluating Potential English Language Proficiency Indicators

Once one or more indicators of progress toward English language proficiency have been selected for consideration by the state, a systematic process for evaluating the policy-based feasibility and technical quality of the each of the potential metrics should be established. This means that in the evaluation of any growth model, states need to not only consider the technical aspects of the model, but the political and contextual considerations such as how it supports the vision for effective instruction of ELs, how the accountability system will use the growth model to support such efforts, as well as the capacity the state has to effectively implement and communicate about a particular model. The following table outlines the policy and technical criteria states could use for evaluating of the potential growth models used for the progress towards English language proficiency indicator.

Table 2.

Policy and Technical Criteria for Evaluating Indicators of Growth[7]

Policy-Based Criteria	Technically-Based Criteria
▪ **Policy Goals/Purpose**: *Alignment with theory of action? Usefulness for supporting changes in behavior?* ▪ **Interpretation/Inference Supported**: *Does the model provide the intended inference related to student growth?* ▪ **Utility:** *How useful are the results for supporting the intended growth inference? Are the results so complex to interpret that they lose meaning for the general public?* ▪ **Resources:** *How much will the model cost to run in terms of time and cost? What is the additional data burden if any?*	▪ **Technical Goals/Purpose:** *How well does it differentiate individually and as part of the system?* ▪ **Consistency:** *What kind of stability does the model support in its classifications? How reliable are school-level scores across years?* ▪ **Equity:** *How unrelated are the model outputs to school and student demographics?* ▪ **Consequences:** *How corruptible are the scores? What are the potential unintended negative consequences of this model choice?*

For the policy-based criteria, states should be holding internal discussions about values and capacity as well has hosting formal stakeholder engagement sessions to allow for public input into the choice of indicator. Discussions with educators will be additionally informative for understanding how a choice in metric may drive intended or unintended changes in behavior. Public listening sessions are another option to gather feedback from interested parties about the complexity and interpretability of the model options.

[7] Adapted from D'Brot & Goldschmidt, 2016.

The technically-based criteria should be evaluated, to the extent possible, by modeling existing data to make predictions about the implications of the different modeling decisions. Using past data on the English language proficiency assessment to test run the different model options will allow state leaders to better understand the utility of each growth model for providing valid and reliable indicator scores that accurately represent the progress of ELs in attaining proficiency in schools across the state.

Incorporating English Language Proficiency into Systems of Accountability

Once the statistical model for quantifying progress towards English language proficiency has been chosen, there are a number of equally consequential decisions that must be made about how the English language proficiency indicator will factor into the larger school accountability system. We suggest that the driving question guiding these decisions should be: *Which solution would lead to outcomes that are most likely to promote growth towards English language proficiency for English learners while also prioritizing fairness?* In accountability design, it is important to consider this question from the lens of the schools, and also from the lens of ELs. While both notions of fairness are essential considerations within the accountability system, they can be sometimes at odds with one another. There will be trade-offs between the sometimes competing goals of incentivizing desirable behaviors to benefit ELs and ensuring fairness for all schools. These tradeoffs are highlighted in the example, high-leverage decision points provided in the following sections.

Including K-2 Students in the Growth Calculation

States may want to incentivize schools to provide intensive and effective English language interventions for young EL students entering in kindergarten and the early elementary grades. To incentivize a focus on early childhood English language programming, states may consider including the growth of K-2 students in the school accountability system in addition to the federally mandated grades of 3-8 and high school. Stakeholders representing schools and educators may also be in favor of this decision because they may feel that they are more likely to be successful at bringing the youngest students to English proficiency and they would want the school accountability system to value this success.

Including K-2 growth within the school accountability model may draw additional attention to the importance of providing high-quality English language supports to the youngest students, but states must be careful that appropriate, research-based targets are carefully set for this age cohort of students. States will need to set targets that take into account the potential differences in growth for younger students in order to set an ambitious yet attainable timeframe for reaching proficiency and rigorous annual targets measuring progress toward that goal. Value-added models or student growth percentiles may be particularly useful for ensuring that the magnitude of student growth is taken in context and compared to academic peers. If value tables are used, states will have to carefully model the effects of setting the same cell values for all elementary

school students. If the same values are used, the school scores may suffer from instability across years for different cohorts of students with variable numbers of EL students. It may make more sense to have separate cell values for K-2 students since their expected growth is likely to be different from their older classmates (Collier, 1987). The decision about whether or not and how to include K-2 students in the growth calculation is a prime example how of states will need to thoughtfully examine the trade-offs between promoting desired behaviors while ensuring fairness for all students and schools.

Creating Reporting Categories

As previously mentioned, the ESSA regulations require that each of the school accountability indicators be reported individually with at least three levels of performance. All of the example growth models provided in this paper will allow for fine-grained reporting of many ordered levels of performance. States will need to decide if it makes sense to report in the scale of the indicators, or if it makes more sense to transform the indicator scores into levels of performance, or index scores. Figure 4 provides an example of how average value table scores could be easily converted into index scores using a rubric.

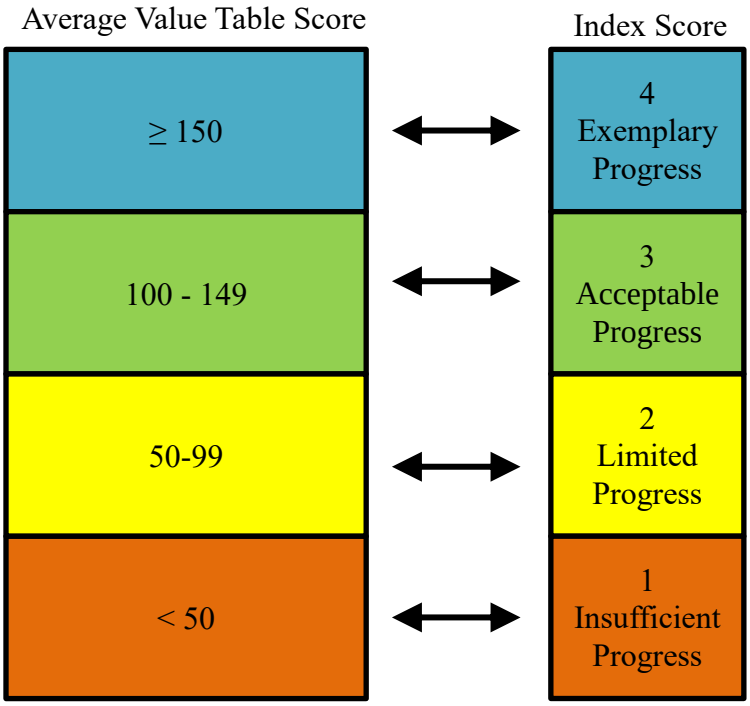

Figure 4. *Example of Value Table Rubric for Reported Index Scores*

The benefit of using index scores to signify levels of performance rather than reporting the raw metric of the growth model is that it improves the interpretability of school performance. Index scores can create common understandings about program effectiveness without the need to fully understand the details of the growth model and its resultant scores. States should undergo a planned standard setting process that defines levels of performance in a way that is reflective of

both the state's values and the empirical distribution of scores across schools in order to create meaningful index scores. Additionally, it may be helpful to provide performance level descriptors along with the index scores to further enhance interpretability.

Regardless of the type of score used—raw or index—states will need to determine how to report performance in a way that does not systematically advantage or disadvantage schools that do not have enough English learners to meet the minimum sample size for reporting on this indicator. In many states, a substantial proportion of schools will not have a reportable score for this indicator. States must design systems that can produce comparable overall annual determinations, regardless of whether or not schools have a reportable score for this indicator. If this is not done carefully, states may end up with a system that, say, disproportionately identifies schools with reportable scores on the EL indicator as poorly (or highly) performing. For example, if states do not report this indicator for schools that service no or few ELs, then by default, the weighting of the other indicators in the accountability system would each increase, and according to the regulations, this must be done proportionally. The regulations specify that the relative weighting among the indicators, for schools that do not reach the minimum-n on the ELP indicator, must remain constant (34 C.F.R. § 200.18(d)(3)(ii)). However, we know that not all indicators are equally difficult to attain, and if the ELP indicator is even slightly more or less difficult for schools to reach an acceptable score than the other indicators, then excluding this indicator for some schools will lead to incomparability in the resulting overall determination. Another way to conceptualize this lack of comparability is accountability model bias—either positive or negative—for schools that serve students who are English learners.

We offer a possible solution for alleviating this potential problem: states could specify the ELP indicator as a "neutral" indicator (Marion, D'Brot & Lyons, 2017). A neutral indicator is one where schools that do not serve enough English learners to meet the minimum n are given the same score as those schools that are making average progress with their EL students. This means that the only schools receiving high marks on this indicator are those that are making above average progress with ELs, and the only schools receiving low marks on this indicator are those that are making below average progress. While the use of a neutral indicator does not completely eliminate the potential for incomparability across schools, it should greatly lessen the issue. A neutral indicator can be operationalized in many ways such as creating three categories of performance—or a three-point index score—or, by using symbols (e.g., +/-) to report on performance instead of numerals. While this method of assigning points is inherently normative, modeling historic or extant data can help states proactively identify any unintended consequences associated with extreme cases and make adjustments as needed (e.g., all schools in the state have low rates of EL progress, meaning that even relatively high performing schools on this indicator should not be rewarded).

Including the ELP Indicator in the Overall Determination

In addition to individual reporting on each of the indicators, ESSA requires that states provide an overall accountability determination for every school that is informed by performance on all of the accountability indicators. This means that states will need to decide on: 1) the method for aggregating indicators into an overall score, and 2) the weight of each of the indicators within the overall score. The law does not provide much guidance relative to the method of aggregation, but it does specify that the English language proficiency indicator must be given "substantial" weight, and when combined with the other academic indicators, "much greater weight" than the additional indicator of school quality or student success (ESSA, § 1111 (c)(4)(C)). Whatever the method of aggregation (e.g., decision table, composite index), states will need to run empirical analyses to evaluate the effective weight of the English language proficiency indicator within the overall determination. Even with a weighted composite index where nominal indicator weights are explicit within the formula for aggregation, the effective weight of the English language proficiency indicator may be more or less depending on the variances of all of the indicators. Decisions regarding the method of aggregation and the weighting of the indicator within the overall accountability determination will ultimately have implications for incentivizing desired behaviors and fairness. If the English language proficiency indicator is given little weight within the final score, states may inadvertently signal that making progress on achieving ELP is not a priority. Alternatively, a highly-weighted English language proficiency indicator could exacerbate the issues with score incomparability and bias described in the previous section.

Some states are considering weighting the ELP indicator according to the percentage of EL students within each school. For example, a state using student growth percentiles for both the growth indicator and the ELP indicator could simply aggregate all the SGPs for a given school together when creating the overall annual determination. This would give the ELP indicator a weight that is proportional to the percentage of students within the school who are English learners.[8] Regardless of the desired weighting and aggregation schemas, states should be providing a forum to engage with stakeholders and host meaningful conversations about the implications and trade-offs of the accountability design decisions for including the English language proficiency indicator within the overall system.

Validating a System of Accountability for Meeting Policy Goals

Our conceptions of validating systems of accountability are evolving beyond the frameworks and tools used to support the validity of psychological measures, or tests (see Betebenner, 2017). The validity of an assessment is premised on the degree of theoretical and evidentiary support for the appropriateness of the intended interpretations and uses of the test scores (Messick, 1989). Because accountability systems are not constructed as measures, and instead comprise composites of indicators, systems of accountability can be better characterized as mechanisms

[8] This option may now be compliant given that the regulations have been repealed. The regulations clarified that the state use a uniform weighting scheme for all public schools within each grade span (34 C.F.R. § 200.18(b)(3)).

for evaluation. The purpose of ESEA accountability is to ensure that public tax dollars are resulting in improved educational programming and the intended student outcomes related to achievement and equity (Bailey & Mosher, 1968). If designed well, accountability systems should help us to determine which types of programs and settings are ultimately best for students. This is distinct from a measurement purpose where the goal would be to place every school on some latent continuum representing school quality or school effectiveness. Therefore, the criteria we use validate a system of accountability are necessarily different from those criteria we are familiar with for evaluating the validity of psychological instruments. Systems of accountability should be evaluated on the degree to which they provide "useful information and constructive responses in support of one or more policy goals… within an education system, without causing undue deterioration with respect to other goals" (Braun, 2008). Thus, Braun argues that *consequential validity* is the primary criterion on which accountability systems must be validated. In other words, does the theory of action in fact play out in the way intended?

The evaluation of an accountability system necessitates a coherent, and well-articulated theory of action outlining how the accountability system will bring about the desired change in service to the stated policy goals. Without the a priori hypothesis about how the accountability system is intended to function and lead to improved outcomes for schools and students, there is no way to evaluate whether or not the system is adequately serving its purpose. The theory of action creates a chain of logic and testable hypotheses that structure the validation activities.

While this framework for accountability system validation can apply to the entire system, it can also be applied to parts of the whole such as the evaluating the functioning of the English language proficiency indicator for supporting state-level EL policy goals. This is played out in the simple example provided in Figure 5. Please note that this example has been substantially simplified from a full theory of action related to English learners for the purposes of clearly illustrating the processes by which a state could design a validation plan.

<table>
<tr><th>Policy Goal(s)</th><th>System Input(s)</th><th>Change Agent(s)</th><th>Outcome(s)</th></tr>
<tr><td rowspan="3">All EL students reach English Language Proficiency within five years upon entering programming and upon exit, be prepared for long-term success.</td><td rowspan="3">1. The accountability system reports on student-level and school-level progress toward the goal of exiting in five years.
2. As a separate subgroup, the accountability system reports on success of former English learners on each of the indicators.</td><td>Schools evaluate their language programming and move to models that are more likely to move students to proficiency and long-term success (e.g., bilingual education).</td><td rowspan="3">1. Low performing students receive improved services and the state observes improved rates of EL students reaching English language proficiency within five years.
2. Former EL students are as successful as non-ELs in measures of long-term success such as content proficiency, graduation rate, and college and career readiness.</td></tr>
<tr><td>Students, parents, and schools receive feedback from the system on student-level progress in order to provide additional intervention when appropriate.</td></tr>
<tr><td>The state identifies schools that are most in need to support with EL students and provides additional resources and programmatic guidance.</td></tr>
<tr><td colspan="4">Schools provide additional supports to re-classified EL students, when necessary, to support long-term success.</td></tr>
<tr><td colspan="4"> 
<u>Example testable assumptions:</u>
• English language proficiency within five years is a reasonable yet ambitious goal for all English learners.
• Schools have information and resources for implementing more effective language programming.
• The ELP indicator accurately and reliably characterizes student-level progress toward proficiency.
• The ELP indicator accurately and reliably identifies schools that are most in need of support.
• Supports and programmatic changes result in improved services and outcomes for ELs.
• Schools have information and resources for implementing additional supports for ELs who have reached language proficiency but who are struggling academically.</td></tr>
</table>

Figure 5. *Example Simple Theory of Action for an EL-related Policy Goal*

The ultimate criterion for validating the utility of the ELP indicator for supporting the policy goal is to measure its effect on the intended outcomes. However, this may be difficult to do in that establishing causality is problematic in an ever-changing education and policy landscape, and it may take significant time before changes in the outcome are observed. Therefore, it is useful for states to begin evaluating the degree to which the assumptions that underlie the theory

of action hold. Violations of any assumptions may lead to a failure to achieve the intended outcomes or, potentially more damaging, lead to unintended negative consequences associated with the accountability system.

We provide two examples to further illustrate how the assumptions can and should be evaluated as part of the accountability system validation plan. First, when testing the assumption related to the accuracy of the ELP indicator for identifying schools that are most in need of support, states will want to consider the effect of minimum n size on school classification. To this end, states could model how the percentage of schools classified in each of the indicator performance levels shifts as the minimum n size varies. If the distribution of ratings changes as the n size changes, this reveals that the minimum n size decision is acutely related to the tenability of the tested assumption. While it is likely that the "true distribution" of school classifications is unknown, deciding on a minimum n size under the conditions described should involve consideration of Type I and Type II error. In the scenario provided, states may be more concerned about Type II error (i.e., failing to identify a struggling school in need of support) than Type I error (i.e., wrongly identifying a school providing adequate services as in need of additional support), since the consequences associated with identification are non-punitive. On the other hand, states with very limited resources that are concerned about spreading the resources too thin across a high number of schools may find that Type I error is preferable—and ultimately more effective for reaching the intended outcome—than over-identifying schools for support. As has been stressed throughout this paper, it is clear that the decisions about business rules need to be made in conjunction with the state context and theory of action, and are in the end inextricably linked to the validity of the entire system.

Another example of a testable assumption provided by the theory of action is that schools have the information and resources they need to effectively implement research-based language programming for EL students. Given the new emphasis on English language proficiency in ESSA, and the increased expectations for students and schools in this area, states may be wise to survey current language practices in their state to better understand the type and quality of supports offered. This may be an area where guidance from the state about best practices for EL programming is welcomed by schools that are now going to be facing heightened expectations and consequences associated with English language proficiency. Additionally, since all schools will now be held publicly accountable for progress in attaining proficiency, this might be a good time for states to consider undergoing a research effort to better understand what type of programming works best for ELs within their particular state context.

In sum, a coherent state vision and theory of action will not only be invaluable in the design of an effective accountability system, but this structure can also be relied upon to test the many assumptions that support the validity of the accountability system as a whole.

References

Bailey, S., & Mosher E. (1968). *ESEA: The Office of Education Administers a Law*. Syracuse, N.Y.: Syracuse University Press.

Betebenner, D. W. (2017, March 7). *A foolish consistency.* Presentation to the National Center for the Improvement of Educational Assessment staff. Retrieved from: https://dbetebenner.github.io/Presentation_030717/#cover

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155–170). New York: Taylor & Francis.

Braun, H. (2008, September). *Vicissitudes of the validators*. Paper presented at the Reidy Interactive Lecture Series, Portsmouth, NH.

Castellano, K. E. & Ho, A. D. (2013). *A practitioner's guide to growth models*. Council of Chief State School Officers: Washington, DC.

Collier, V. P. (1987). Age and rate of acquisition of second language for academic purposes. *TESOL quarterly*, *21*(4), 617-641.

Council of Chief State School Officers (2016, February). *Major provisions of every student succeeds act (ESSA) related to the education of English learners.* Washington, DC: Author.

D'Brot, J., & Goldschmidt, P. (2016, June). *Improvement and Growth Measures*. Workshop presented at the Chief Council of State and School Officer's ESSA Accountability Systems Technical Assistance Meeting, Tempe, AZ.

Goldschmidt, P., Choi, K., & Beaudoin, J. P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance*. Council of Chief State School Officers, Washington DC.

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

Hakuta, K., Goto Butler, Y., & Witt, D. (2000). ''*How long does it take English learners to attain proficiency?*'' University of California Linguistic Minority Research Institute Policy Report 2000–1.

Hakuta, K., & Pompa, D. (2017). *Including English Learners in Your State Title I Accountability Plan*. Washington DC: Council of Chief State School Officers.

Hill, R., Gong, G., Marion, S., DePascale, C., Dunn, J., & Simpson, M. A. (2005, November). *Using Value Table to Explicitly Value Student Growth*. Paper presented at the Conference on Longitudinal Modeling of Student Achievement, College Park, MD.

MacSwan, J., & Pray, L. (2005). ‘‘Learning English bilingually: Age of onset of exposure and rate of acquisition among English language learners in a bilingual education program.’’ *Bilingual Research Journal, 29*(3), 653– 678

Marion, S., D'Brot, J., & Lyons, S. (2017). *Improvement-Based School Accountability in New Hampshire: Recommendations from the New Hampshire Department of Education's Accountability Task Force.* Center for Assessment.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18*(2), 5-11.

Motamedi, J.G. (2015). ‘‘*Time to reclassification: How long does it take English language learners in the Washington Road Map school districts to develop English proficiency?*’’ U.S. Department of Education, Institute of Education Sciences.

Slavin, R.E., Madden, N.A., Caldero´ n, M.E., Chamberlain, A.,& Hennessy, M. (2011). ‘‘Reading and language outcomes of a five-year randomized evaluation of transitional bilingual education.’’ *Educational Evaluation and Policy Analysis, 33* (1), 47–58.

Ujifusa, A. (2017, March 1). *Measure to overturn ESSA accountability rules introduced in Senate.* Politics K-12, Education Week. Retrieved from: http://blogs.edweek.org/edweek/campaign-k-12/2017/03/essa_accountability_rules_overturn_sen_alexander.html

U. S. Department of Education. Rule; Elementary and secondary education act, as amended by the Every Student Succeeds Act: Accountability and state plans. 81 Fed. Reg. 86076-86248 (November 29, 2016).

Appendix A: Statute and Regulations for Long-Term Goal and Measures of Interim Progress

Statute	Regulations
§1111 (b)(4)(A) Establishment of Long-Term Goals. Establish ambitious State-designed long-term goals, which shall include measures of interim progress toward meeting such goals— (ii) for English learners, for increases in the percentage of such students making progress in achieving English language proficiency, as defined by the State and measured by the [ELP assessments].	**§ 200.13(c) (c) English language proficiency.** (1) Each State must, in its State plan under section 1111 of the Act- (i) Identify its ambitious State-designed long-term goals and measurements of interim progress for increases in the percentage of all English learners in the State making annual progress toward attaining English language proficiency, as measured by the English language proficiency assessment required in section 1111(b)(2)(G) of the Act; and (ii) Describe how it established those goals and measurements of interim progress. (2) Each State must describe in its State plan under section 1111 of the Act a uniform procedure, applied to all English learners in the State in a consistent manner, to establish research-based student-level targets on which the goals and measurements of interim progress under paragraph (c)(1) of this section are based. The State-developed uniform procedure must-- (i) Take into consideration, at the time of a student's identification as an English learner, the student's English language proficiency level, and may take into consideration, at a State's discretion, one or more of the following student characteristics: (A) Time in language instruction educational programs. (B) Grade level. (C) Age. (D) Native language proficiency level. (E) Limited or interrupted formal education, if any; (ii) Based on the selected student characteristics under paragraph (c)(2)(i) of this section, determine the applicable timeline, up to a State-determined maximum number of years, for English learners sharing particular characteristics under paragraph (c)(2)(i) of this section to attain English language proficiency after a student's identification as an English learner; and (iii) Establish student-level targets, based on the applicable timelines under paragraph (c)(2)(ii) of this section, that set the expectation for all English learners to make annual progress toward attaining English language proficiency within the applicable timelines for such students. (3) The description under paragraph (c)(2) of this section must include a rationale for how the State determined the overall maximum number of years for English learners to attain English language proficiency in its uniform procedure for setting research-based student-level targets, and the applicable timelines over which English learners sharing particular characteristics under paragraph (c)(2)(i) of this section would be expected to attain English language proficiency within such State-determined maximum number of years. (4) An English learner who does not attain English language proficiency within the timeline under paragraph (c)(2)(ii) of this section must not be exited from English learner services or status prior to attaining English language proficiency.

Appendix B: Statute and Regulations for Annual Indicator

Statute	Regulations
§1111 (b)(4)(B)(iv) For public schools in the State, progress in achieving English language proficiency, as defined by the State and measured by the [state ELP assessments], within a State-determined timeline for all English learners— (I) In each of the grades 3 through 8 and (II) In the grade for which such English learners are otherwise assessed…during the grade 9 through grade 10 period, which such progress being measured against the [state's ELP assessments] taken in the previous grade.	§ 200.14(a)(4) (4) For all schools, a Progress in Achieving English Language Proficiency indicator, based on English learner performance on the annual English language proficiency assessment required under section 1111(b)(2)(G) of the Act in at least each of grades 3 through 8 and in grades for which English learners are otherwise assessed under section 1111(b)(2)(B)(v)(I)(bb) of the Act, that-- (i) Uses objective and valid measures of student progress on the assessment, comparing results from the current school year to results from the previous school year, such as student growth percentiles; (ii) Is aligned with the applicable timelines, within the State-determined maximum number of years, under § 200.13(c)(2) for each English learner to attain English language proficiency after the student's identification as an English learner; and (iii) May also include a measure of proficiency (e.g., an increase in the percentage of English learners scoring proficient on the English language proficiency assessment required under section 1111(b)(2)(G) of the Act compared to the prior year).