U.S. Department of Education

National Evaluation of Title III Implementation Supplemental Report— Exploring Approaches to Setting English Language Proficiency Performance Criteria and Monitoring English Learner Progress

National Evaluation of Title III Implementation Supplemental Report— Exploring Approaches to Setting English Language Proficiency Performance Criteria and Monitoring English Learner Progress

---

Submitted to

U.S. Department of Education
Office of Planning, Evaluation and Policy Development
Policy and Program Studies Service

Prepared by

Gary Cook, Wisconsin Center for Educational Research
Robert Linquanti, WestEd
Marjorie Chinen, American Institutes for Research
Hyekyung Jung, American Institutes for Research

American Institutes for Research
Washington, DC

2012

This report was prepared for the U.S. Department of Education under Contract Number ED-04-CO-0025/0017. Elizabeth Eisner and Andrew Abrams served as the contracting officer's representatives. The views expressed herein do not necessarily represent the positions or policies of the Department of Education. No official endorsement by the U.S. Department of Education is intended or should be inferred.

U.S. Department of Education
Arne Duncan
Secretary

Office of Planning, Evaluation and Policy Development
Carmel Martin
Assistant Secretary

Policy and Program Studies Service
Stuart Kerachsky
Director

March 2012

This report is also available on the Department's website at http://www.ed.gov/about/offices/list/opepd/ppss/index.html.

On request, this publication is available in alternative formats, such as Braille, large print, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-0852 or 202-260-0818.

Content Contact
Andrew Abrams
202-401-1232
Andrew.abrams@ed.gov

Contents

Exhibits

Appendix D. Education Agency 3

Appendix E. Event History Analysis

Appendix F. Education Agency 1

Acknowledgments

We wish to thank several individuals who contributed to the completion of this study. First, we are grateful for the guidance and support of the U.S. Department of Education. In particular, we thank Elizabeth Eisner and Andrew Abrams, of the Policy and Program Studies Service, who served as our project officers on this study. We also thank Stuart Kerachsky, the Director of the Policy and Program Studies Service for his guidance on the study as well as Supreet Anand, of the Office of Elementary and Secondary Education, and Scott Sargrad, of the Office of Planning, Evaluation and Policy Development. We are also grateful to the staff of the American Institutes for Research for their assistance in producing this report. In particular, we thank James Taylor for his guidance in writing and analysis issues, Jennifer O'Day for her leadership of the study, and Megan Petroccia for her support in analyses and production. We also recognize the assistance of Kenji Hakuta, of Stanford University, for his advice on analytic approaches and data issues. Furthermore, we are grateful for the expertise of David Francis, of the University of Houston, Edward Haertel, of Stanford University, and Scott Marion, of the Center for Assessment, as part of the study's technical working group. Of course, any errors in judgment or fact remaining are those of the authors.

Executive Summary

The *Elementary and Secondary Education Act (ESEA)*, as amended by the *No Child Left Behind Act of 2001* inaugurated important changes in assessment and accountability for English Learner (EL) students. Specifically, Title III of the law required states to develop or adopt English-language proficiency (ELP) standards aligned with language demands of academic content standards. An annually administered ELP assessment based on those standards was also required by the *ESEA* (*NCLB* 2002). Title III also instituted new accountability requirements for districts and states. These new EL accountability provisions required states to define criteria for progress in learning English, establish a performance standard for English proficiency, and set annually increasing performance targets for the number and percentage of ELs meeting these criteria.

As has been well documented, the new law's requirements exceeded the technical capacity of many states and districts to comply with it (Abedi 2004; Government Accountability Office 2006). Over the past several years, empirical research with more rigorous ELP assessments, systematic technical assistance efforts, and federal guidance have helped to reduce confusion and increase coherence in state Title III accountability systems (Abedi 2007; Linquanti and George 2007; Cook et al. 2008; Federal Register 2008). Nevertheless, a significant need remains to develop the capacity of state and technical assistance providers to utilize empirical data for performance standard setting and accountability policy development in these areas. Even as *ESEA* reauthorization draws closer, prevailing civil rights laws and sustained focus on improving EL education suggest these policy-making needs will grow in importance, particularly given broad adoption of Common Core State Standards and establishment of related multistate academic and ELP assessment consortia.

This document is intended to contribute to that capacity development by describing and illustrating several empirical methods and conceptual or theoretical rationales to help state policy-makers, standard-setting panels, and the technical advisory panels and assistance providers to (1) determine a meaningful ELP performance standard; (2) establish a realistic, empirically anchored time frame for attaining a given ELP performance standard; and (3) take into account an EL's ELP level when setting academic progress and proficiency expectations. This is *by design* a technical document intended to assist those charged with providing empirical information germane to developing or revising EL accountability models, using ELP and academic assessments.

This volume does not focus on several additional basic issues around EL student achievement because there is a companion volume (Taylor, Chinen, et al., forthcoming) that analyzes similar state- and district-provided student-level longitudinal achievement data and descriptively addresses those issues. That companion volume describes (1) the heterogeneity of the EL population and the different achievement statuses and trajectories of ELs with different characteristics; (2) the estimated achievement gaps among ELs, former-ELs and non-ELs; (3) a basic description of the typical time frame for attaining English language proficiency; and(4) the nature of the relationship between assessment scores measuring language acquisition and academic-content-area learning. The chapters of this current volume represent logical extensions building on those more basic descriptive analyses.

Chapter I positions data-analysis methods illustrated in the report within a larger deliberative process of setting meaningful, ambitious, and realistic performance standards and accountability criteria for EL students. The chapter offers guidelines for enacting best practices in standard setting, and highlights limitations in using empirical data.

Chapter II illustrates three methods (decision consistency, logistic regression, and descriptive box plots) for analyzing empirical data to assist policymakers in determining an ELP performance standard for English Learners. These methods are used in conjunction in three states, and the results—which vary in their degree of convergence within each state—are interpreted to show how they might be utilized to support each state's decision-making process.

Chapter III illustrates methods for conducting empirical analyses to inform setting expected time frames for EL students to attain the ELP performance standard. Specifically, a descriptive approach and an event history approach are applied with various adjustments. Results are compared for EL students in different grade clusters with different levels of initial English language proficiency. As students' initial ELP level influences the expected time frame for their attaining the English-proficient criterion, these data are used to illustrate the ways in which more refined time-to-English-proficiency criteria could be derived.

Finally, Chapter IV discusses two methods for taking into account an EL's ELP level when setting academic progress and proficiency expectations, and one method that explicitly ignores it. First, progressive benchmarking methods are illustrated that adjust either EL students' content achievement scale scores or their weight (individual "count"), based on each student's ELP level relative to their initial ELP level and time in the state school system. Second, an indexed progress method utilizes ELs' ELP growth as a proxy for English language arts performance on a weighted, time-sensitive basis for more newly-arrived ELs who enter the state's school system at lower initial ELP levels. Third, a status and growth accountability matrix method credits both a predetermined level of student academic growth as well as attainment of academic proficiency, without considering an English Learner's ELP level. Each method is carefully described and applied using the same education agency's sample data set.

All the approaches presented in this document—many of which have been employed by the principal authors in working with states on their EL accountability systems—are intended to stimulate discussion and further exploration of additional methods among state data analysts, technical assistance providers, and researchers. The ultimate goal is to support the development and regular use of empirical methods that inform ambitious, realistic, and meaningful performance standards and accountability policies, which will foster EL students' linguistic and academic progress and attainment.

I. Introduction

The *Elementary and Secondary Education Act (ESEA),* as reauthorized by the *No Child Left Behind Act of 2001*, inaugurated important changes in assessment and accountability for English Learner (EL) students. Specifically, Title III of the *Elementary and Secondary Education Act (ESEA)* required states to develop English-language-proficiency (ELP) standards aligned with language demands of academic content standards, and an ELP assessment based on those ELP standards that would measure English Learners' progress in developing the language needed to attain academic proficiency (Public Law 107-110). Moreover, Title I and Title III of the amended *ESEA* required states to assess each EL's ELP annually. Title III also required states to define criteria for progress in learning English, establish a performance standard for the English-proficient level and set annually increasing performance targets for the number and percentage of ELs meeting these criteria. Specifically, states were required by 2003 to set annual measurable achievement objectives (AMAOs) for the percentage of ELs in each Title III–funded local education agency making progress (AMAO 1), attaining the English-proficient level (AMAO 2), and attaining the adequate yearly progress (AYP) targets in English reading or language arts (ELA) and mathematics, as required of the EL subgroup for school districts under Title I (AMAO 3).

As has been well documented, the law's requirements exceeded the technical capacity of many states and districts to comply with it (Abedi 2004; Government Accountability Office 2006). Most states lacked empirical data—and indeed, even standards-based ELP assessments—with which to determine progress criteria and performance targets. The majority of states initially used off-the-shelf ELP assessments and set arbitrary, often unrealistic performance criteria and targets for AMAOs 1 and 2. For example, some states expected ELs to progress equally in every domain annually, while others judged progress as advancing in any language domain, or used mean scale score gains of the entire cohort. One state set the highest language proficiency level as its expected performance standard for all ELs; another implicitly allowed ELs under AMAO 1 to progress for 10 years, but required them under AMAO 2 to attain English proficiency within five years. Many states assumed that AMAO 1 and 2 targets needed to reach 100 percent by 2014, as with Title I AYP targets.

Over the past several years, empirical research with more rigorous ELP assessments, systematic technical assistance efforts, and federal guidance have each helped to reduce confusion and increase coherence in state Title III accountability systems (Abedi 2007; Linquanti and George 2007; Cook et al. 2008; Federal Register 2008). These efforts notwithstanding, there is still a significant need to develop the capacity of state and technical assistance providers to utilize empirical data for performance standard setting and accountability policy development in these areas. The current document seeks to contribute to that capacity development.

Specifically, this document offers several empirical methods that state policy-making authorities can use as part of a larger deliberative process to set ELP performance standards and operationalize ELP assessment and accountability criteria. This document describes and illustrates several empirical methods and conceptual or theoretical rationales to help state policymakers, standard-setting panels, and the technical advisory panels and assistance providers supporting them to

—Determine a meaningful ELP performance standard.

—Establish a realistic, empirically anchored time frame for attaining a given ELP performance standard.

—Take into account an EL's ELP level and time in the school system when setting academic progress toward proficiency expectations.

This is *by design* a technical document intended to inform those charged with implementing a larger policy-making judgmental process. The intended audience is state assessment and accountability directors with responsibility for overseeing state performance standard setting, and establishing progress criteria and performance target structures. Secondary audiences include technically inclined Title I and Title III state program directors, senior state education agency leaders, technical assistance providers, and those directly advising governance boards and state boards of education.

This volume, also by design, does not focus on several additional basic issues around EL student achievement because there is a companion volume (Taylor, Chinen, et al., forthcoming) that analyzes similar state- and district-provided student-level longitudinal achievement data and descriptively addresses those issues. That companion volume describes (1) the heterogeneity of the EL population and the different achievement statuses and trajectories of ELs with different characteristics; (2) the estimated achievement gaps among ELs, former-ELs and non-ELs; (3) a basic description of the typical time frame for attaining English language proficiency; and (4) the nature of the relationship between assessment scores measuring language acquisition and academic-content-area learning. The chapters of this current volume represent logical extensions building on those more basic descriptive analyses.

Using Multiple Methods and Empirical Data to Set Performance Standards

The data methods illustrated in this report offer tools for states to use as part of a larger policy process of setting meaningful, ambitious, and realistic performance standards for EL students. The testing and measurement field has developed a substantial literature around methods, procedures and protocols for appropriately setting performance standards (e.g., see Hambleton and Pitoniak 2006 for a review of these).

Establishing performance standards is not of concern just to testing and measurement professionals; it is also of interest to those who establish criteria for educational accountability. Educational accountability requirements are like performance standards, but instead of being for individual students, they are for educational entities such as schools, districts or states. When states established accountability models under the requirements of the *No Child Left Behind Act of 2001,* they employed many common performance-standard-setting procedures. In fact, federal guidance requires that states employ common standard-setting procedures when developing *ESEA* Title I accountability systems.[1] The following paragraphs briefly review common standard-setting practices and discuss the ways in which they might be used to support establishing the accountability requirements for Title III (i.e., AMAOs 1 and 2).

Hambleton and Pitoniak (2006, 435), in their chapter on setting performance standards in *Educational Measurement, 4th Edition*, note, "[T]he setting of performance standards is a blend of judgment, psychometrics and practicality." The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association and National Council of Measurement in Education 1999, 54) also state, "[D]etermining cut scores … cannot be purely a technical matter, although empirical studies and statistical models can be of great value in informing the

1. See U.S. Department of Education, Office of Elementary and Secondary Education. (Jan. 12, 2009). Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001. Downloaded at http://www2.ed.gov/policy/elsec/guid/saaprguidance.pdf, p.27 (Critical Element 2.6).

process." Given high-stakes consequences that flow from performance standards, assessment and accountability experts have warned that outcomes of performance-standard-setting processes can vary by the way in which judges are selected and trained, and by methods used to set standards (Linn 2003; Haertel, 2002; Haertel 2008). Thus, a foundational principle to consider when establishing standards (or accountability models) is that standard setting is not merely an empirical procedure; rather, it requires *both* empirical information germane to the intended uses of the performance standard and informed judges engaged in a rational, coherent and transparent deliberative process. These requirements cannot be overemphasized in setting expectations for EL performance, given the complex interrelationship of second-language progress and proficiency to academic-content-area progress and proficiency.

Hambleton and Pitoniak (2006, 436) also note, "[S]ometimes it is suggested that two or even three [standard-setting] methods be implemented so that the results can be compared, but this is usually an expensive and time-consuming process. …" That is, if feasible, more than one method should be employed. Another foundational principle to consider when establishing standards is to use several approaches or methods to the extent practicable. Multiple approaches provide more information for the judgment process. Akin to Hambleton and Pitoniak's point, the *Standards for Educational and Psychological Testing* (AERA, APA and NCME, 1999, 146) standard 13.7, states,

> In educational settings, a decision or characterization that will have a major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision.

The notion of using more than a single data point as a basis for important decisions about students is also applicable to teachers, schools and districts. Additional empirical evidence can be used to support important assessment and accountability policy decisions, such as establishing performance criteria and reasonable targets for AMAOs.

Hambleton and Pitoniak provide further guidance via a list of procedures to follow in order to develop reasonable performance standards. They write, (436) "[T]he defensibility of the resulting performance standard is considerably increased if the process reflects careful attention to: (1) the selection of method; (2) the selection and training of panelists; (3) the sequence of activities in the process; (4) validation; and (5) careful documentation of the process." The authors argue that these procedures will support the defensibility (acceptability) of newly established performance standards.

An appropriate standard-setting method (or multiple methods) should be used to support establishing standards (accountability criteria). Panelists (or judges) should be selected and trained. Regarding panelists, Standard 1.7 of the *Standards for Educational and Psychological Testing* requires that they have appropriate qualifications (i.e., experiences to support setting standards, lack of vested interest in standards set, and relevant training). A clear process for establishing standards should also be developed and implemented. A mechanism should be in place to validate the decisions made by established criteria, and the entire process and decisions made should be carefully documented and reported.

Applied to Title III AMAOs, the following guidelines, derived from best practices in performance-standard setting, can productively support establishing reasonable ELP performance criteria and, as argued here, strong accountability policies:

—To the extent practicable, employ a variety of analytic methods and procedures when establishing AMAO criteria.

—Select analytic methods that appropriately inform the decisions to be made on the basis of intended uses of the performance standards; this includes the generation of outcome data to model likely outcomes based on current student performance.

—When establishing AMAOs, use both empirical data analyses and expert judges in a well-defined deliberative process.

—Select panelists who are sufficiently qualified (e.g., expert and experienced with ELs) to support setting AMAO-related criteria.

—Train panelists adequately. In that training, panelists should understand that different analytic methods will likely yield different results. It is an essential part of their charge to utilize this information and their expert judgment to make the best possible decision.

—Clearly define and follow a rational and coherent sequential process for establishing AMAOs.

—Document the process clearly. This documentation should provide relevant stakeholders information about how AMAOs were set; who panelists were and why and how they were selected; what analyses and procedures were used; what options and scenarios were considered; what deliberations occurred, and what final recommendations were made.

—Design validity studies on newly established AMAO criteria. Such research is necessary for examining outcomes (consequences) relative to the intent of criteria.

Adherence to these guidelines will support more realistic and productive decisions on ELP performance criteria and AMAO accountability policies. As Mehrens and Cizek (2001, 484) note,

> In one way or another, setting performance standards is unavoidable. Categorical decisions will be made. These decisions can be made capriciously or they can be accomplished using the sound procedures at hand today, or via the almost assuredly better methods that continue to be the product of psychometric research and development.

In recent years, many states have revisited and revised their ELP assessments, ELP performance standards, and AMAO accountability provisions. While it is beyond the scope of this document to fully illustrate the guidelines delineated above, the material that follows will provide several analytic methods applied to authentic state and district data sets. These worked examples illustrate ways to support establishing rigorous and realistic ELP performance criteria and meaningful AMAOs.

Limitations to Empirical Approaches

As this document illustrates several empirical methods to support policy decision making, it is important to note the limitations of these methods and the value of using several appropriate methods in conjunction, when possible. No one empirical method gives a complete picture. As Box and Draper famously note regarding statistical models (1987, 74), "Essentially, all models are wrong, but some are useful." Statistical models are imperfect representations. The methods used to support decision making seldom provide definitive answers. Nonetheless, they do provide valuable evidence to inform deliberations and decision making and to understand the potential consequences of decisions.

In a related vein, data are almost always imperfect. For example, as will be seen in the following chapters, data sets used for AMAO analyses often have missing cases, and reasons for missing data may not be apparent. Furthermore, some data patterns are influenced by the way the EL construct is defined and operationalized. Two key implications of the latter phenomenon merit attention here:

First, EL status is intended to be temporary and to change as a direct result of high-quality language-instruction educational services. Therefore, more successful ELs exit language instructional programs as they reach required levels of ELP (and often, academic performance). A natural consequence of this fact is that faster progressing, higher attaining students exit the EL category sooner than slower growing students.

Second, students who reach English proficiency no longer participate in the state language-proficiency assessments. As a result, when looking at longer term, longitudinal results for current ELs, the observed growth is likely an underestimate relative to that of the total cohort of students who entered the system needing to learn English as a second language, because it contains only the results of those EL students who continued not to meet exit criteria.

It is therefore important for panelists to understand these kinds of limitations in all analytic methods and data utilized in policy-setting deliberations. The examples provided in the following chapters attempt to illustrate how several empirical methods can be used in combination with actual state and school district data, and how the limitations of both methods and data can be highlighted for decision makers that utilize them.

II. Determining an English-Language-Proficient Performance Standard

Overview

This chapter illustrates how empirical data might be analyzed to assist policymakers in determining an English-language-proficient (ELP) performance standard for English Learners (EL). Specifically, the chapter presents three methods for analyzing data related to this issue, applies these methods in three states, and discusses how the results might be interpreted and utilized to support decision making. As the analyses to come illustrate, state context matters greatly. Applying the methods in states with different EL populations and different ELP and academic assessments and cut scores will yield different degrees of convergence across the methods, and results need to be interpreted carefully within this context. However, data from these methods can ground policy deliberations and help clarify the implications of different options under consideration.

Implicit in analyses presented in this chapter are two assumptions: (1) State academic-content-area performance standards are generally set independent of ELP performance standards, so empaneled judges must accept the current performance standards for a state's academic assessments as unmodifiable; and (2) the state's academic-content performance standards have been established appropriately and will lead to valid inferences of students' academic knowledge, skills, and abilities. Interpretations of findings from the methods shared below hinge particularly on this second assumption. If content-area performance standards are set inappropriately or validity concerns exist about the performance standards or content assessment from which performance standards were created, the methods shared below will be correspondingly compromised.

The *Elementary and Secondary Education Act of 1965 (ESEA),* as amended by the *No Child Left Behind Act of 2001 (NCLB),* defines a limited-English-proficient (LEP) student[2] as an elementary or secondary school student

> whose difficulties in speaking, reading, writing, or understanding the English language may be sufficient to deny the individual the ability to meet the *State's proficient level of achievement on State* [*academic*] *assessments* [italics added] described in section 1111(b)(3); the ability to successfully achieve in classrooms where the language of instruction is English or the opportunity to participate fully in society. §9101(25)(D)

This definition implies that a key indicator of having sufficiently addressed the linguistic needs of EL students is their performance on state content assessments, specifically, how able these students are to attain the state's proficient performance standard on its academic content assessments. *NCLB* requires that states' English-language-proficiency (ELP) standards and associated ELP assessments be "aligned with" academic content and performance standards (§3113(b)(2)). These aspects of the federal law imply an expected relationship between students' ELP and levels of academic proficiency when content is assessed in English. Moreover, this relationship is reinforced in the recently announced federal enhanced assessment grant program for next-generation ELP assessment systems. These new ELP assessments are expected to "indicate whether individual [EL] students have attained the English proficiency necessary to

2. Federal law uses "limited English proficient," or LEP, to designate linguistic-minority students whose English-language skills inhibit their ability to benefit from mainstream instruction in English. In the research literature and in most states, the term "English Learner" (EL) is used, as it is in this report.

participate fully in academic instruction in English and *meet or exceed college- and career-ready standards* [italics added]" *(Federal Register* 2011, 21978).

State policymakers, therefore, need to examine the relationship between their state's ELP assessments and academic-content assessments as they determine what levels of linguistic and academic performance will be used to operationalize a definition of EL.[3]

To date, studies investigating the relationship between ELP and academic-content performance have found positive relationships between both assessment types.[4] These studies use either correlational or regression-based approaches to confirm relationships between ELP and academic-content assessments.

This chapter illustrates how states can examine the relationship between EL performance on ELP assessments and academic-content assessments for the purpose of identifying where a proficient performance standard on the ELP assessment might be established. States are required to define this performance standard under Title III of *ESEA* (specifically, to address the title's second annual measurable achievement objective [AMAO 2]), and several states have used empirical methods to explore this definition (see Linquanti and George 2007; Cook and others 2008). To do so, policymakers must first clarify what is meant by the term *English-language proficient*.

The federal definition of limited English proficiency specified in the law suggests that, when students' English-language skills are sufficient to (1) no *longer be denied* the ability to meet a state's proficient performance standard on its academic-content assessments and (2) be *able* to achieve success in English-only classrooms, these students may be classified as fully English proficient and exited from specialized language and academic support services. That is, when ELs' English proficiency no longer inhibits their meaningful participation on state assessments or in the classroom using English, they may be classified as fully English proficient. Note that the federal definition does *not* require that ELs be academically proficient in order to be classified as fully English proficient. Clearly, many native-English-speaking students are also not proficient on state content assessments. These students' lack of academic proficiency may not be related at all to their English-language skills. ELs, therefore, must have sufficient ability in academic English to meaningfully participate in the classroom and on content assessments.[5]

Empirically, researchers can define "English language proficient" as the point at which EL students' academic content achievement assessed using English becomes *less* related to their ELP. That is, there is a point at which EL students have sufficient English language skills to adequately function in English on content assessments; accordingly, there should be observable decreases in the relationship between the two assessments. At or beyond this point is where the ELP performance standard might be considered, and empirical procedures can help to identify this level of performance. Because academic language demands vary by academic content area and grade level, this performance point will likely vary as well. Yet state policymakers are usually required to select one ELP performance level. They, therefore, need to examine the data carefully to clarify tradeoffs and attempt to make an optimal decision.

3. Although *ESEA* Title I specifically requires states to set a goal of 100 percent of all students' attaining academic proficiency in reading or language arts and mathematics by the 2013–14 school year, recent policy discussions regarding *ESEA* reauthorization emphasize more empirically based growth models and performance targets (see Linn 2008).

4. See Stevens and others (2000), Butler and Castellon-Wellington (2000 and 2005), Kato and others (2004), Francis and Rivera (2007), Parker, Louie, and O'Dwyer (2009), Cook and others (2009), and Taylor and others (forthcoming).

5. Correspondingly, some EL policy experts argue that the academic performance of ELs attaining ELP should be *comparable* to that of their native-English-speaking counterparts, notwithstanding the fact that EL students generally experience higher poverty rates and often have proportionally fewer instructional resources (Working Group on ELL Policy 2010).

Key Approaches

We used three approaches to explore the relationship between ELP level and meeting the state's grade-level performance standard on academic-content assessments:

1. **Decision consistency analysis,** which analyzes linguistic and academic proficiency-level categorizations and seeks to optimize consistent categorization of ELP students at the state's preestablished academic proficient cut score.[6]
2. **Logistic regression analysis,** which estimates the probability of being proficient on academic-content assessments for each ELP score.[7] This approach could identify ELP scores for which students have a probability of equal to or greater than 50-50 (0.5) of being proficient on the content assessment.[8]
3. **Descriptive box plot analysis,**[9] which identifies the ELP level at which at least half the assessed EL students are above the academic-content proficient score cut point.[10] At this point, students equally distribute above and below the state's proficient performance standard in academic content, which may suggest that, above this point, more than just language proficiency is contributing to observed scores.

Taken together, these three approaches provide multiple sources of evidence to investigate and corroborate the point at which an ELP performance standard might be set. All three approaches should be used, when feasible, in order to provide policymakers with more complete, possibly "triangulated" empirical evidence for delimiting a range of performance and defining options to establish an ELP performance standard for ELs.

The following section applies these analytic methods to examine data for EL students within each of three very different states at different grade levels for two academic years.[11] Specifically, the analyses illustrate grade 4 outcomes from Education Agency 1 for 2007–08 and 2008–09, grade 7 outcomes from

6. As this approach is relatively new, a detailed description and step-by-step illustration of it is provided in Appendix A.

7. For this approach, the outcome is a dummy indicator that equals 1 if the student is proficient on academic-content assessment and 0 otherwise. The predictors of the model are dummy indicators, which take the value of 1 if the student's score falls within the particular ELP proficiency level and 0 otherwise.

8. The English-proficient performance standard is conceptualized here as the point where students' language proficiency becomes less related to content proficiency. A 50-50 criterion is selected because students with ELP scores at this level have an equal likelihood of attaining content-area-proficient performance. ELP assessments are not designed or intended to strongly predict content proficiency per se but rather to ascertain if students have the requisite language needed to meaningfully participate in academic-content-area learning using English. Certainly, a different criterion could be adopted. States should be careful, however, not to set the criterion too high. If our assumption about the English-proficient performance standard is correct, then ELP assessments will be less predictive of content proficiency at higher ELP levels. This imprecision would add greater error and might lead to inaccurate inferences. The imprecision might also lead to establishing higher content performance expectations for EL exit decisions than the performance currently attained by non-EL students.

9. A box plot shows graphically five-number summaries: the smallest observation, lower quartile, median, upper quartile, and largest observation. The box plot graph may also indicate outliers observations. But for simplicity, the following graphs do not show outliers observations.

10. There are significant concerns (including noncomparability) surrounding states' selected academic-proficient performance standards for high-stakes accountability purposes (Ho 2008; Dietz 2010). However, the present study assumes that the sample states' academic-proficient performance standards are rigorous and defensible.

11. Results from both years' analyses were very similar within each state; therefore, only the more recent year is discussed here, and all results for both years analyzed are provided in appendixes B–D.

Education Agency 2 for 2007–08 and 2008–09, and grade 10 outcomes from Education Agency 3 for 2009–10 and 2010–11.

Each "worked example" is illustrated in turn, employing all three methods to explore ranges of EL performance on the ELP assessment and the relationship of these ranges to content-area assessment performance. The purpose of doing this is to provide input from multiple methods to inform policy discussions on determining an acceptable ELP performance standard.

State results are shown to illustrate how the different methods are used together to examine relationships between assessment results *within* a state, and how policymakers can interpret these results to establish options. *Readers should not compare outcomes across states:* Each state has different content and ELP assessments based on state-specific content and performance standards, and each state administers its assessments at different times and utilizes different scaling and equating methodologies. Thus, drawing comparisons and inferences across states would be inappropriate and misleading.

Example 1. Education Agency 1

This first example shows results from Education Agency 1. Given in grades 2 through 11, the state's academic content test assesses English or language arts (ELA) and mathematics, and is administered in the spring of each school year. It provides five content performance standard categories: Far Below Basic, Below Basic, Basic, Proficient, and Advanced. The state ELP assessment is administered from mid-summer through early fall of each school year. This assessment provides five ELP performance levels: Beginning, Early Intermediate, Intermediate, Early Advanced, and Advanced.[12]

Method A. Decision Consistency Analysis

As explained, the decision consistency analysis (Exhibit 1, below) illustrates the cumulative percentage of consistent decisions derived from ELP and academic-content-assessment classifications. Assessment results provided are from fourth-grade students who took both exams in the same academic year. The objective of this approach is to identify the ELP performance in which the maximum percentage of consistent decisions is found (see Exhibit 1, below).

As described in Appendix A, the first step in a decision consistency analysis is to create comparative bands for more refined analysis, as performance levels on standardized large-scale assessments can represent a broad range of performance. Accordingly, 10 comparative bands are created from the state ELP assessment composite proficiency scores to provide sufficient gradation for analysis: Beginning Low, Beginning High, Intermediate Low, Intermediate High, and so on. For each band, a decision consistency value is calculated, and those values are plotted in Exhibit 1. In this example, the greatest percentage (78 percent) of consistent decisions for fourth-grade ELs in ELA is obtained at the Early Advanced Low (Early Adv Low) performance level, while for mathematics, the greatest percentage (66 percent) occurs in the upper half of the scale score range of the Intermediate High (Int High) performance band. This analysis suggests that state policymakers consider the low end of Early Advanced as a possible range for optimizing consistent decisions, given the state's current academic

12. The scale score ranges for the Beginning, Early Intermediate, and Intermediate proficiency levels of the overall scores for grade 4 are 230–432, 433–472, and 473–530, respectively. The ranges for the Early Advanced and Advanced proficiency levels are 531–574 and 575–700, in that order.

performance standard at this grade level.[13] Relevant grade levels can be analyzed similarly, and results from analyses can be aggregated and presented for deliberations.

Exhibit 1.
Education Agency 1, Grade 4: ELP and English or Language Arts and Mathematics Decision Consistency Analysis (2007–08)

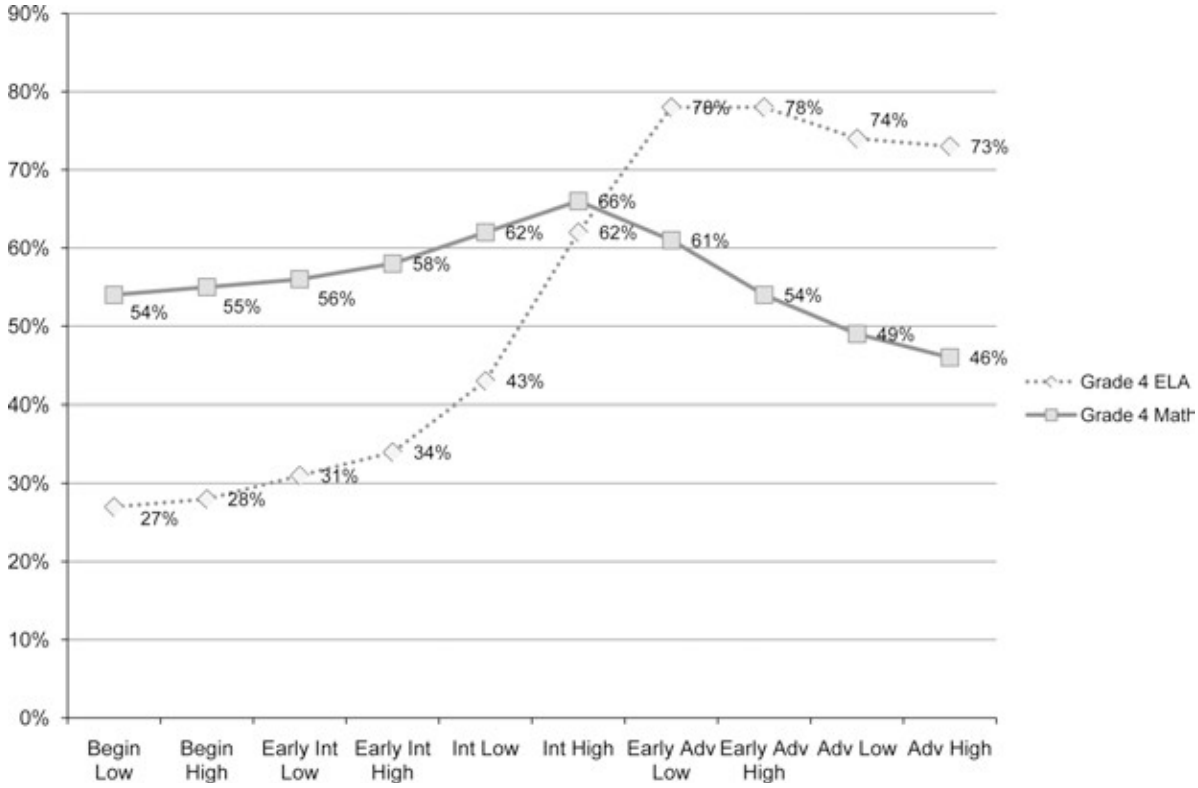


Exhibit reads: The percent of consistent decisions obtained for fourth-grade EL students in mathematics up through the High Intermediate ELP performance level is 66 percent. In English or language arts, the cumulative rate up through the lower Early Advanced scale score range is 78 percent.

Notes: The composite ELP scale score ranges for the Beginning (level 1), Early Intermediate (level 2), and Intermediate (level 3) proficiency levels of the overall scores for grade 4 are 230–432, 433–472, and 473–530, respectively. The ranges for the Early Advanced (level 4) and Advanced (level 5) proficiency levels are 531–574 and 575–700, in that order. The scale score point at the midpoint of each proficiency level range is then demarcated. Students below the midpoint are classified as "low." Students at or above that point, are classified as "high." Appendix Exhibits B.5 and B.6 present the number of students with ELP and English or language arts and mathematics proficiency level scores and the percentage of consistent decisions.

Decision consistency formula: DC% = (QII + QIII)/(QI + QII + QIII + QIV).

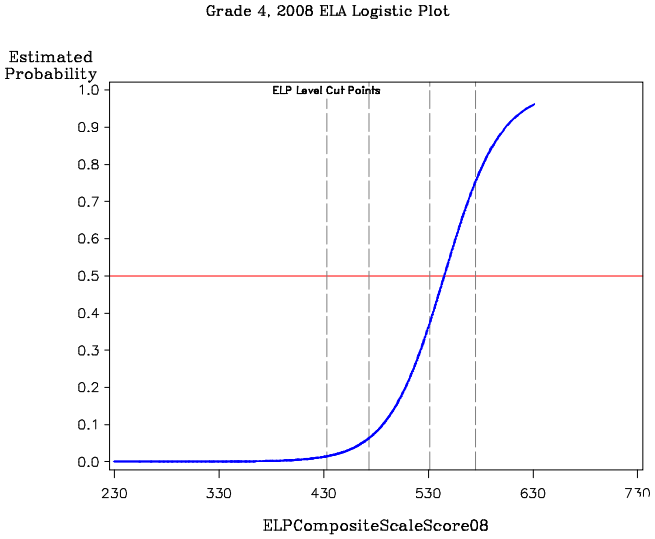
Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

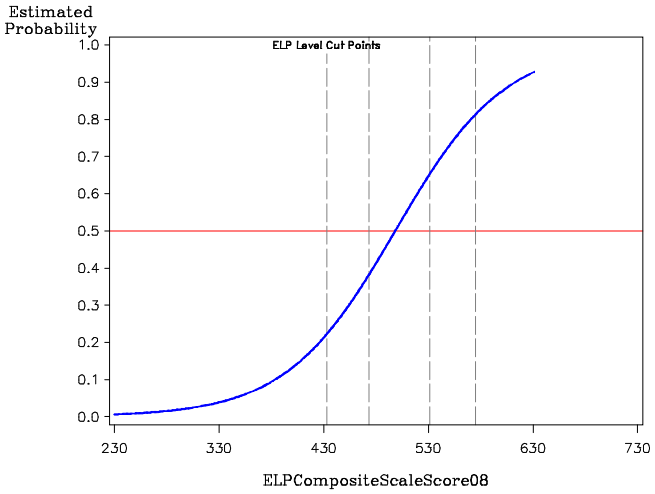
13. As noted above, state policymakers are usually required to select one English-language-proficient performance standard to apply to ELs across all grade levels, and so would identify the higher ELP performance standard to ensure adequate performance on ELA, as well as mathematics.

Method B. Logistic Regression Analysis

The logistic regression (probability) curves in Exhibit 2 (below) illustrate the likelihood of scoring at or above the academic proficient performance standard, as currently defined by the state for ELA and mathematics, respectively, as a function of increasing composite ELP scale scores. The horizontal line in the middle of each exhibit marks the point at which there is an equal (50-50) probability of attaining that academic performance standard, while the vertical lines mark the scale score cut points that distinguish the five ELP performance levels on the state ELP assessment. Similar to the decision consistency analyses, these regression analyses suggest that the state ELP assessment scale score value (545 points) corresponding to the lower half of the Early Advanced ELP performance level, and that corresponding to the midpoint of the Intermediate ELP level (501 points) would be sufficient to obtain that likelihood of performance in ELA and mathematics, respectively. This approach yields corroborating evidence and would therefore increase confidence in suggesting the lower end of Early Advanced as a range to consider for the English-proficient performance standard.

Exhibit 2.
Education Agency 1, Grade 4: Logistic Regression Plots for English or Language Arts and Mathematics (2007–08)





ELP performance levels for grade 4 (in scale score points)

Level 1 = Beginning (230–432)
Level 2 = Early Intermediate (433–472)
Level 3 = Intermediate (473–530)
Level 4 = Early Advanced (531–574)
Level 5 = Advanced (575–700)

Exhibit reads: There is an equal probability for grade 4 EL students who score in the lower half of the Early Advanced scale score range to achieve the proficient performance standard in English or language arts, specifically, when obtaining 545 scale score points.

Notes: The plot represents the estimated logistic curve where the outcome is the dummy indicator English or language arts proficient and the predictor is the continuous ELP composite scale score. The logistic regression (probability) curves illustrate the likelihood of scoring at or above the academic proficient performance standard, as currently defined by the state for ELA and mathematics, respectively, as a function of increasing composite ELP scale scores.

The vertical dashed lines correspond to minimum scale scores for Early Intermediate, Intermediate, Early Advanced, and Advanced proficieny levels.

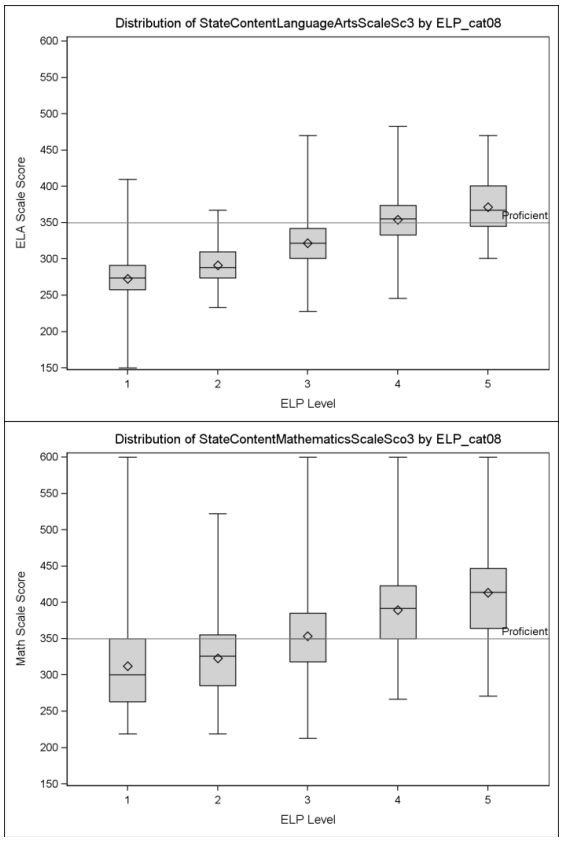
The point estimate and standard error are presented in Appendix Exhibit B.8.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Method C. Descriptive Box Plot Analysis

Finally, the box plot analyses for ELA and mathematics (Exhibit 3) show the distribution of scale score performance in those academic subject areas, respectively, for students at each of the state ELP assessment's five composite ELP performance levels. The data reveal findings congruent with the decision consistency and logistic regression analyses. Specifically, as seen in Exhibit 3 (below), EL students performing at the Early Advanced ELP performance level (or level 4) have a median ELA performance of 355 scale score points (just above the proficient standard in ELA), and more than 50 percent of ELs at this ELP performance level attain this academic performance standard. For mathematics (Exhibit 3), a similar result (median performance of 350 scale score points) occurs for ELs at the Intermediate ELP performance level (or level 3), indicating that exactly half the EL students at this ELP performance level attained the mathematics academic performance standard.

Exhibit 3.
Education Agency 1, Grade 4: Box Plots of English or Language Arts and Mathematics Scale Scores, by ELP Performance Level (2007–08)



ELP performance levels
for grade 4 (in scale score points)

Level 1 = Beginning (230–432)
Level 2 = Early Intermediate (433–472)
Level 3 = Intermediate (473–530)
Level 4 = Early Advanced (531–574)
Level 5 = Advanced (575–700)

Exhibit reads: Grade 4 EL students performing at the Early Advanced ELP performance level (or level 4) have a median performance of 355 scale score points in English or language arts, five points above the proficient performance standard.

Note: Appendix Exhibit B.10 presents the descriptive statistics associated with this graph.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

For Education Agency 1, results of all three methods converge, and suggest that policymakers might consider setting the ELP performance standard in the Early Advanced range of the state ELP assessment. Doing so would maximize the percentage of consistent decisions relative to the students' ELA performance, and would also—per regression analysis—yield a 50 percent or greater probability that EL students attaining this level of language proficiency on the state ELP assessment would also attain the current proficient standard on the state's ELA test. As corroborated by the box plots, slightly more than 50 percent of EL students at the Early Advanced level on the ELP assessment exceed the proficient performance standard on the state's ELA test.

Performance on mathematics appears to require less ELP, as consistent decisions are maximized in the upper half of Intermediate on the state ELP assessment, with the logistic regression curve indicating that the fourth-grade EL students above the midpoint of the Intermediate level exceed the 0.50 probability of attaining mathematics proficiency.The box plots indicate that fully 50 percent of students at the Intermediate ELP level attain proficiency on mathematics in the fourth grade. Since policymakers must choose only one performance standard for the ELP level, these analyses suggest a need to consider the higher ELP level indicated by the ELA analyses.

Example 2. Education Agency 2

The state content test used for Education Agency 2 is administered in grades 3 through 8 in the areas of literacy and mathematics in the spring of each school year. These assessments provide four performance categories: Below Basic, Basic, Proficient, and Advanced. Education Agency 2 is one of a consortium of states that shares an English language-proficiency assessment. This assessment is administered in the late spring of each school year and provides five ELP performance levels: Pre-Functional, Beginning, Intermediate, Advanced, and Fully English Proficient.

Because the ELP assessment composite proficiency levels are not created from composite scale scores, creating more than five comparative bands of performance for decision consistency analysis is not possible.[14] Therefore, the ELP assessment's five proficiency levels are used.

Method A. Decision Consistency Analysis

Decision consistency analyses for Education Agency 2's EL students in 2008 (Exhibit 4, below) show that the greatest percentage of consistent decisions for seventh-grade ELs in literacy (in the exhibit legend, identified as ELA) (77 percent), as well as mathematics (76 percent), is obtained at the *same* ELP performance level—namely, the Advanced level on the ELP assessment, the fourth of five performance levels defined on this ELP test. This finding holds for both years examined (with slightly higher results—81 percent—for ELA occurring in 2007). In addition, the percentage of consistent decisions in Education Agency 2 at this grade level varies little between the academic subject areas.

14. Composite scale scores are not aligned with the composite proficiency levels. While the former is created by averaging results from the four domains, the latter is generated using a weighting system that combines proficiency-level results from two synthetic categories—comprehension (composed of listening and reading domains) and production (composed of speaking and writing domains).

Exhibit 4.
Education Agency 2, Grade 7: ELP and Literacy/Mathematics Decision Consistency Analysis (2007–08)

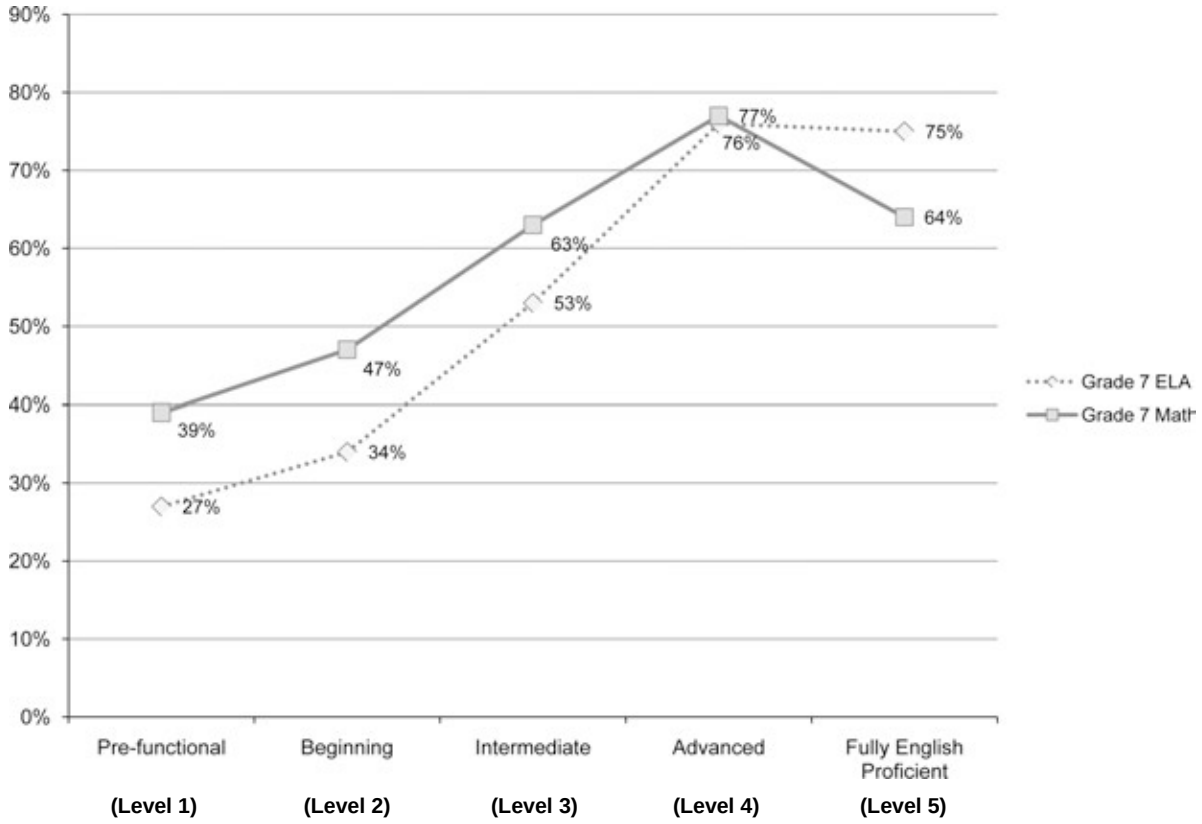


Exhibit reads: The percent of consistent decisions obtained for seventh-grade EL students in mathematics up through the Intermediate ELP performance level is 63 percent. In English or language arts, the cumulative rate up through the Advanced scale score range is 77 percent.

Notes: ELP composite scale scores are not aligned to the composite proficiency levels in Education Agency 2. Although the former is created by averaging results from the four domains, the latter is generated using a weighting system that combines proficiency-level results from two synthetic categories—comprehension (composed of listening and reading domains) and production (composed of speaking and writing domains). Appendix Exhibits C.5 and C.6 present the number of students with ELP and English or language arts and mathematics proficiency-level scores and the percentage of consistent decisions.

Decision consistency formula: DC% = (QII + QIII)/(QI + QII + QIII + QIV).

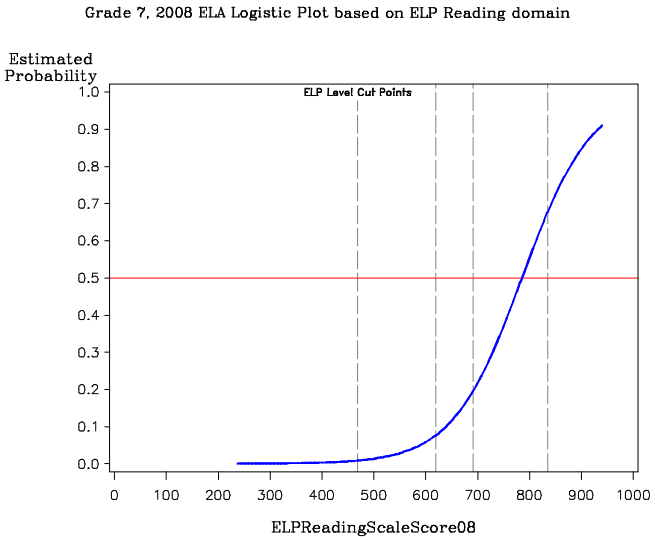
Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

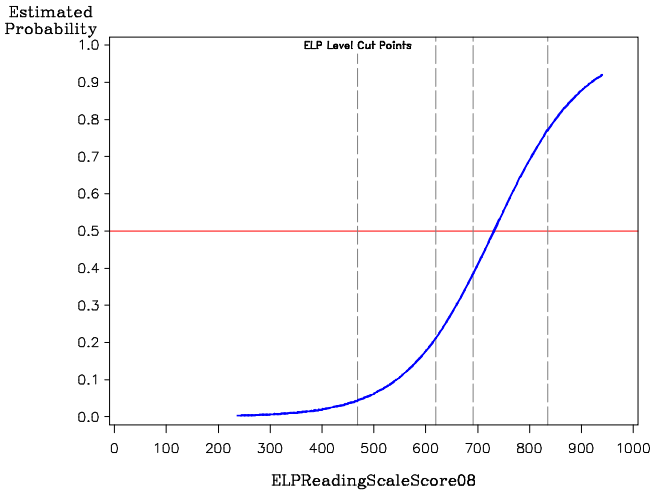
Method B. Logistic Regression Analysis

The logistic regression (probability) curves for Education Agency 2's seventh-grade EL students are displayed in Exhibit 5 (below). Given the way the ELP assessment constructs its overall composite scale scores, they are not used in this analysis.[15] Instead, the ELP assessment reading scale score result is used, as the reading domain has been found to be the domain with highest predictive validity on some ELP assessments (e.g., see Parker, Louie, and O'Dwyer 2009). Like decision consistency analyses, logistic regression analyses suggest that the scale score value (784 points) corresponding to the upper portion of the Advanced ELP reading performance level and that corresponding to the lower portion of the Advanced ELP reading level (734 points) are sufficient to obtain the equal likelihood of attaining the state's proficient performance standard in literacy and mathematics, respectively. Again, this corroboration, strengthens confidence in pinpointing a range of ELP performance for the ELP performance standard.

15. Composite performance levels are derived from a weighted combining of two synthetic categories—comprehension (composed of listening and reading domains) and production (composed of speaking and writing domains). Because these composite performance levels do not align with the composite ELP level scale score ranges, logistic regression analyses could not be conducted using the ELP assessment composite scale score.

Exhibit 5.
Education Agency 2, Grade 7: Logistic Regression Plots for Literacy and Mathematics (2007–08)





ELP Reading performance levels for grade 7 (in scale score points)

Level 1 = Prefunctional (Below 469)
Level 2 = Beginning (469–618)
Level 3 = Intermediate (619–690)
Level 4 = Advanced (691–834)
Level 5 = Fully English Proficient (835 or above)

Exhibit reads: There is an equal probability for grade 7 EL students who score in the upper portion of the Advanced ELP performance level in reading to achieve the proficient performance standard in English or language arts, specifically, when obtaining 784 scale score points.

Notes: The plot represents the estimated logistic curve from a model in which the outcome is the dummy indicator English or language arts and mathematics proficiency and the predictor is the continuous ELP reading scale score. The logistic regression (probability) curves illustrate the likelihood of scoring at or above the academic proficient performance standard, as currently defined by the state for ELA and mathematics, respectively, as a function of increasing composite ELP scale scores.

The vertical dashed lines correspond to minimum ELP reading scale scores for Beginning, Intermediate, Advanced, and Fully English Proficient levels.

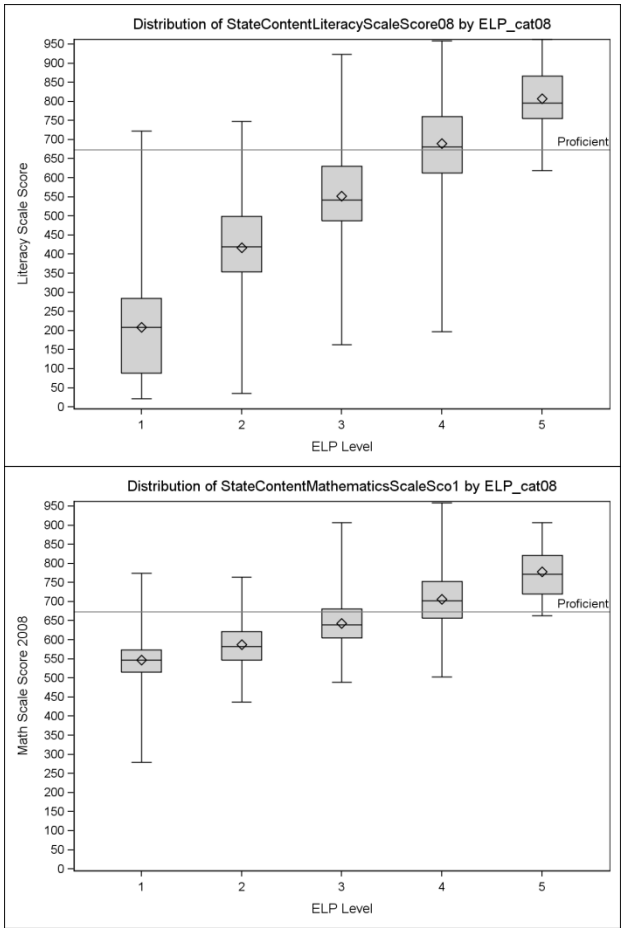
The point estimate and standard error are presented in Appendix Exhibit C.8.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Method C. Descriptive Box Plot Analysis

Finally, the box plot analyses for literacy and mathematics (Exhibit 6) for Education Agency 2's seventh-grade EL students reveal findings consistent with the decision consistency and logistic regression analyses. Specifically, as seen in Exhibit 6 (below), EL students performing at the Advanced composite ELP performance level (or level 4) have a median literacy performance of 681 scale score points (a little higher than the proficient standard of 673), and slightly more than 50 percent of ELs at this ELP level attain this academic-performance standard. For mathematics (Exhibit 6), EL students performing at the Advanced composite ELP level attain a higher median performance (703 scale score points) and attain academic performance standard at higher rates (nearly 75 percent of the EL students at the advanced ELP level do so).

Exhibit 6.
Education Agency 2, Grade 7: Box Plots of Literacy and Mathematics Scale Scores, by ELP Performance Level (2007–08)



ELP composite performance levels

Level 1 = Prefunctional
Level 2 = Beginning
Level 3 = Intermediate
Level 4 = Advanced
Level 5 = Fully English Proficient

Exhibit reads: Grade 7 EL students performing at the Advanced ELP level (or level 4) have a median performance of 681 scale score points in English or language arts, 8 points above the proficient performance standard.

Note: ELP composite scale scores are not aligned to the composite proficiency levels. While the former is created by averaging results from the four domains, the latter is generated using a weighting system that combines proficiency-level results from two synthetic categories—comprehension (composed of listening and reading domains) and production (composed of speaking and writing domains). Appendix Exhibit C.10 presents the descriptive statistics associated with this graph.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

In Education Agency 2, the percentage of consistent decisions tends to be higher for mathematics than for literacy at a given ELP level in seventh grade, but the percentage of consistent decisions for ELs in this grade is maximized at the same ELP reading level (Advanced) on the ELP assessment for both academic subjects. The logistic regression curves indicate that the 0.50 probability level for attaining academic proficiency in mathematics versus ELA is attained at a lower ELP assessment reading scale score (734 scale score points, versus 784 scale score points, respectively), but these are both within the Advanced ELP assessment reading level. The box plots for the same students also show that, while slightly more than 50 percent of those at the ELP assessment's Advanced reading level score proficient on the state literacy test, close to 75 percent do so on the state mathematics exam. The evidence across these analytic methods triangulates, and suggests that policymakers should consider this range of performance on the ELP assessment as a starting point for discussion.

Example 3. Education Agency 3

Education Agency 3's academic assessment tests students in reading and mathematics in grades 3 through 8 and in grade 10. These assessments provide academic-proficiency performance-standard categories of Minimal, Basic, Proficient, and Advanced. These assessments are administered in mid-fall (November), to assess knowledge gained in the prior academic year.

The ELP assessment used in Education Agency 3 is administered in the early fall and winter of each school year and provides six language-proficiency performance levels: Entering, Beginning, Developing, Expanding, Bridging, and Reaching. In addition, the ELP assessment provides composite proficiency scores in decimal values from 0.0 to 0.9 for each proficiency category (e.g., 3.0, 3.1, 3.2, 3.3). The decimals represent 10 equidistant scale score points between each proficiency level and the next.

Method A. Decision Consistency Analysis

To create bands for the decision consistency analysis, ELP assessment composite scores lower than 0.5 (e.g., 4.4) are categorized in the Low group. Thus the composite proficiency score of 4.4 would be categorized as Expanding Low. Scores at or above 4.5 would be Expanding High. Other proficiency bands are created similarly.

Exhibit 7.
Education Agency 3, Grade 10: ELP and Reading/Mathematics Decision Consistency Analysis (2009–10)

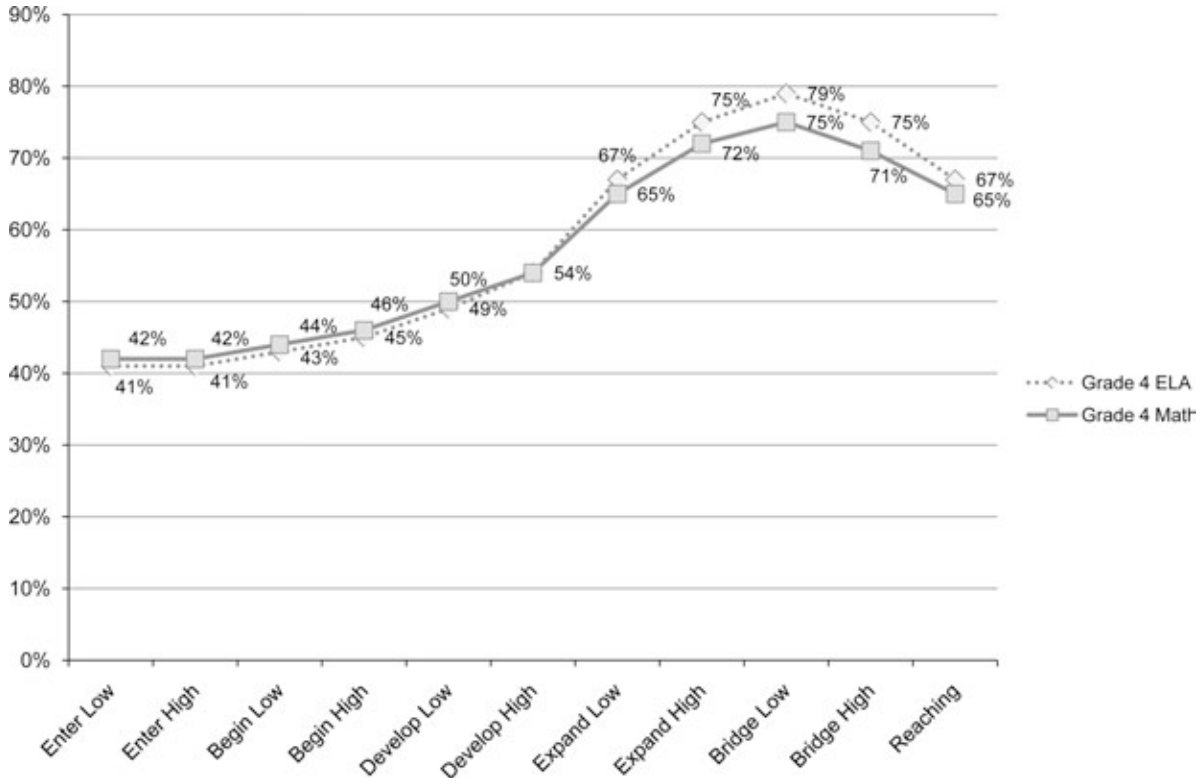


Exhibit reads: The percent of consistent decisions obtained for 10th-grade EL students in mathematics up through the Bridging Low ELP performance level is 75 percent. In English or language arts, the cumulative rate up through the Bridging Low level is 79 percent.

Note: This state has six ELP levels: Entering, Beginning, Developing, Expanding, Bridging, and Reaching. The composite ELP scale score ranges for the Entering (level 1), Beginning (level 2), and Developing (level 3) proficiency levels of the overall scores for grade 10 are 100–332, 333–362, and 363–386, respectively. The ranges for the Expanding (level 4), Bridging (level 5), and Reaching (Level 6) proficiency levels are 387–404, 405–423, and 424–600 in that order. The scale score point at the midpoint of each proficiency level range is then demarcated. Students below the midpoint are classified as "low." Students at or above that point, are classified as "high." Appendix Exhibits D.5 and D.6 present the number of students with ELP and reading and math proficiency-level scores and the percentage of consistent decisions.

Decision consistency formula: DC% = (QII + QIII)/(QI + QII + QIII + QIV).

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

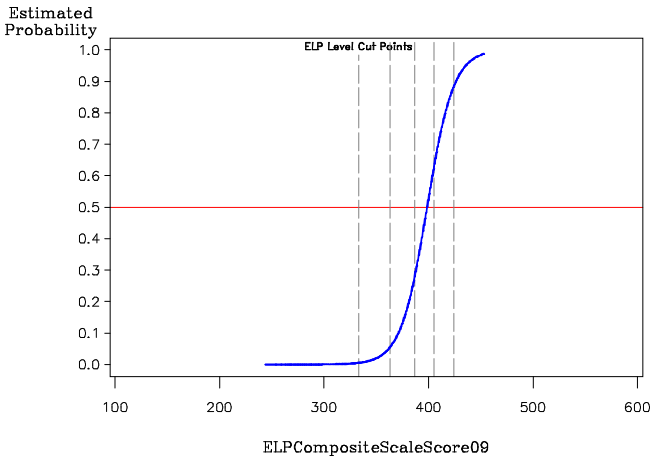
Decision consistency analyses for Education Agency 3 EL students in 2010 (Exhibit 7) show that the greatest percentage of consistent decisions for 10th-grade ELs in reading (79 percent), as well as mathematics (75 percent), is obtained at the same ELP score range—Bridging Low on the ELP assessment. This pattern holds for both years examined. Also, the percentage of consistent decisions in Education Agency 3 in this grade is similar in the two academic subjects.

Method B. Logistic Regression Analysis

The logistic regression (probability) curves for Education Agency 3's 10th-grade EL students, displayed in Exhibit 8 (below), yield results that *differ* from those of the decision consistency analyses for the same population. Specifically, the probability curves for both reading and mathematics yield a scale score value corresponding to the upper portion of the Expanding (i.e., Expanding High) composite ELP assessment performance level (398 and 399 scale score points, respectively), which is sufficient to obtain a 50-50 likelihood of proficient performance on the state's reading and mathematics content exams. *Expanding is an ELP level lower than that predicted in the above decision consistency analyses.* This kind of discrepancy requires careful interpretation and possibly further data analysis.

Exhibit 8.
Education Agency 3, Grade 10: Logistic Regression Plots for Reading and Mathematics (2009–10)

Grade 10, 2010 Reading Proficient Logistic Plot



ELP composite performance levels for grade 10 (in scale score points)

Level 1 = Entering (100–332)
Level 2 = Beginning (333–362)
Level 3 = Developing (363–386)
Level 4 = Expanding (387–404)
Level 5 = Bridging (405–423)
Level 6 = Reaching (424–600)

Grade 10, 2010 Mathematics Proficient Logistic Plot

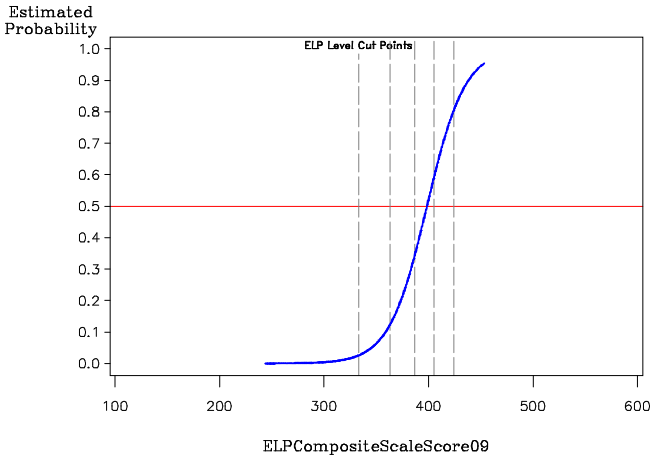


Exhibit reads: There is an equal probability for grade 10 EL students who score in the upper portion of the Expanding composite ELP performance level to achieve the proficient performance standard in English or language arts, specifically, when obtaining 398 scale score points.

Note: The plot represents the estimated logistic curve from a model in which the outcome is the dummy indicator English or language arts and mathematics proficiency and the predictor is the continuous ELP composite scale score. The logistic regression (probability) curves illustrate the likelihood of scoring at or above the academic proficient performance standard, as currently defined by the state for ELA and mathematics, respectively, as a function of increasing composite ELP scale scores.

The vertical dashed lines correspond to the cut scores between two continuous ELP performance levels such that a value of 333 is the point between Entering (Level 1) and Beginning (Level 2).

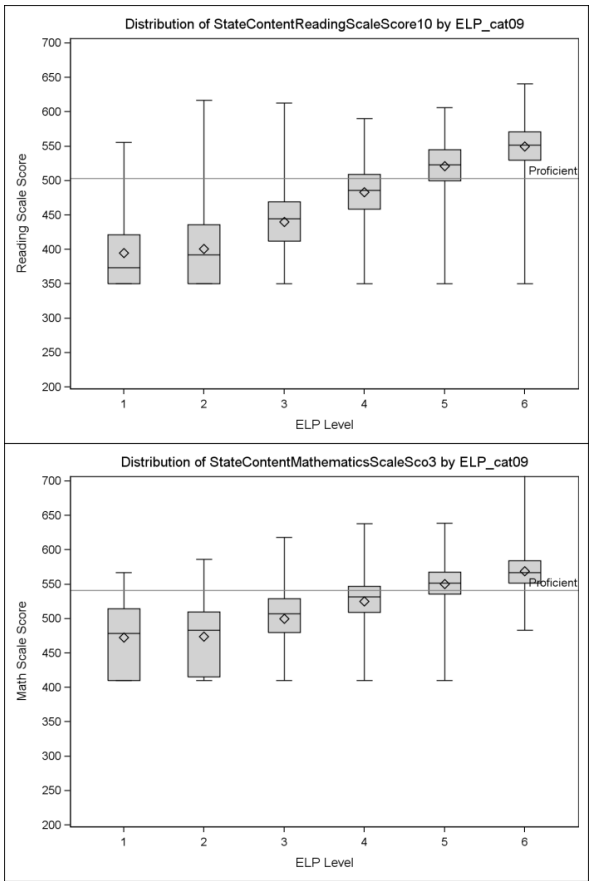
The point estimate and standard error are presented in Appendix Exhibit D.8.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Method C. Descriptive Box Plot Analysis

Finally, the box plot analyses for reading and mathematics (Exhibit 9) among Education Agency 3's 10th-grade EL students reveal findings consistent with the decision consistency method, although not with that of the logistic regression method. Specifically, as seen in Exhibit 9 (below), EL students performing at the Bridging composite ELP performance level (or level 5) on the ELP assessment have a median performance of 523 scale score points on the academic content exam (well above the proficient standard of 503 in reading), and about 75 percent of ELs at this Bridging performance level meet this academic performance standard. For mathematics (Exhibit 9), EL students performing at the Bridging composite ELP level on the ELP assessment attain a higher median mathematics performance (552 scale score points) but meet the mathematics performance standard at slightly lower rates (nearly 70 percent of the EL students at the Bridging ELP level do so).

Exhibit 9.
Education Agency 3, Grade 10: Box Plots of Reading and Mathematics Scale Scores, by ELP Performance Level (2009–10)



ELP composite performance levels for grade 10 (in scale score points)

Level 1 = Entering (100–332)
Level 2 = Beginning (333–362)
Level 3 = Developing (363–386)
Level 4 = Expanding (387–404)
Level 5 = Bridging (405–423)
Level 6 = Reaching (424–600)

Exhibit reads: Grade 10 EL students performing at the Bridging composite ELP performance level (or level 5) have a median performance of 523 scale score points in English or language arts, 20 points higher than the proficient performance standard.

Note: Appendix Exhibit D.10 presents the descriptive statistics associated with this graph.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Education Agency 3's three analyses, therefore, yield a more complex picture. Specifically, the percentage of consistent decisions is maximized for 10th-grade EL students at the Bridging Low ELP level on the ELP assessment for both reading and mathematics academic subject areas. In contrast, the logistic regression curves indicate that the 0.50 probability level of attaining academic proficiency in the state's 10th-grade reading and mathematics tests is reached at the upper end of the scale score range defining the Expanding ELP level on the ELP assessment; yet less than 50 percent of EL students at the expanding ELP level attain academic proficiency in reading or mathematics, as seen in the box plots—in fact, only about 27 percent of students at the Expanding level attain academic proficiency in either reading or mathematics.

In this example, how might analysts informing policymakers proceed? In this case, more careful analysis and interpretation of the data addresses the issue. The apparent discrepancy in the outcomes of the logistic regression method relative to the other two methods is resolved through more careful review of the empirical evidence. Specifically, the steep slopes seen in the logistic regression curves within the scale score range of the Expanding ELP level on the ELP assessment suggest that performance on Education Agency 3's 10th-grade academic content assessments is particularly sensitive to linguistic gains in this range of the ELP test. Indeed, the logistic regression curves estimate that an EL student at the high end of Expanding is much more likely to attain academic proficiency than a student at the low end of Expanding. For example, for Education Agency 3's 10th-grade reading academic assessment in 2010, the probability runs from 25 percent at the lower end (385 scale score points) to 62 percent at the upper end (405 scale score points) of the Expanding ELP level on the ELP assessment; for mathematics, the probability of academic proficiency runs from 32 percent to 59 percent for the same scale score point range of Expanding on the ELP assessment. Because the box plot analyses indicate that nearly 75 percent of EL students at the Bridging ELP level on the ELP assessment meet the proficient performance standards for reading and mathematics content assessments, the preponderance of evidence suggests that policymakers should consider setting the ELP performance standard somewhere between the Expanding High and Bridging Low ELP levels on the state's ELP assessment. Next steps would involve analysis of additional grades, to increase confidence in recommending a particular performance standard.

Summary

The three examples highlight how these distinct analytic methods can assist state policymakers in examining empirical evidence on EL student academic performance relative to ELP level, and provide input for policy decisions on determining an acceptable ELP performance standard. Even when utilizing these methods in very different "worked examples," on EL student outcomes from each of three states over two academic years at three different grade levels, these methods yield largely convergent results within each Education Agency, although results for each require different degrees of interpretation.

In effect, these empirical methods are tools to gather empirical evidence to inform and support policy discussions. They are not intended to mechanically determine or constrain policy decisions. Moreover, for simplicity of presentation, analyses were performed on grade-level data. In a complete application, states would review the empirical outcomes of all grades for which data are available. In addition, states could choose to aggregate these data into grade spans (e.g., 3–5, 6–8, 9–12) in order to examine performance patterns by elementary, middle, and high school segments, respectively. As noted previously, policymakers are usually constrained to choose a single English-language-proficient performance standard. The evidence presented above suggests that these three analytic methods used in conjunction can greatly assist policymakers in conducting informed discussions and making defensible decisions. While the three methods presented here provide useful information, they are intended to

stimulate reflection and do not represent all analytic possibilities. In fact, this work should motivate more exploratory research in this area to generate new, more powerful analytic and visual approaches.

Caveats

One key caveat that policymakers should keep in mind as they review state analyses concerns the timing of administration of the ELP and academic content assessments. Specifically, in instances such as Education Agencies 1 and 3, where the ELP assessment is administered several months before the academic achievement assessment, EL students' *actual* level of English language proficiency *when they take the academic content assessment* may be quite different from that indicated by their ELP assessment result. Furthermore, the direction of this difference may vary, in part, according to the exact time of year when each assessment is given, and the linguistic environment of the EL student population.

In Education Agency 1, because the state ELP assessment is given July through October and the academic assessment is given the following spring, one could argue that the academic achievement results for a given ELP level will be systematically overstated relative to performance on concurrent assessments, since EL students will have received almost an entire year of instruction in English as a second language or English language development and will likely be at a higher ELP level when they take the academic content assessment. In Education Agency 3, where the ELP assessment is given in late spring and the academic achievement assessment is administered in the *following* fall, one might imagine that the same phenomenon as that just described could result. However, if some EL students spend the summer months in isolated linguistic communities, with little or no exposure to native English speakers and English texts, it is possible that this overestimation effect is canceled out, or even that an underestimation effect is generated. Given the complexities of interpreting these effects for different populations and contexts, findings from these analytic approaches will be more robust and interpretable to the extent that the ELP and academic content assessments are administered in closer time proximity.

Another, related issue concerns the possible systematic exclusion of data of students tested on the ELP and academic assessments at different times. For example, since Education Agency 1 administers the state ELP assessment in the fall, some of the higher performing EL students in Education Agency 1 had their language classifications changed from EL to reclassified fully English proficient (RFEP) between the administration of the state ELP assessment in the fall and the state's academic content tests in the following spring. Because results for all annual state ELP assessment examinees were available, these students (listed as RFEP in the state academic content test data file for the same academic year) could be identified and included in the analysis. Data analysts assisting states to utilize these analytic techniques should investigate these kinds of issues and address them as possible.

This last caveat raises a more general issue for analysts and decision-makers to keep in mind. As an empirical fact, states have a given EL population at a given point in time according to the current entry and exit criteria they employ. Any empirical exploration to define (or redefine) the English-language proficient level will be affected by the current EL classification or reclassification criteria that define the population of a given state.

For example, states currently using an ELP performance standard below the highest possible performance level will lose ELP assessment data on students above that cut point. Moreover, in states with multiple reclassification criteria that include academic achievement measures, ELs at higher ELP levels (including the state's currently defined English-language-proficient level) who also perform higher academically are more likely to exit EL status and therefore no longer be assessed on the state's language proficiency assessment. This means that even students who remain EL at higher ELP levels in these states are likely to perform less well academically *by definition*. Such censoring can result in a systematic

underestimation of academic performance of ELs because it reports (at the higher grades especially) only the results of those who remain EL—often long-term ELs—as well as the results of newcomers. This skimming bias is now more widely recognized, and it particularly affects accountability decisions on overall EL subgroup performance (see Working Group on ELL Policy 2010). The skimming bias also likely yields empirical findings suggesting higher ELP assessment cut points for the English-language-proficient level at high school grade levels. However, decision-makers generally reject such an extreme for the state performance criterion, as results from elementary and middle school grades tend to be more consistent and signal a lower ELP performance standard.

III. Establishing a Time Range for English Learners to Attain an English-Language-Proficient Performance Standard

Overview

In defining accountability provisions regarding English Learners (ELs) attaining English proficiency (i.e., annual measureable achievement objectives, or AMAOs), federal law specifically mentions time:

> Such annual measurable achievement objectives shall be developed in a manner that reflects the *amount of time* [italics added] an individual child has been enrolled in a language instruction educational program. §3122(a)(2)(A).

The statute effectively requires states to pay careful attention to the amount of time ELs are expected to be in language instruction educational programs. This in turn has implications for states in setting annual progress expectations (AMAO 1) and establishing time frames to attain English-language proficiency (AMAO 2). Because state accountability systems need to reflect such expectations, this chapter illustrates approaches for establishing a reasonable yet rigorous time frame for ELs to attain the proficient performance standard on state ELP assessments.

Given its importance, empirical research on this topic has been surprisingly limited, but is nevertheless instructive. For example, Hakuta and others (2000, 13) examined this question and concluded that

> even in districts that are considered the most successful in teaching English to EL students, … [attaining] academic English proficiency can take 4 to 7 years.

Others have derived from empirical research similar time estimates (e.g., Genesee et al. 2006; Linquanti and George, 2007; Cook et al. 2008; Taylor et al. forthcoming). What emerges from these studies is that time frames to reach ELP vary based on several factors (e.g., initial English-proficiency level, particular language domain(s) assessed; age or grade on entry; primary language literacy level; type of language-proficiency assessment; background of EL students within a school, district or state; instructional program goals) with estimated time frames ranging from three to seven years.

Given the variety of factors that influence time to attain English proficiency, states should conduct empirical analyses of their own existing data to derive a time-to-English-proficiency expectation for their EL populations. In doing so, the states need to consider several issues:

1. **Longitudinal data is often limited.** Many states have a limited number of years of standardized longitudinal data for ELs (e.g., three to five years). In these cases, many students in a state's data set have not yet attained ELP. In calculating time to an English-proficient performance standard, how does a state account for students who have not yet attained the standard? For example, if there are 1,000 EL students in the first year of a record system and 250 of those have not attained English proficiency after five years of language instruction, how does a state calculate time to English proficiency? Using only the 750 students who have attained the performance standard to calculate time to ELP will generate an *underestimate* of how long it actually takes, as the analysis effectively excludes the 250 students who have not yet attained the

goal. The not-yet-English-proficient students need to be accounted for in any method calculating time to ELP.

2. **State policies defining ELP evolve over time.** Proficiency standards change; ELP assessments change or are restandardized; and criteria defining ELP performance levels are adjusted. The expected English-proficient standard can therefore change for different cohorts of EL students, and the resulting "move in the goal post" can make year-to-year comparisons problematic.
3. **EL language proficiency growth rates vary significantly.** Empirical studies of EL student growth on ELP assessments (Cook et al. 2008, Linquanti et al. forthcoming) illustrate a general pattern: Students starting at lower proficiency levels will likely take longer to attain the English-language-proficient performance standard than will students starting at higher proficiency levels. Also, students of equivalent levels of language proficiency at higher grade levels are likely to take longer to attain the standard than their counterparts in lower grade levels. Given this general pattern, time frames to attain ELP may be sensitive to ELs' initial English-proficiency levels and their grade span.[16]
4. **Data are often missing for EL students.** EL student records often have missing data, even when these should be available. For example, EL students may have three years of ELP assessment information, but the subsequent two years of records lack information. Some of these students may have left the district, state or country, while others continue to be enrolled but do not have current records. Should the available information be discarded or incorporated? If the latter, what methods should be used to incorporate the data in ways that do not distort the time estimates?

Certainly there are other issues that arise when attempting to determine time to ELP. The point is to identify the most salient ones and offer methods that adequately address these issues in answering the time-to-proficiency question.

Key Approaches

We present two approaches for establishing a target time frame for ELs to attain a preidentified ELP performance standard:

1. **Descriptive analysis**, which follows over time EL students who start at a prespecified date at varying English-proficiency levels. The proportions of EL students who annually attain the ELP criterion are then shown in a bar chart. The goal of this approach is to get a sense of percentages attaining language proficiency, by time, initial ELP level and grade span.
2. **Event history analysis**, which is also known as survival analysis,[17] is used extensively in the fields of engineering and medicine to estimate the time required for an event of interest to occur (Klein and Moeschberger 1997).[18] For analyses here, the event is an EL student's attaining the given ELP performance standard. The goal of this approach is to calculate a time frame that incorporates students for whom the event of interest does not occur. The following section

16. However, current federal regulations interpreting Title III permit states to set target performance percentages for cohorts of ELs only on the basis of their time in U.S. schools.

17. The term "survival function" is used here; yet the values presented in the following analyses are linear transformations of the survival function (called the failure rate) and are calculated as 1 minus the survival function. The event of interest represents a positive outcome for ELs, while "surviving" means not experiencing that event.

18. In engineering, that event is often the failure of a mechanical part or component. In medicine, it is often the death of a patient. Covariates can be used with this procedure to address such questions as, "If this metal is used, how much longer will it be before this component fails?" or "If patients receive this treatment, will their lives be extended? If so, by how much?"

applies these analytic approaches to a robust EL data set, in order to generate estimates that suggest a range of options and then illustrate how these can be applied by decision-makers.

As in the previous chapter, these methods are recommended for use in a deliberative process by expert stakeholders empaneled to make recommendations by considering empirical data from the population of interest, in addition to their experiences. The methods effectively offer two information points for consideration. The following section applies these analytic methods to examine data for EL students within one education agency referred as Education Agency 1.

Example: Education Agency 1

Student data from Education Agency 1 (EA 1) are used to model both the descriptive and the event history analysis approaches. This education agency was chosen because relevant data were available for five years. ELs from kindergarten to fifth grade are included in these analyses. The sample was restricted to these grades for convenience and for illustration purposes. Because samples are reduced when ELs are divided by grade and ELP levels, results are presented by grade clusters, combining the data from grades K–2 and grades 3–5. Doing so provides some sense of how time estimates can vary by grade span.

From these clusters, only students first designated as ELs between July and October 2003 (the ELP testing window for this education agency) were selected. Although there are small numbers of students entering and identified as ELs later in the school year (e.g., spring 2004), confining the sample to the July to October testing window permits a clearer interpretation of how many years it takes to attain the English-proficient level according to the given ELP assessment's performance standard.

The 2003–04 school year was used as the starting point because this is the first year for which data are available for EA 1. EA 1 has EL student records from the 2003–04 to 2007–08 (i.e., five successive years). The state's ELP assessment defines the English-proficient performance standard, using a conjunctive approach: Students must attain an overall composite score of 4 and have no domain score (i.e., reading, writing, speaking or listening, each weighted 25 percent) lower than 3. For both analytic approaches shown, EL students meeting this performance standard on the ELP assessment are considered to have experienced the event of interest and are categorized as English proficient.[19]

Method A. Descriptive Analysis Approach

In this approach, ELs who start English-language instruction educational programs at a specific date are identified. These students are then followed over time. The cumulative percentage of students reaching the English-proficiency criterion is shown for each successive year. In the last year of available data, the proportion of EL students in the cohort not reaching the English-proficient criterion is also shown. Proportions of students reaching the proficiency criterion each year are identified by initial ELP level and grade span, to highlight the differences in ELP attainment rates of groups based on these key variables.

19. As this agency uses multiple criteria to exit students from EL status and students must take the state ELP assessment annually, per Title III requirements, until they leave EL status, there are students who meet the ELP performance standard but continue to take the ELP assessment annually, as they have not met the other criteria (which include academic criteria from standardized tests and classroom performance) needed to exit. Moreover, some of these students may not meet the ELP performance standard in a subsequent test administration. In the present analyses, once an EL student attains the ELP assessment criterion, he or she is considered to have experienced the event of interest, and any subsequent ELP assessment result, even if available, is not used.

The following table (Exhibit 10) shows the number and percentage of students who meet the assessment criterion by number of years in program.

Exhibit 10.
(Method A)
Number and Percent of Students Identified EL in Kindergarten to Second Grade From 2003–04 Attaining the English-Proficient Performance Standard, by Initial ELP Level and Time in Program, Education Agency 1

Elapsed Time (in Years)	Total Number of Remaining Students With Complete Data	Cumulative Number of Students Becoming English Proficient	Cumulative Percent of Students Becoming English Proficient
ELP Level 1			
1	6,506	809	12%
2	5,697	1,593	24%
3	4,913	2,023	31%
4	4,483	2,860	44%
ELP Level 2			
1	6,671	2,184	33%
2	4,487	3,254	49%
3	3,417	3,636	55%
4	3,035	4,393	66%
ELP Level 3			
1	9,328	5,271	57%
2	4,057	6,889	74%
3	2,439	7,353	79%
4	1,975	8,021	86%

Exhibit reads: Among grade K to 2 students identified as EL during 2003–04, who were designated as such between July and October 2003 (i.e., at the beginning of the school year) and who started at ELP level 1, 12 percent became proficient in one year.

Note: Sample sizes are adjusted to prevent distorting the analysis with missing cases. For example, an initial sample size of 7,728 students was identified as ELs at ELP level 1 during 2003–04 and who were designated as such between July and October 2003 (i.e., at the beginning of the school year). Among them, those missing assessment data in years 1 through 3 (before year 4) are excluded from the analysis (n = 1,222), yielding a total sample of 6,506. Students without data in year 4—the last year examined—are not excluded, as they are considered "censored," that is, not having attained the English-proficient performance standard by year 4.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

The last column in Exhibit 10 provides a cumulative percent of EL students becoming English proficient in a given time frame. For example, for those in grades K–2 whose initial ELP level is 1, the percentage of students becoming proficient after one year is obtained by dividing the number of students becoming proficient in year 1 by the total number of students with complete data in the initial year (i.e., 809/6,506 = 0.12). Thus, the cumulative percent of students becoming proficient after three years (i.e., at year 3) for those with initial proficiency level of 1 is 31 percent (i.e., 2,023/6,506).

Using data from Exhibit 10, Exhibit 11 shows that the proportion of ELs at each initial English-proficiency level who attain English proficiency increases over four years. Note also that the lower the initial proficiency level, the lower the percentage of students becoming proficient over the same time period. For example, of EL students whose initial proficiency level during 2003–04 was level 1, 44 percent became proficient in four years, whereas 86 percent of EL students whose initial proficiency level was level 3 became proficient in four years.

Exhibit 11.
(Method A)
Cumulative Percentage of Students Attaining English Proficiency, by Year, Kindergarten to Second Grade (Without Missing Records), Education Agency 1

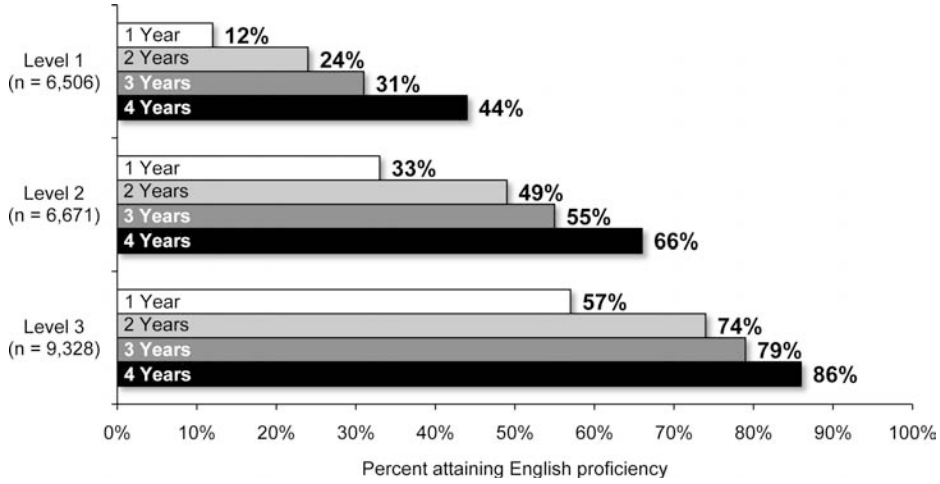


Exhibit reads: Among grade K to 2 students identified as ELs during 2003–04, who were designated as such between July and October 2003 (i.e., at the beginning of the school year) and who started at ELP level 1, 12 percent became English proficient in one year.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibits 12 and 13 present the same analyses for the grade cohort 3 through 5.

Exhibit 12.
(Method A)
Number and Percent of Students Identified EL in Third to Fifth Grade From 2003–04 Attaining the English-Proficient Performance Standard, by Initial ELP Level and Time in Program, Education Agency 1

Elapsed Time (in Years)	Total Number of Students With Complete Data	Number of Students Becoming Proficient	Cumulative Percent of Students Becoming Proficient
ELP Level 1			
1	714	60	8%
2	654	151	21%
3	563	263	37%
4	451	359	50%
ELP Level 2			
1	168	75	45%
2	93	93	55%
3	75	119	71%
4	49	133	79%
ELP Level 3			
1	281	215	77%
2	66	238	85%
3	43	252	90%
4	29	263	94%

Exhibit reads: Among grade 3 to 5 students identified as EL during 2003–04, who were designated as such between July and October 2003 (i.e., at the beginning of the school year) and who started at ELP level 1, 8 percent became proficient in one year.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit 13.
(Method A)
Cumulative Percentage of Student Proficiency, by Year, Third to Fifth Grade (Without Missing Records), Education Agency 1

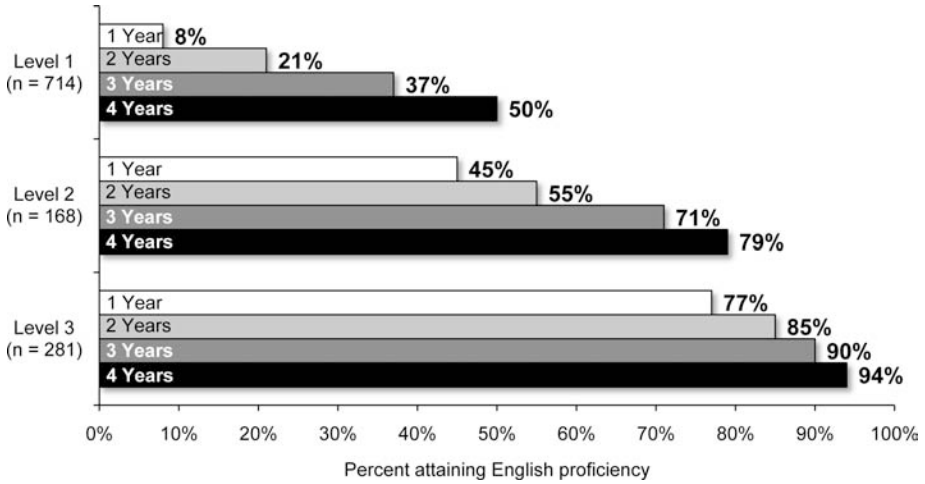


Exhibit reads: Among grade 3 to 5 students identified as EL during 2003–04, who were designated as such between July and October 2003 (i.e., at the beginning of the school year) and who started at ELP level 1, 8 percent became proficient in one year.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

A similar pattern was observed in the third- to fifth-grade cluster to that seen in the grades K to 2 cluster. The lower the initial proficiency level, the longer it took to become proficient on the ELP assessment. For those whose initial proficiency level was 1 in 2003–04, about 50 percent became proficient in four years; whereas 94 percent of EL students whose initial proficiency level was 3 in 2003–04 became proficient in four years. According to the descriptive results, the percentages of students becoming English proficient, conditioned on initial proficiency level, are higher in the grade 3 to 5 cluster than in the grade K to 2 cluster in all but two instances (i.e., those at initial ELP level 1 after one or two years).

Method B. Event History Analysis Approach

Event history analysis is an approach for estimating the probability that a particular event of interest will occur in a given time frame. The event of interest here is an EL student's attaining the ELP criterion. Several event history analysis methods are available. The one adopted here is the product-limit estimator (Klein and Moeschberger 1997, 84), commonly known as the Kaplan–Meier estimator. This event history analysis creates a variety of estimates. Of particular interest here is the survival function shown in Appendix E.

The survival function provides the probability that students will become English proficient at a particular time. This statistic is well suited to all students who experience the event of becoming proficient within a given time frame (for the current analysis, in four years). However, a survival function estimate cannot be calculated for students who do not attain the English-proficient criterion. This is a problem, since substantial numbers of students do not meet the criterion after five years. (In event history analysis, students who do not experience the event within the observed time frame are said to be "censored." Similarly, students who have experienced the event of interest are said to be "noncensored.") The use of

only noncensored ELs (i.e., those who meet the criterion) for the event history analysis will very likely generate an underestimate of the time it takes to become English proficient (i.e., to experience the event). To overcome this shortfall, two corrective procedures have been adopted. The first procedure (*Censored Adjustment 1*) yields an underestimate of the time it will take, while the second (*Censored Adjustment 2*) yields an overestimate of that time. The use of both adjustments provides comparative information on time frames that ELs are likely to need to attain the English-proficient criterion. For Censored Adjustment 1, all students who are censored are assumed to attain the English-proficient criterion *in the following year*. For example, all students censored at year 4 (4th year in the program) are assumed to attain the criterion at year 5 (the 5th year). Adjustment 1 is an underestimate because it is unlikely that all students in EL programs in EA 1 will attain English proficiency in year 5 and instead some students will take a longer time until they become proficient. Thus, event history analysis estimates will be shorter than what would be observed if all students could be followed until they actually reached the English-proficient criterion. Censored Adjustment 2 draws on Hakuta and colleagues (2000) and Cook and colleagues (2008). Hakuta and colleagues state that students can attain academic ELP somewhere between four and seven years. For this analysis, seven years is used as the maximum time for students to attain the English-proficiency criterion. Cook and colleagues demonstrate that students at different English-proficiency levels grow at different rates. For the Censored Adjustment 2, the empirical observations made by Hakuta and colleagues and Cook and colleagues are combined. That is, students who started at the lowest proficiency level (level 1) in the 2003–04 school year are assumed to take seven years to attain the English-proficient criterion. Students starting at level 2 will take six years, and students starting at level 3 will take five years. Adjustment 2 is assumed to be an overestimate, as "maximum time frames" from prior empirical research are utilized.[20]

Data from the noncensored and censored students (with imputed times) are then analyzed, using event history analysis. However, the number of students becoming proficient over time and the number of censored students are exactly the same for the two methods described above. The only difference is the year imputed for the censored cases.[21]

The exhibits below (14 and 15) graphically display results from the event history analysis for both methods. The horizontal axis of each figure displays years in language instruction educational programs. The vertical axis shows the probability of an EL's becoming proficient. (Results tables can be seen in Appendix E.)

Two grade clusters are displayed: kindergarten to second grade and third to fifth grade. The event history graphs are typically displayed as step functions. The horizontal axes in the graphs displaying Censored Adjustment 1 (Exhibit 14) and Censored Adjustment 2 (Exhibit 15) show seven years. Note that Censored Adjustment 1 has a maximum time of five years. The seven-year time line is provided in displayed graphs for comparative purposes.

20. For the Censored Adjustment 2, because there are no additional data after year 4, the imputed values do not provide additional information to calculate further probabilities. Thus, from years 4 to 7, the probabilities are effectively unchanged.

21. By design, the two methods provide the same results for the first two years, as n_i in the survival function is the same across the two methods because no censored cases are assumed in year 0, and the total number of students in year 1 is determined by subtracting the number of students becoming proficient and the number of students censored in year 0, which is zero, from the total number of students in year 0. (See Exhibit E.1 in comparison with Exhibit E.2.)

Exhibit 14.
(Method B)
Censored Adjustment 1
Probability of ELs Identified During 2003–04 Becoming Proficient, by Grade Cohort and ELP Level, Education Agency 1

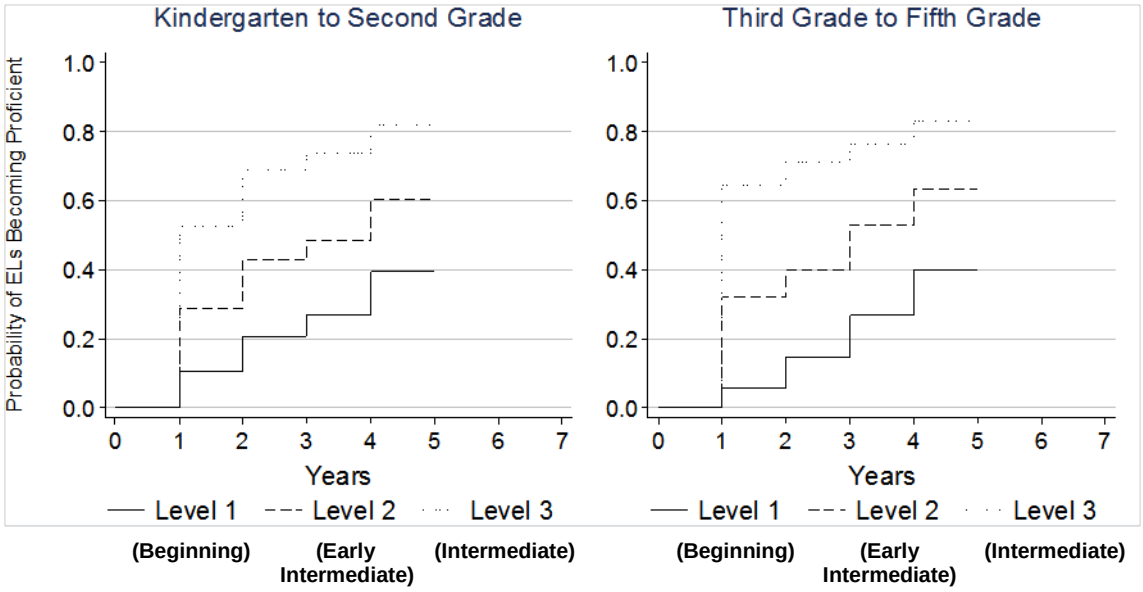


Exhibit reads: For students identified as ELs during 2003–04 who were designated as such between July and October 2003 (i.e., at the beginning of the school year) and who began at ELP level 1, there is a 10 percent probability of attaining English proficiency in one year.

Note: The state ELP assessment provides five ELP performance levels: Beginning (Level 1) , Early Intermediate (Level 2), Intermediate (Level 3), Early Advanced (Level 4), and Advanced (Level 5). Level 4 or higher represents this agency's English-Language-Proficient performance standard. The corresponding scale score range for each ELP performance level varies across grades. For grades K to 2, the number of students at the Beginning ELP level was 7,728, at the Early Intermediate ELP level was 7,603, and at the Intermediate ELP level was 10,045. The total number of students was 25,376. For grades 3 to 5, the number of students at the Beginning ELP level was 1,043, at the Early Intermediate ELP level was 235, and at the Intermediate ELP level was 335. The total number of students was 1,613.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

The observed event history curves from both methods (Exhibits 14 and 15) are similar. As expected, Censored Adjustment 1 shows students meeting the criterion in higher proportions than in Censored Adjustment 2. That is, Censored Adjustment 1 predicts a shorter time for students to reach the criterion than Censored Adjustment 2 does. But these differences are slight. With both methods, higher proportions of level 3 students meet the criterion in the third- to fifth-grade cluster than in the kindergarten to second-grade cluster.

Exhibit 15.
(Method B)
Censored Adjustment 2
Probability of ELs Identified During 2003–04 Becoming Proficient, by Grade Cohort and ELP Level, Education Agency 1

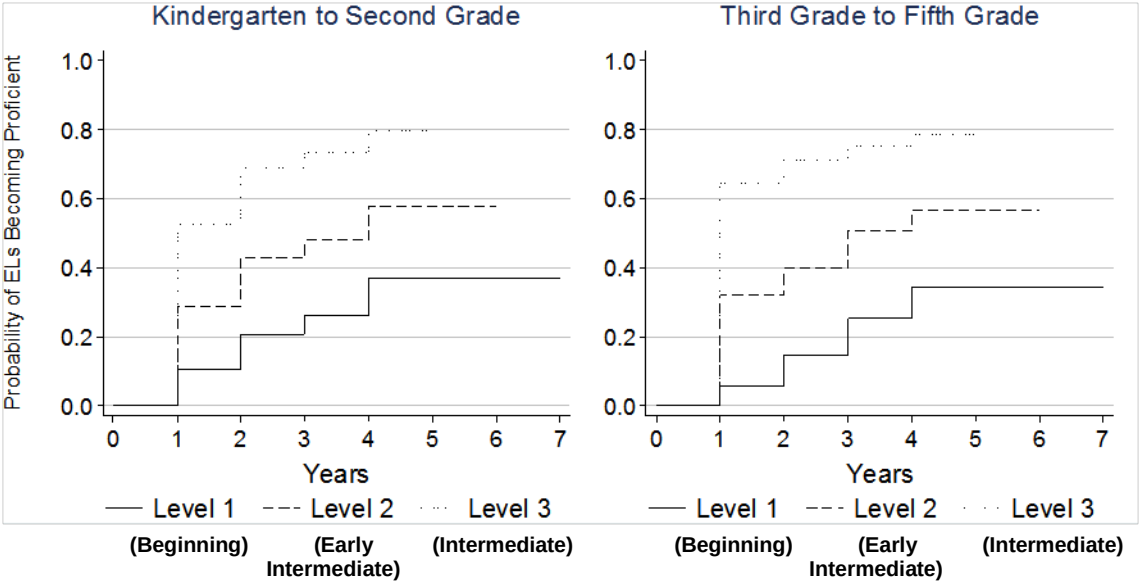


Exhibit reads: For students identified as ELs during 2003–04 who were designated as such between July and October 2003 (i.e., at the beginning of the school year) and who began at ELP level 1, there is a 10 percent probability of attaining English proficiency in one year.

Note: The state ELP assessment provides five ELP performance levels: Beginning (Level 1) , Early Intermediate (Level 2), Intermediate (Level 3), Early Advanced (Level 4), and Advanced (Level 5). Level 4 or higher represents this agency's English-Language-Proficient performance standard. The corresponding scale score range for each ELP performance level varies across grades. For grades K to 2, the number of students at the Beginning ELP level was 7,728, at the Early Intermediate ELP level was 7,603, and at the Intermediate ELP level was 10,045. The total number of students was 25,376. For grades 3 to 5, the number of students at the Beginning ELP level was 1,043, at the Early Intermediate ELP level was 235, and at the Intermediate ELP level was 335. The total number of students was 1,613.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Application of Methods for Decision Making

How might results of these two methods be used by stakeholders for decision making? Results from these approaches provide insight into the proportion of EL students who actually attain, as well as the probability that they will attain, the English-proficient criterion within certain time frames. These methods, however, are not intended to provide policymakers with a simple definitive answer to the "how long?" policy question. The analyses should support but not determine such a policy decision. Professional judgment by informed panelists and policymakers is required to weigh these outcomes and utilize them to set expected time frames that are rigorous and defensible. Also, it is important to point out that both analyses provide information about how long it is observed or estimated for EL students to attain English proficiency *based on current practices*. The two methods say nothing about how long it *should* take with improved instructional practice. In establishing expectations for accountability purposes, this expectation of improved instructional practice should also be taken into account to avoid unintentionally setting lower expectations.

The following discussion applies results from Education Agency 1 to illustrate how these methods could be used to set adjusted time frames for EL students by initial English-proficiency level.

For example, a time frame where 60 percent (or a probability of 0.6) of a state's ELs attain the English-proficient criterion (arguably a clear majority) might be considered as an initial place to begin deliberations among stakeholders. A starting point time frame in which a clear majority of EL students have met the criterion provides both an "existence proof" that the target time frame is attainable and also provides a justifiable "stretch" in target performance for both educators and students.

To illustrate how this time frame could be identified from the empirical data, Exhibit 16 shows proportions (actual or predicted) of students attaining the proficiency criterion by each approach. Cells having proportions above 60 percent or 0.60 are shaded.

Exhibit 16.
Combined Outcomes From Descriptive Approach and Event History Analyses, by Grade Cluster, 2003–04 Initial Proficiency Level and Time

Level/Time	Kindergarten to Second Grade			Third to Fifth Grade		
	Descriptive Approach (Proportion)	Censored Adjustment 1 (Probability)	Censored Adjustment 2 (Probability)	Descriptive Approach (Proportion)	Censored Adjustment 1 (Probability)	Censored Adjustment 2 (Probability)
ELP Level 1						
1	12%	0.10	0.10	8%	0.06	0.06
2	24%	0.21	0.21	21%	0.14	0.14
3	31%	0.27	0.26	37%	0.27	0.25
4	44%	0.39	0.37	50%	0.40	0.34
ELP Level 2						
1	33%	0.29	0.29	45%	0.32	0.32
2	49%	0.43	0.43	55%	0.40	0.40
3	55%	0.48	0.48	71%*	0.53	0.51
4	66%*	0.60*	0.58	79%*	0.63*	0.57
ELP Level 3						
1	57%	0.52	0.52	77%*	0.64*	0.64*
2	74%*	0.69*	0.69*	85%*	0.71*	0.71*
3	79%*	0.74*	0.73*	90%*	0.76*	0.75*
4	86%*	0.82*	0.80*	94%*	0.83*	0.79*

Exhibit reads: For example, for grade K to 2 students identified as ELs during 2003–04, who were designated as such between July and October 2003 (i.e., at the beginning of the school year) and who started at ELP level 1, the descriptive approach shows that 12 percent became English proficient in one year, and event history analyses in both Censored Adjustments 1 and 2 show that there is a 10 percent probability of attaining English proficiency in one year.

* Cells having proportions above 60 percent or .60 are shaded.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit 16 (above) presents four years of results. (Recall that there were five data points representing change over four instructional years). Estimated probabilities beyond year 4 in the event history analyses are identical, as these analyses adjust estimated proportions within the observed time frame on the basis of censored case assumptions.

For all levels and clusters, the descriptive approach yields higher proportions of students' attaining the ELP criterion than do the event history analyses' probabilities. Similarities between observed proportions and estimated probabilities might be expected. But the descriptive approach does not take censoring into account. As mentioned above, the event history analysis using Censored Adjustment 1 has at least the same probabilities as Censored Adjustment 2.

None of the approaches shows more than 50 percent, or 0.50, probability of the lowest level (ELP level 1) EL students' attaining the English-proficient criterion within the observed four-year time frame. For ELs entering Education Agency 1 at an initial ELP level 2, both the descriptive and Censored Adjustment 1 approaches show more than 60 percent, or 0.60, of students in the kindergarten to second-grade cluster meeting the criterion within four years. For the third- to fifth-grade cluster, the descriptive approach shows 60 percent of students meeting the criterion in three years, and the Censored Adjustment 1 shows 0.60 occurring in four years. For Censored Adjustment 2, neither grade cluster has a probability of more than 0.60 within the observed time frame. Students starting at the higher initial English-proficiency level (ELP level 3) reach the 60 percent or 0.60 threshold far sooner than students starting at lower proficiency levels.

On the basis of the above results, Education Agency 1 might consider the following approach for setting an initial expected time frame for reaching the ELP performance standard. The students at the lowest initial ELP level are not observed or predicted to attain the English-proficient performance standard at the 60 percent, or 0.60, threshold within the four-year time frame.[22] Thus, a process is needed to predict when these students may reach this threshold. To calculate how long it will take to reach this threshold, a simple regression procedure can be employed using the available data. (See Exhibit 17, below.)

22. Alternatively, states could rank districts on the basis of the proportion of EL students attaining the English-language-proficient standard each year. Following this, an nth percentile rank criterion could be established. A graph could be plotted and smoothed and used to identify the target attainment rate from that district-ranked performance.

Exhibit 17.
Percent of Initial ELP Level 1 ELs Attaining the English-Proficient Threshold Across Analytic Approaches and Grade Clusters Predicted Beyond Observed Years

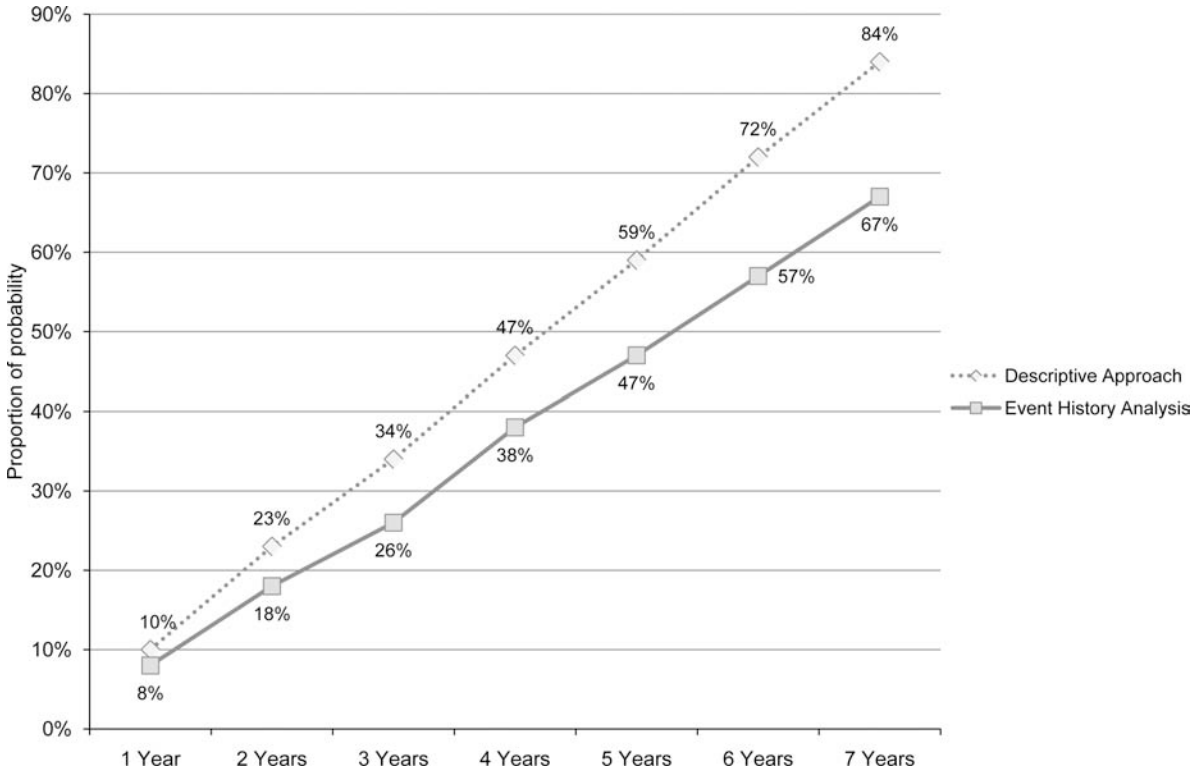


Exhibit reads: Using a descriptive approach, 10 percent of students across grade clusters at ELP level 1 were estimated to have attained the English-proficient standard in year 1. For the event history analysis approach the estimated value was 8 percent in year 1.

Note: Values up to year 4 were calculated by averaging across approaches and grade clusters in Exhibit 16. Values beyond year 4 were obtained by using the year 1 to 4 averages and calculating a slope, using a linear regression. Therefore, the values for years 5 through 7 are *predicted values.*

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

The graph and table displayed in Exhibit 17 show an example of a simple regression procedure. Values in Exhibit 17 are derived by averaging across approaches and grade clusters. For example, the percentage of students at Level 1 attaining the English-proficient standard in the grade K to 2 cluster, in the first year of English language instruction, is 12 percent (see Exhibit 16). The proportion attaining proficiency in the grade 3 to 5 cluster is 8 percent. The average proportion across clusters for level 1 ELs in year 1 is 10 percent, which is the value shown in the descriptive approach in Exhibit 17 for year 1. For the event history analysis, averages are calculated across Censored Adjustments 1 and 2 and across clusters. Values up to year 4 in the table are calculated similarly. Values beyond year 4 are obtained by using the year 1 to 4 averages and calculating a slope, using a linear regression. The slope then is added to subsequent years. The slope value of the descriptive approach is 12 percent, and the slope of the event history approach is 0.097. These values are then added to the year 4 average, which results in values of 59 percent, or 0.47, for year 5. The slope values are then added to the predicted year 5 results, and so on. On the basis of the descriptive approach, the lowest level students are predicted to reach the 60 percent, or 0.60, threshold in 5 to 6 years, and on the basis of the event history approach, 6 to 7 years.

For initial ELP level 2 students, two of the three approaches show that the 60 percent, or 0.60, threshold is met by or before the fourth year. Thus, a four-year time line could be adopted for level 2 students.

Level 3 students have noticeably different time frames by grade cluster. A one-year time line could be adopted, but that may not be reasonable for ELs in the kindergarten to second-grade cluster, especially because literacy is strongly developing and measured substantially differently at these grades. Were this agency to set a single criterion for both clusters, a time frame of two years might be more reasonable.

Clearly, students' initial ELP level strongly influences the expected time frame for their attaining English proficiency. Using these analyses, more refined time-to-English-proficiency criteria could be derived for EA 1. For example, EL students at initial ELP level 3 might be expected to attain the English-proficient criterion in two years, initial ELP level 2 students in four years and initial ELP level 1 students in six or seven years.

Summary

Several considerations should be addressed when interpreting findings. First, only one cohort of students was used in analyses. Ideally, states would use as many cohorts as is practicable. Second, states should use empirical approaches, along with the expert judgment of EL educators and informed policymakers, to determine an ambitious but reasonable time range for attaining the ELP performance standard. Using only empirical analysis to establish time frames is insufficient. Third, EA 1 uses multiple criteria to determine whether ELs require further English language instructional support. Analyses here only identify time to when students might be considered proficient. Fourth, two grade clusters within elementary grades were analyzed. Higher grades might exhibit very different time-to-proficiency characteristics. States should examine all grade clusters when conducting these types of analyses. Fifth, students with censored data were included in the event history analyses. That is the advantage of using event history analysis. Nonetheless, assumptions were made about how censored students performed. In performing event history analysis, states should deliberate on the assumptions they choose to make about censored cases. These assumptions may differ from those presented in the EA 1 example. Sixth, two approaches are used to explore this question. Other statistical methods could be employed, as well (e.g., mixed linear models). Approaches presented here should stimulate discussion among states and researchers on more precise methods of determining a time range for EL students to attain the English-proficient performance standard. Finally, the results presented here describe how long it is *observed* to take to reach the English-proficiency criterion—*not* how long it *should* take. Policymakers should take this into account when determining target time frames and percentages of EL students expected to meet them.

IV. Taking Into Account English Learners' English-Language Proficiency Level When Establishing Academic Progress and Proficiency Expectations

Overview

The preceding two chapters have illustrated empirical methods states can use to inform their deliberations on (1) determining an English-language-proficient performance standard on the state ELP test in relation to English Learners' (ELs') performance on academic content tests at different ELP levels; and (2) establishing a challenging and realistic time range to attain that English-language-proficient performance standard. This final chapter explores empirical methods that states can use to inform deliberations on setting academic progress and proficiency expectations for ELs that reasonably take into account students' ELP levels and their time in the state school system. Viewed through the lens of Title III requirements, where the prior two chapters focused on issues related to ELP progress over time (AMAO 1), and a rigorous English-language-proficient performance standard (AMAO 2), this chapter bears directly on AMAO 3, and therefore necessarily on Title I academic performance criteria and targets for the EL subgroup because these achievement expectations and results (called adequate yearly progress, or AYP) are applied as AMAO 3 for Title III subgrantees.

There has long been concern among researchers and policy analysts regarding *ESEA* Title I AYP's status bar academic performance expectation and its 100 percent academic-proficiency-by-2014 target (see, for example, Linn, 2005; Ho, 2008; Ryan and Shepard, 2008). In partial response to the first concern, the U.S. Department of Education's *Growth Model Pilot Project* (U.S. Department of Education, 2005) allowed approved states to refine their Title I accountability systems in order to recognize and receive credit for students meeting predefined academic growth criteria as being "on track" to meet or exceed the state's academic proficiency performance standard within a reasonable time frame, and therefore making AYP, the law's 2014 100 percent proficiency requirement notwithstanding (U.S. Department of Education 2011). In none of these pilot states, however, are growth expectations set for ELs by their current or expected ELP level.

The issue of setting academic progress and proficiency expectations for ELs conditioned to some degree by their ELP level must be approached with extreme care. On the one hand, researchers have consistently noted that academic accountability provisions of the *ESEA* are particularly problematic when applied to the EL population, as limited English language proficiency fundamentally affects an EL student's capacity both to benefit from academic content instruction delivered in English and to demonstrate knowledge and abilities on academic assessments given in English (Abedi 2004; Francis and Rivera 2007). On the other hand, educational rights advocates have been wary of proposals to establish differential expectations in academic progress and performance for different subgroups of students, out of concern that these could easily lead to lower expectations and diminished attention by educators to these students' academic needs (National Council of La Raza 2006; Education Trust 2006). Moreover, the now well-recognized, long-term English Learner phenomenon (Olsen 2010)—whereby students are unable to meet the linguistic and other (often academic) criteria required to exit EL status despite many years in the U.S. school system—suggests that conditioning academic expectations or results solely on

students' ELP level, without equal regard to their time in the state's school system, could have unintended negative consequences for this population.[23]

In light of these issues, and anticipating *ESEA* reauthorization, a national group of EL researchers has recently argued for the law to incorporate time into accountability provisions for the acquisition of English language proficiency, and to require states to establish expected time frames for the development of ELP. Moreover, these researchers have argued that, for each EL assessed in English, states should incorporate ELs' ELP into accountability provisions for content area achievement *using these expected time frames* (Working Group on ELL Policy 2010, 2011).

Two key points are inherent in these interrelated recommendations: First, content area achievement results should be adjusted for EL students' ELP level. Second, given the importance of setting an expected time frame for EL students to attain ELP, this time frame should be part of the content area performance result adjustments. Two methods are mentioned in the Working Group's documents as possibilities for satisfying these requirements: progressive benchmarking and indexed (i.e., weighted) progress.

This chapter explores these methods by illustrating how they might be operationalized with empirical data for a state's policymaking deliberations. Specifically, it examines two approaches to the progressive benchmarking method, and one approach to the indexed progress method. Additionally, given potential equity concerns (or stakeholder reluctance) regarding conditioning academic performance expectations or results specifically for ELs, this chapter also considers a third method, which does not rely on EL students' ELP level—the status and growth accountability matrix approach. This method acknowledges student attainment of academic proficiency (i.e., the AYP performance standard) or a predetermined, acceptable level of student growth toward academic proficiency (e.g., a level of academic progress to be considered "on track" to proficiency in a reasonable time frame).

Key Approaches

We use two approaches to progressive benchmarking, and one approach to indexed progress; both methods take into account an EL's ELP level and time in the state school system when setting his or her academic progress and proficiency expectations. The third approach explored explicitly ignores an EL's ELP level when judging academic progress and proficiency. Each method is described in detail as follows.

1. The ***progressive benchmarking*** methods adjust either (a) EL students' content achievement scale scores or (b) EL students' weight (their individual "count"), based on each student's ELP level relative to his or her initial ELP level and time in the state school system. In this method, there is an expectation that (1) students will increase in English language proficiency annually from their level of initial English proficiency and that (2) students will increase in content achievement annually. Thus, while recognizing the effect of limited English proficiency on ELs' academic performance on tests given in English, scale score or calculative weight adjustments lessen as students increase in ELP level, as expected, or, if they do not, as they continue in EL status over time. At the end of the time frame expected for ELs to attain English language

23. For example, an EL at an "intermediate" ELP level might have his or her English or language arts (ELA) performance result adjusted in recognition of this level of English proficiency; yet the student may have spent several years in the state's school system at this same ELP level. It would be an unintended negative consequence if accountability policy conferred to a school system an "on track" academic judgment for this EL student as a result of her *lack* of expected ELP progress.

proficiency (or sooner if they attain that level), students' content achievement scores or calculative weights are no longer adjusted.[24] In essence, expected performance benchmarks progressively increase (and corresponding adjustments progressively decrease) to the point at which no adjustments are made at all.

2. The ***indexed progress*** method uses an ELs' ELP growth as a proxy for academic content performance on a weighted, time-sensitive basis for more newly arrived ELs who enter the state's school system at lower initial ELP levels. These weights and time frames are empirically derived for each subject matter and grade tested because "the impact of limited English proficiency on academic performance varies by subject matter and grade [e.g., ELs with lower levels of language proficiency have more difficulty demonstrating content knowledge in English or language arts compared with mathematics, and this difficulty increases at higher grade levels])" (Working Group on ELL Policy 2010, p. 5).
3. The ***status and growth accountability matrix (SGAM)*** method acknowledges student attainment of academic proficiency (i.e., the AYP performance standard) or a predetermined, acceptable level of student growth toward academic proficiency (e.g., a level of academic progress to be considered "on track" to proficiency in a reasonable time frame), without considering an EL's ELP level.

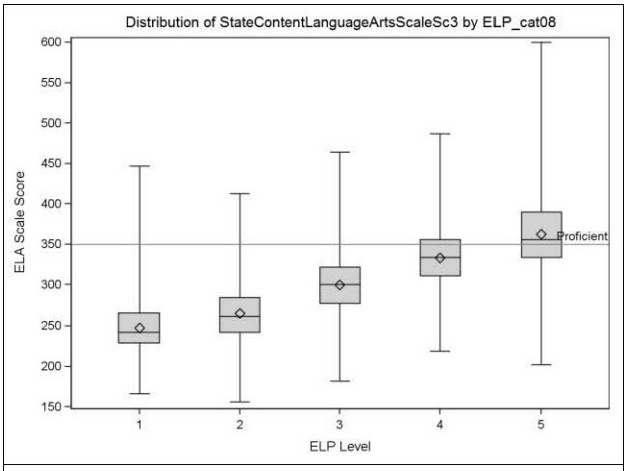
Before applying these methods using empirical data, the following presents a more in-depth explanation of each method.

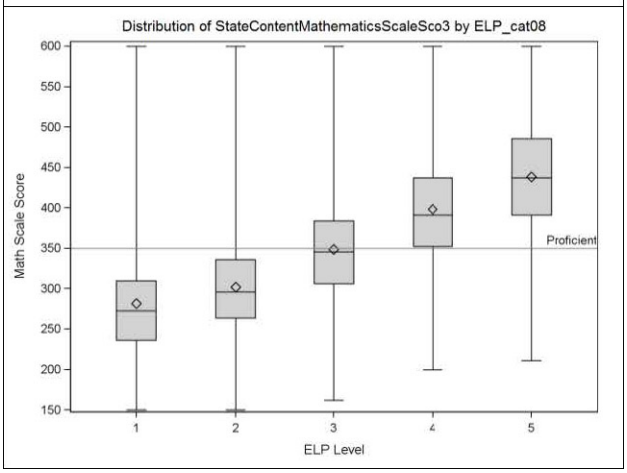
Method 1 (progressive benchmarking) adjusts either the scale scores or the way students are counted (weighted) in calculating results. To support the creation of adjustments, states need some knowledge of the distribution of EL students' content achievement scores over time and by ELP level. The box plots in Exhibit 18 below show the distribution of scores for EL students in grade 3 of Education Agency 1, the Education Agency selected for this chapter's illustrative analyses.

Two sets of box plots are shown, one for mathematics and one for English or language arts. On the horizontal axis, ELP levels are displayed. The vertical axis shows the content assessment scale score range. A horizontal line is drawn at the scale score value of 350, which represents the "proficient" cut score on the assessment for this grade. The boxes at each ELP level show the distribution of scale scores. The bottom line of each box is drawn at the 25th percentile of the distribution, and the top line represents the 75th percentile. The line drawn within each box shows the median score for that level, and the diamond displays the mean. The T-shaped lines ("whiskers") above and below each box show extreme scores for that distribution.

24. The importance of ensuring that the English-language-proficient performance standard is carefully examined relative to EL students' likelihood of attaining academic content area achievement standards is explored in Chapter II.

Exhibit 18.
Education Agency 1, Grade 3: Box Plots of English or Language Arts and Mathematics Scale Scores, by ELP Performance Level (2007–08)





ELP performance levels
for grade 3 (in scale score points)

Level 1 = Beginning (230–414)
Level 2 = Early Intermediate (415–459)
Level 3 = Intermediate (460–513)
Level 4 = Early Advanced (514–556)
Level 5 = Advanced (557–700)

Exhibit reads: Grade 3 EL students performing at the Early Advanced ELP performance level (or level 4) have a median performance of 334 scale score points in English or language arts, 16 points below the proficient performance standard.

Note: Appendix Exhibit F.1 presents the descriptive statistics associated with this graph.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Notice that the lower the ELP level, the lower the distribution of English or language arts and mathematics scale scores. At the lowest ELP levels, a vast majority of EL students are performing well below the proficient standard. The poor content test performance by low ELP-level ELs seen in these exhibits has been observed across many other grades and education entities (i.e., districts and states) and with other academic content and ELP assessments. While these exhibits do not show the length of time ELs have been in the system, they do provide a starting point for considering the setting of ambitious and realistic progress expectations.

Specifically, the distributions shown in the exhibits above can be used to establish "benchmarks" to adjust expectations based on ELP level. Adjusting content scores based on language proficiency level alone is insufficient, however. Students are expected to grow in their ELP (as discussed in Chapter III)

and those time-based expectations should be factored into content score adjustments. For illustrative purposes, a timeline like the one shown in Exhibit 19 below is adopted here.

Exhibit 19.
Expected English-Language Proficiency (ELP) Level Growth, by Year in State Schools

ELP Level	Expected ELP Level by Year in School			
	Initial Year	2nd Year	3rd Year	4th Year
Level 1	Level 1	Level 2	Level 3	Level 4
Level 2	Level 2	Level 3	Level 4	Proficient
Level 3	Level 3	Level 4	Proficient	–
Level 4	Level 4	Proficient	–	–

Exhibit reads: EL students starting at ELP level 1 in the initial year are expected to move to level 2 in the 2nd year, level 3 in the 3rd year, and level 4 in the 4th year; whereas students starting at ELP level 4 at outset are expected to become English language proficient in the 2nd year.

Students starting at the lowest level (Level 1) could potentially receive four years of adjustments. But each year these students would receive lesser adjustments based on the expectation that they will attain progressively higher ELP levels. Students at Level 4 would receive an adjustment only in their initial year. This table illustrates the progressively increasing expectations of ELP over time. Accordingly, academic content score adjustments lessen either as ELP level increases or, failing that, as ELs' time in the school system progresses.[25] And as will be seen in the worked examples below, adjustments can be made either to ELs' actual scale score result, or to their calculative weight.

25. Note that these linear ELP growth expectations are posited for the purpose of adjusting content performance expectations in relation to students' ELP level and time in the school system. Empirical analyses of ELP growth suggest that several factors may influence the way ELP growth occurs. (See the text box below for a brief elaboration.)

How Might Variability in Growth on ELP Assessments Inform the Setting of Annual Progress Expectations?

Although research about how ELs grow on ELP assessments is limited, the existing evidence suggests students at different ELP levels grow at different rates (Cook and Zhao 2011; Linquanti and others, forthcoming; Taylor and others, forthcoming). Exhibit 20 below depicts growth rates (in vertically-scaled ELP assessment scale score points) over three years for ELs at the third, fourth, or fifth grade in the initial year. (Only EL students with four years of ELP assessment scores were included.) EL students at lower initial ELP levels have much steeper growth rates than students at higher initial ELP levels. This characteristic has been observed on many different types of ELP assessments, motivating Cook and others (2008) to introduce the descriptive mnemonic, "lower is faster, higher is slower." That is, students at lower ELP levels tend to grow at higher rates than those at higher ELP levels. This observation also extends to grades: For a given initial ELP level, EL students at lower grades tend to grow at higher rates than their counterparts at higher grades.

Exhibit 20.
Rates of Growth in ELP Scale Score, Grades 3, 4, 5, by ELP Level in Base Year

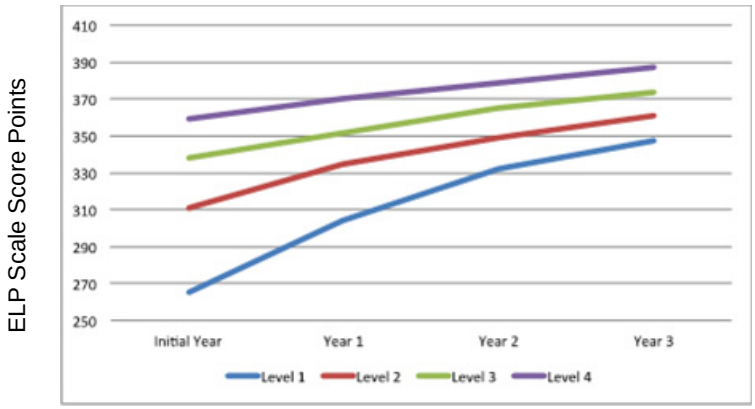


Exhibit reads: An EL in grade 3, 4, or 5 beginning at ELP level 4 in the initial year of the analysis is estimated on average to grow from an ELP scale score of 361 to 389 over a three-year period.

Source: Adapted from Cook and Zhao (2011).

These patterns suggest that setting a "one-size-fits-all" progress expectation for ELs may not be realistic. Doing so could set lower expectations for some ELs and potentially unattainable expectations for others. Nor does this exhibit reveal the whole story—many other factors might also influence ELP growth. For example, ELs who read at grade level in their native language will likely grow at different rates than students not literate in their native language—so too for students with interrupted formal education compared with those consistently enrolled in school. Beyond student characteristics, different EL instructional services and settings may influence ELP growth rates. All said, we need to better understand how ELs grow on ELP assessments and what characteristics and factors influence that growth.

Method 2 (indexed progress) takes an EL's growth in ELP and applies it as a proxy measure of ELA proficiency. That is, for those ELs at the lowest levels of initial English proficiency, sufficient growth in ELP scores can be used—on a weighted, time-sensitive basis—in lieu of (or combined with) their ELA proficiency scores. It should be noted that this method has in the past been disallowed under *ESEA* Title I AYP regulatory requirements. Nevertheless, given the Department's recent flexibility policy allowing states to request waivers from key *ESEA* accountability provisions, and *ESEA* Title I and Title III requirements that states' English language proficiency standards both reflect and support development of academic language found in state academic content standards, a rationale remains for arguing that, for more newly arrived, low ELP-level ELs, progress in ELP can reasonably indicate proximal progress in ELA.[26]

Specifically, for recently arrived ELs at the beginning stages of learning English, this "proxy" approach may offer a more valid application of test results than use of their performance on ELA assessments alone. Again, the use of ELP gains as a proxy for ELA should be only for a limited time.[27]

How might progress expectations be generated for consideration? Exhibit 21 shows second-to-third-grade growth in ELP for four language proficiency levels in Education Agency 1.

Exhibit 21.
English Language Proficiency Composite Scale Score Growth, by ELP Level, From Second to Third Grade, by Second-Grade ELP Level

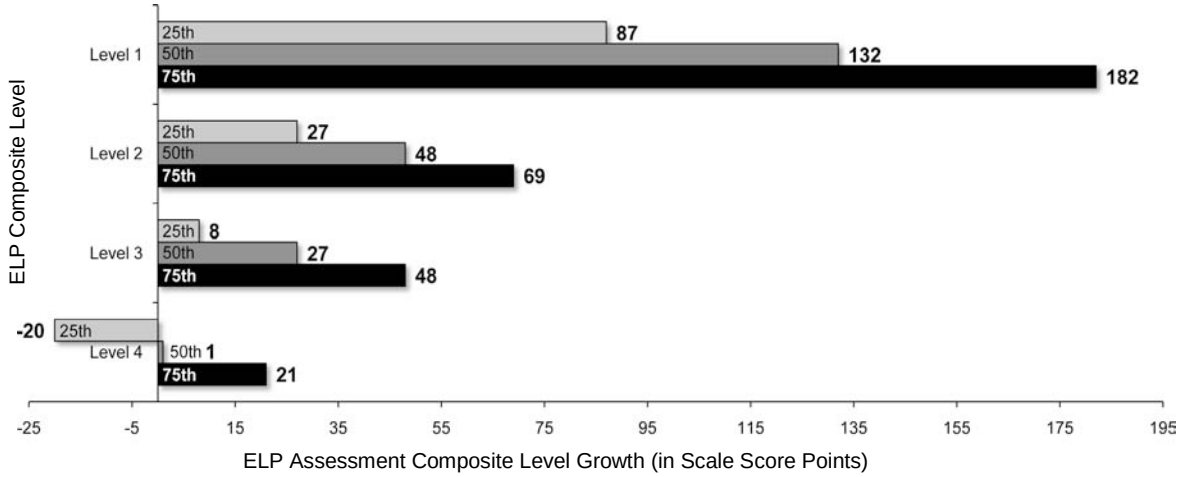


Exhibit reads: Among third-grade EL students who scored at ELP level 1 in second grade, 25 percent grew at a rate of 87 score points or less from second to third grade (i.e., such an EL at the 25th percentile of the distribution was estimated to have ELP composite scale score growth of 87 score points).

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

26. While a case might be made for ELP growth's signaling the potential for growth in mathematics, the relationship between these two subject areas is less compelling for arguing that ELP growth may serve as proxy for math performance.

27. Interestingly, current federal law allows states to waive ELs' ELA performance results in AYP accountability calculations for their first year in the state school system, regardless of ELP level. This method builds on this tacit acknowledgment of the problematic measurement issues in ELA for newly arrived low-ELP-level ELs, and uses empirical evidence of the relationship between ELP and ELA, as well as time-based ELP growth expectations.

The bar chart displays results for three percentile ranks: the 25th, 50th, and 75th percentiles. At ELP level 1, third-grade EL students with a composite growth score of 87 scale score points are at the 25th percentile (i.e., 25 percent of ELs grew at that rate or less). ELP 1 students with a composite growth score of 182 scale score points are at the 75th percentile. Notice that as ELP levels increase, growth rates decrease.[28] If an indexed progress method were adopted, growth expectations should differ based on ELP level. Potential acceptable target growth values generated from this approach could be used alone or in conjunction with a weighted or index formula with EL students' ELA scores to create a composite indexed progress result (e.g., ELP level 1 students might receive the following weights/index values: 70 percent ELP growth score, 30 percent ELA score). Time frames for applying these weights would also need to be set based on empirical evidence relative to expected or actual ELP growth.

Method 3 (status and growth accountability matrix) acknowledges student attainment of academic proficiency (i.e., the AYP proficient performance standard) or a predetermined, acceptable level of student growth toward academic proficiency (e.g., a level of academic progress to be considered "on track" to proficiency in a reasonable time frame), without considering an EL's ELP level. This method posits that when "on track" growth on content assessments is considered in concert with content proficiency (i.e., status), the need to adjust EL outcomes based on ELP level diminishes. This approach can be illustrated by the matrix in Exhibit 22, below. Note that there are four quadrants, which represent two dimensions of accountability: status (rows) and growth (columns).

Exhibit 22.
Status and Growth Accountability Matrix

Status on Content Assessment		Growth on Content Assessment	
		Low Growth	**High Growth**
	Proficient or Above on Content Assessment	I Students in this cell are proficient or advanced but are growing at lower rates than other students.	II Students in this cell are proficient or advanced and are growing at adequate rates compared with all other students.
	Not Proficient on Content Assessment	III Students in this cell are not proficient and are growing at lower rates than other students.	IV Students in this cell are not proficient but are growing at adequate rates compared with other students.

Exhibit reads: Students in quadrant I are proficient or advanced but are growing at lower rates than other students.

In this method, schools and districts are evaluated by how many students are proficient or how well students are growing. Students would either have to demonstrate academic proficiency in tested content (status) or adequate growth toward proficiency in content. Quadrant I characterizes students who have met the status requirement but have demonstrated low growth. Quadrant II characterizes students who have met both the proficiency and growth expectations. Quadrant III characterizes students who have met neither the status nor the growth requirement, and quadrant IV captures those students who have not met the status requirement but have met the growth expectation. Conceptually, students in quadrants I, II, and IV are meeting at least one requirement (status or growth). Students in quadrant III are meeting neither. Schools or districts that have substantial numbers of students in quadrant III might be identified

28. The negative growth reflected in the performance of EL students at the 25th percentile (i.e., the bottom quartile of the performance distribution) is not unusual for students at higher ELP levels. (See Cook et al. 2008; Linquanti et al. forthcoming.)

by the accountability system for further examination, and perhaps provided support. The proportion of students in quadrants I, II, and IV could be used in lieu of the percent proficient "status bar" in content area performance. Schools or districts doing well would have acceptable percentages of students in quadrants I, II, and IV.

The underlying argument of the status and growth accountability matrix approach is that ELs' demonstrating sufficient growth in academic content test performance obviates the need to adjust for their actual or expected ELP level. All students would be held to the same status or growth expectations.[29]

Method Examples

To illustrate their application, the methods described above are presented through worked examples. Only one grade is used (third grade) in these examples for economy of presentation, but potential users of these methods should be prepared to apply them to all tested grades. Further, as has occurred with methods illustrated elsewhere in this report, there will likely be variability in findings across grades. Thus, expert counsel and deliberation among appropriate stakeholder representatives will be necessary to establish final recommendations in the application of any of these methods. Given the complexities involved, the following worked examples are presented step-by-step in order to clearly describe both the procedures and decisions made in carrying out each method. Certainly, different procedures or decisions might be made that would prove equally plausible, and possibly better, given local preferences and constraints. Descriptions provided here are intended to be illustrative only and to stimulate discussion, inspire alternative approaches, and spur further research. After each method application, descriptive statistics are provided that highlight differences among content proficiency results with and without the application of the methods.

Method 1 (Progressive Benchmarking)

Two progressive benchmarking methods are applied in this section. The first method adjusts scale scores based on student distributions and creates a scaling factor. The second adjusts how ELs at different language proficiency levels are counted based on students' likelihood of being proficient on content assessments.

1.a. Adjusted Scale Score Method

This method takes the scale score distribution of EL students on content assessments (see Exhibit 18) and creates adjustments to those scores, which will in turn affect content proficiency scores. The following steps outline how this method is applied.

1. **Identify the EL and non-EL scale score distributions on relevant content assessments.** The two tables in Exhibit 23 below show the 25th, 50th, and 75th percentile ranks for non-EL and for EL students by ELP level for mathematics and ELA. Note that the ELs used to create this distribution have been part of EL programs for three years or less (i.e., only students in EL programs from first grade or later were used). The shaded cells highlight those points where scores are above the proficient level (proficient = 350)

29. Determining how much academic growth is sufficient to be judged "on track" is a critical decision in employing this method.

Exhibit 23a.
Distribution of Grade 3 Mathematics Scales Score for ELs (by ELP Level) and Non-ELs, in Education Agency 1

Groups	25th Percentile	50th Percentile	75th Percentile
ELs Level 1	245	282	336
ELs Level 2	287	325	377*
ELs Level 3	320	364*	407*
ELs Level 4	346	377*	437*
ELs Level 5	361*	417*	469*
Non-ELs	341	399*	451*

Exhibit reads: Third-grade EL students at ELP level 1 with a mathematics scale score of 245 points were at the 25th percentile, whereas the 25th percentile score for non-EL students was 341.

* The shaded cells highlight those points where scores are above the proficient level (proficient = 350).

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit 23b.
Distribution of Grade 3 ELA Scales Score for ELs (by level) and Non-ELs, in Education Agency 1.

Groups	25th Percentile	50th Percentile	75th Percentile
ELs Level 1	224	242	266
ELs Level 2	250	272	296
ELs Level 3	281	307	330
ELs Level 4	296	326	356*
ELs Level 5	315	347	384*
Non-ELs	311	342	377*

Exhibit reads: Third-grade EL students at ELP level 1 with a ELA scale score of 224 points were at the 25th percentile, whereas the 25th percentile score for non-EL students was 311.

* The shaded cells highlight those points where scores are above the proficient level (proficient = 350)

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

2. **Determine a scale score to apply to the adjustment formula.** That is, what percentile rank and corresponding scale score will be used to make the adjustment? For illustrative purposes, the 75th percentile rank and its associated scale score will be used.
3. Create a scale score adjustment factor[30] using the following formula:

30. A different and potentially more rigorous procedure might use linked and equated item response theory (IRT) theta values corresponding to the 75th percentile rank (if IRT is the measurement model employed) to make adjustments. Such information was not available in the current dataset.

$$\frac{Proficient\ Scale\ Score\ Value}{EL\ Scale\ Score\ Value\ at\ 75th\ percentile}$$

Note that for mathematics, only ELs at ELP level 1 have a scale score value less than 350 at the 75th percentile. In this case the scale adjustment factor in mathematics will only apply to level 1 students. In ELA, ELs up to ELP level 3 will receive the scale score adjustment.

4. **Determine the appropriate timelines for EL students to receive adjustments** (i.e., apply adjustments to students who are either "on track" in their ELP growth or apply what should be the value for their "on-track" ELP levels on the basis of time in the school system). For this analysis, the expected ELP growth timelines depicted in Exhibit 19 are used.
5. **Apply the scale score adjustment factor to EL students' original scale scores to generate adjusted scale score values.**

The following example illustrates how adjusted scores are generated and applied. Only ELP level 1 students who are in their initial year of EL program enrollment would receive a mathematics scale score adjustment (see Exhibit 23a). The adjustment factor for these students is 1.04 (350/336). If an ELP level 1 student in her initial year in an EL program received a mathematics scale score of 320, her adjusted score would be 333 (1.04 x 320). This would not be sufficient to be classified as proficient (meeting AYP). If a similar ELP level 1 student received a mathematics scale score of 338 (not proficient), her adjusted score would be 352 (1.04 x 338). This student would be identified as proficient for accountability purposes.

In ELA, EL students up to ELP level 3 would receive adjustments because ELP level 4 students at the 75th percentile have scores above the proficient cut point of 350 (see Exhibit 23b). To illustrate: An EL student in her second year in an EL program is at ELP level 2. Last year, she was also at level 2 (i.e., she did not advance in ELP level from last year to the year in consideration). She would not receive the ELP level 2 adjustment, rather, she would receive the level 3 adjustment, 1.06 (350/330). If she received an ELA scale score of 315, her adjusted score would be 334 (not proficient). To repeat: Only those EL students within expected ELP levels and time frames would receive scale score adjustments. The two tables in Exhibit 24 below show the scale score adjustments (shaded) by years in program and ELP level.

Exhibit 24a.
ELP Level Scale Score Adjustment Factor to be Applied to Grade 3 Mathematics Results

ELP Level	Years in Program			
	0 to 1	2	3	4
1	1.04*	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00

Exhibit reads: The scale score adjustment factor in grade 3 mathematics for EL students at ELP level 1 is 1.04 (350/336). Given that the factors in all other cells in the table are 1.00 the adjustment would not be applied to any other EL students.

Note: * Cells having a value above 1.00 are shaded.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit 24b.
ELP Level Scale Score Adjustment Factor to be Applied to Grade 3 English or Language Arts Results

ELP Level	Years in Program			
	0 to 1	2	3	4
1	1.32*	1.18*	1.06*	1.00
2	1.18*	1.06*	1.00	1.00
3	1.06*	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00

Exhibit reads: The scale score adjustment factor in grade 3 ELA for EL students at ELP level 1 is 1.32. Any other cells in the table with the value of 1.00 indicate that the adjustment would not be applied to EL students in that cell.

Note: * Cells having a value above 1.00 are shaded.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

1.b. Adjusted Counts Method

The Adjusted Counts Method differs from the Adjusted Scale Score Method in that *how EL students are counted* is what is adjusted, not students' scale scores. The usual percent proficient calculation on a content assessment is obtained by taking the total number of students receiving a proficient score and dividing that number by the total number of students participating (or who should be participating) on that assessment. The Adjusted Counts Method draws on empirical evidence showing that the likelihood of ELs at the lowest ELP levels reaching content proficiency is very small (see Exhibit 18). If the likelihood of attaining content proficiency is known, it can be used to adjust the way students are

counted in the denominator of the content proficiency calculation. The following steps describe how to apply this method.

1. **Identify the likelihood of attaining content proficiency for ELs who are "on track" to meeting the English language proficient performance standard.** Using logistic regression, probabilities are calculated, and Exhibit 25 below shows results from such an analysis. The exhibit displays the probability for ELs and non-ELs being proficient on content assessments of Education Agency 1. From this third-grade example, non-EL students have a 0.692 probability of being proficient in mathematics. They have a lower probability of being proficient in ELA (0.437).

Exhibit 25.
Probability of Being Proficient on Grade 3 Content Assessment for ELs and Non-ELs, Education Agency 1

Groups	Mathematics	ELA
ELs: Level 1	0.211	0.016
ELs: Level 2	0.392	0.034
ELs: Level 3*	0.629	0.152
ELs: Level 4	0.753	0.438
ELs: Level 4*	0.890	0.786
Non-ELs	0.692	0.437

Exhibit reads: Based on logistic regression, the predicted probability for being proficient on grade 3 mathematics is 0.211 for EL students at ELP level 1, and 0.692 for non-EL students. In the next column, both EL and non-EL students show a lower probability of being proficient in ELA.

Notes: * A conjunctive minimum criterion is used in defining this agency's English language proficient performance standard. That is, to be considered English proficient, ELs must attain a minimum composite score and minimum domain scores. Level 4* represents this agency's English-Language-Proficient performance standard. Level 3* (representing a conjunctive minimum of level 3 composite, with all domains level 2 or higher) is used to increase the difference in probability between ELP Levels 2 and 3 for adjusting content proficiency expectations upward.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

As expected, ELs at lower ELP levels have lower probabilities of being proficient on academic subject tests. Interestingly, for this Education Agency at this grade level, students at the English-language-proficient performance standard (level 4*) have greater probabilities of being proficient in ELA and in math than non-ELs.

2. **Establish counting factors based on probability estimates.** At this juncture in the process, deliberation with EL experts and stakeholders is critical to decide adjustment values for application. In the current example, adjustments for probability estimates in mathematics are employed by taking the obtained probability estimate rounded to the nearest decimal place: Level 1 = 0.2, Level 2 = 0.4, Level 3 = 0.6, and Level 4 (not meeting conjunctive minimum) = 0.8. Deciding adjustments for probability estimates at each ELP level in ELA, however, are more challenging because the change in probability from ELP levels 1 to 2 is quite modest, and from levels 2 to 3, levels 3 to 4, and level 4 to 4* is quite dramatic (see Exhibit 25). Rounding to

the nearest decimal place in ELA generates the following adjustments: Levels 1 and 2 = 0.0,[31] Level 3 = 0.2, and Level 4 = 0.4. For this example, therefore, slightly more rigorous adjustments are adopted: Level 1 = 0.1, Level 2 = 0.2, Level 3 = 0.4, and Level 4 (not meeting conjunctive minimum) = 0.5. Ultimately, adjustment values, though informed by empirical methods, are policy decisions; accordingly, others may arrive at different values for equally defensible reasons.

3. **Determine the appropriate timelines for EL students to receive adjustments** (i.e., apply adjustments to EL students who are "on track" in their ELP progress), while for those not on track, utilize the value that should be used were they at their "on-track" ELP levels based on time in the state school system. As above, for this analysis, the timelines outlined in Exhibit 19 are used.
4. **Create count adjustment tables for calculating adjusted student weights.** The tables in Exhibit 26a and 26b show adjustments (shaded) made by ELP level and time in program.

Exhibit 26a.
ELP Count Adjustment Values for Mathematics

ELP Level	Years in Program			
	0 to 1	2	3	4
1	0.20†	0.40†	0.60†	0.80†
2	0.40†	0.60†	0.80†	1.00
3	0.60†	0.80†	1.00	1.00
4	0.80†	1.00	1.00	1.00
4*, 5	1.00	1.00	1.00	1.00

Exhibit reads: The count adjustment value in mathematics for ELs at ELP level 1 and with 0 to 1 year in program was estimated to be 0.20. Any other cells in the table with the value of 1.00 indicate that the adjustment would not be applied to EL students in that cell.

Notes: The obtained probability estimates were rounded to the nearest decimal place.
4* represents this agency's English-Language-Proficient performance standard.
† Cells having a value above 1.00 are shaded.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

31. That is, this statistical technique yields values suggesting that there is virtually no probability of ELs at ELP levels 1 and 2 attaining proficiency on the state's ELA assessment. Indeed, this fact may lead some policy analysts to advocate the indexed progress method, discussed below.

Exhibit 26b.
ELP Count Adjustment Values for English or Language Arts

ELP Level	Years in Program			
	0 to 1	2	3	4
1	0.10†	0.20†	0.40†	0.50†
2	0.20†	0.40†	0.50†	1.00
3	0.40†	0.50†	1.00	1.00
4	0.50†	1.00	1.00	1.00
4*, 5	1.00	1.00	1.00	1.00

Exhibit reads: The count adjustment value in ELA for ELs at ELP level 1 and with 0 to 1 year in program was estimated to be 0.10. Any other cells in the table with the value of 1.00 indicate that the adjustment would not be applied to EL students in that cell.

Notes: The obtained probability estimates were rounded to the nearest decimal place.
4* represents this agency's English-Language-Proficient performance standard.
† Cells having a value above 1.00 are shaded.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

5. **Apply count adjustment values to calculate EL progress-to-proficiency trajectory results.** To calculate results for the EL subgroup,[32] the Adjusted Count formula would be

$$\frac{Number\ of\ [Eligible\ Former\ EL\ +\ Current\ EL]\ Students\ Proficient\ on\ Assessment}{Number\ of\ Eligible\ Former\ EL\ Students + \sum Count\ Adjustment\ Values\ for\ Current\ ELs}$$

To calculate results for current ELs only, the Adjusted Count formula would be

$$\frac{Number\ of\ Current\ EL\ Students\ Proficient\ on\ Assessment}{\sum Count\ Adjustment\ Values\ for\ Current\ ELs}$$

6. **Calculate EL progress-to-proficiency trajectory results based on one of the above formulas.** As a simple example, imagine that a school has 40 third-grade language-minority students, of whom 30 are current ELs and 10 are former (exited) ELs.[33] On this year's ELA assessment, eight of the 10 former-EL students scored proficient, while five of the 30 current ELs received a proficient score. What is the result for the EL subgroup? Without adjustment, 32.5 percent (13/40) of third-grade EL subgroup students in this school are deemed proficient in ELA. Of the 30 current ELs, however, 10 are new students this year and are at ELP level 1. Another 10 are ELs who have been in an EL program for two years and are at ELP level 2, and the third 10 are ELP level 4 students who have been in an EL program for three years. When the count adjustments are applied to these students (Level 1 student = 0.1, Level 2 student = 0.4, and Level 4 student = 1.0), the percentage of the EL subgroup meeting the expected performance standard changes from 32.5 percent to 52 percent:

32. Under current federal regulation, former ELs may be included in EL subgroup results for the two years following their exit from EL status.

33. For this example, we assume all former EL students exited EL status within the past two years.

$$\frac{8+5}{10+\sum[(10*.1)+(10*.4)+(10*1.0)]}$$

Although this hypothetical example includes eligible former ELs in the calculation, results also can be calculated to examine outcomes for current ELs only in order to illustrate more clearly the effect of the method's application. Without adjustment, the results for this school's third-grade current EL students would be 16.7 percent (5/30). Applying the Adjusted Count formula using the results and characteristics of current ELs described above, the percentage of current ELs meeting the expected performance standard changes from 16.7 percent to 33.3 percent:

$$\frac{5}{\sum[(10*.1)+(10*.4)+(10*1.0)]}$$

In both cases—for the EL subgroup including eligible former ELs, and for current ELs only—the adjustments increase the overall percentage meeting the expected performance standard because ELP levels 1 and 2 students are counted using an ELP level/time-sensitive weight compared with former-EL students or current ELs at higher ELP levels or in EL programs for extended time periods.

The ELP expected growth timelines (Exhibit 19) and the scale score or count adjustment tables (Exhibits 24a and 24b, or Exhibits 26a and 26b, respectively) were used in applying the two progressive benchmarking methods to Education Agency 1's third-grade EL data. Exhibit 27 presents outcomes of applying these methods to results of *current EL students* in third grade in order to give a sense of their effect.

Exhibit 27.
Content Proficiency Outcome Comparisons of Progressive Benchmarking Methods, for English Learners in Grade 3 (N = 18,101), Education Agency 1

Method	Percent Proficient
Mathematics Proficiency (no method applied)	39.3%
1.a. Mathematics Proficiency using Scale Score Adjustments	39.4%
1.b. Mathematics Proficiency using Count Adjustments	42.0%
ELA Proficiency (no method applied)	6.3%
1.a. ELA Proficiency using Scale Score Adjustments	7.6%
1.b. ELA Proficiency using Count Adjustments	7.0%

Exhibit reads: For EL students in third grade, 39.3 percent scored proficient or above in mathematics without adjustments, and progressive benchmarking methods 1 and 2 were applied higher proportions of EL students were classified as proficient (39.4 percent and 42.0 percent, respectively).

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

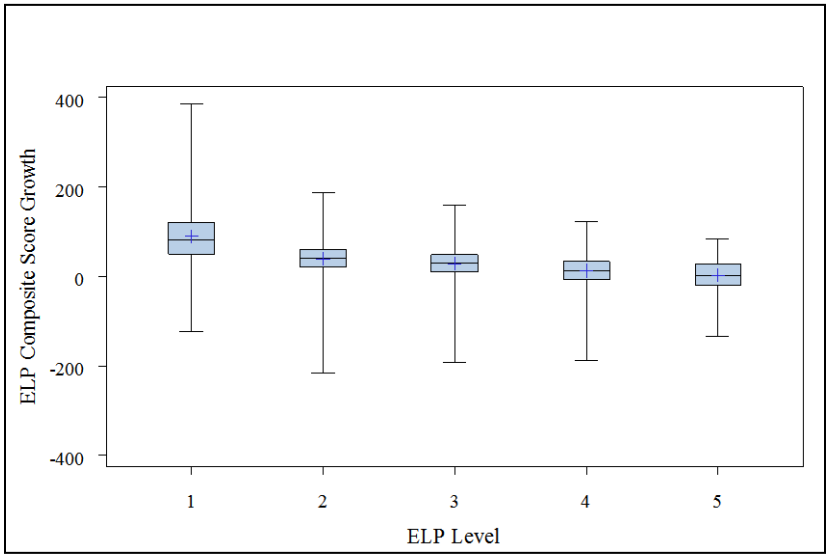
As expected, both progressive benchmarking methods yield higher proportions of EL students classified as proficient (meeting AYP). The scale score adjustment method shows the smallest percentage point increase in mathematics (0.1 percentage points) and the largest in ELA (1.3 percentage points). The student count adjustment method yields an increase of 2.7 percentage points for mathematics and 0.7 percentage points for ELA.

Method 2 (Indexed Progress)

In the worked example, this method is applied only to ELA. The intent is to use gain in ELP as a limited-term proxy for emerging performance in ELA. The following steps illustrate the application of this method.

1. **Identify the progress model to be used to define ELP growth.** As previously noted, there has been much recent state experimentation with progress and growth models, especially relating to AYP. Also, some exploratory research has been conducted regarding growth models and ELs (e.g., Cook and others 2008; Cook and Zhao 2011). Several growth-based procedures are available (e.g., simple gain-based models, percentile growth charts, student growth percentiles, value tables, value-added models). The model applied here is percentile growth charts, which calculate composite scale score gains and rank them by ELP level. The purpose is to identify the distribution of growth scores for each ELP level. In Exhibit 28, the box plot displays composite ELP assessment scale score gains for ELs from second to third grade in Education Agency 1.

Exhibit 28.
Composite ELP Assessment Scale Score Gains for ELs From Second to Third Grade (2007–08), Education Agency 1



ELP performance levels for grade 2 (in scale score points)

Level 1 = Beginning (215–396)
Level 2 = Early Intermediate (397–446)
Level 3 = Intermediate (447–495)
Level 4 = Early Advanced (496–539)
Level 5 = Advanced (540–635)

Exhibit reads: The box plots show a tendency that as students' ELP level increases, and the median and average ELP assessment growth values decrease.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Note that as a student's ELP level increases, the median and average ELP assessment growth values decrease. (Recall this pattern was also observed earlier in Exhibit 21.) This suggests that states might well need to set different growth expectations for EL students at different ELP levels.

2. **Once an ELP progress model is adopted, establish expectations for "acceptable" ELP growth.** In prior worked examples, the 75th percentile level was adopted as a criterion. For consistency, it is adopted for this approach as well. Using values from Exhibit 21, this education agency could posit that to meet ELP growth expectations, EL students must demonstrate the

following composite scale score gains (in scale score units): ELP Level 1 = 182, ELP Level 2 = 69, ELP Level 3 = 48, and ELP Level 4 = 21.[34]

3. **Determine the time frame for EL students to receive adjustments** (i.e., apply adjustments to students who are "on track" in their ELP development or what adjustment should be applied were they at their "on-track" ELP levels by time in the state school system). As with prior methods, the expected time frames outlined in Exhibit 19 are used.
4. **Create an Indexed Progress Gain Table, from which acceptable ELP scale score gain can be determined for ELA proxy purposes.** A sample table is shown below in Exhibit 29. The expectations for ELP progress by year in the state school system are similar to previous methods. ELP level 1 students in their initial year are expected to make the greatest ELP gains. Level 4 students in their initial year are expected to make the smallest ELP gains. The table cells with dashes are not part of the Indexed Progress Method (i.e., ELP growth is not used as a proxy for ELA performance). For EL students represented by these cells with dashes, their ELA assessment result is used.

Exhibit 29.
Indexed Progress Gain Values (in ELP Assessment Composite Scale Score Units) as Proxy for English or Language Arts, by Student ELP Level and Years in State School System

ELP Level	Years in Program			
	0 to 1	2	3	4
1	182	69	48	21
2	69	48	21	—
3	48	21	—	—
4	21	—	—	—
5	—	—	—	—

Exhibit reads: The indexed progress gain value estimated for ELs at ELP level 1 and with 0 to 1 year in program was 182 ELP assessment composite scale score units.

Note: This table was based on 75th percentile values from Exhibit 21and appropriate timelines for EL students to receive adjustments from Exhibit 19. Thus if an EL student who started at ELP level 1 has been in an EL program for two years, a value of 69 index progress gain score would be used. If the student's ELP scale score gain value was equal to or higher than that value of 69, the student would be counted as meeting the ELP performance standard for accountability calculations.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

5. **Apply indexed progress results as proxy for ELA assessment performance for eligible ELs.** The ELP levels and timelines as displayed in Exhibit 29 are applied for these EL students. If a student is at or exceeds the expected ELP scale score gain value, she is counted as meeting the expected ELA performance standard (meeting AYP) for accountability calculations. Exhibit 30 below shows the difference in ELA percent proficient outcomes with and without the Indexed Progress method applied to third-grade ELs.

34. These values correspond to the value of the top line of each box plot in Exhibit 27.

Exhibit 30.
ELA Proficiency Outcome With and Without Indexed Progress Method Applied, for ELs in Grade 3 (N = 18,101), Education Agency 1

Method	Percent Proficient
ELA Proficiency (no method applied)	6.3%
2. ELA Proficiency using Indexed Progress	17.4%

Exhibit reads: Without an adjustment, 6.3 percent of EL students at grade 3 scored proficient or above in the ELA performance assessment. After applying the indexed progress method, 17.4 percent of EL students were counted as being proficient in ELA.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

The Indexed Progress Method identifies much greater proportions of third-grade EL students as meeting the ELA performance standard (being "proficient") because it places greater value on the ELP progress of more recently-arrived ELs at lower ELP levels,as a temporary proxy for ELA performance. The percentage point increase in the ELA outcome is much greater (11.1 percentage points) than that observed for the two Progressive Benchmarking methods above.

Method 3 (Status and Growth Accountability Matrix)

As previously noted, this approach does not use an EL students' ELP level as a mediating factor. This method assumes that either growth in content performance or attainment of content proficiency is sufficient for accountability for all students, including ELs. This approach assumes a growth model is applied to content area assessments. Its application requires the following steps.

1. **Identify an appropriate growth model.** As mentioned earlier, there is much current experimentation with different types of gain or growth models.[35] Generally, states currently use three types of growth models: value tables, student growth percentiles, or value-added models. Detailed descriptions of these models are beyond the scope of this chapter.
2. **Apply a growth model to all students.** This example uses student growth percentiles (SGP) as applied by Betebenner (2008). The SGP tool was used to calculate growth percentile results for all students in our sample education agency's second-to-third-grade dataset.
3. **Establish acceptable growth values.** The SGP tool provides growth percentile scores for all students with sufficient data. Following previous examples, a student growth percentile of 75 was adopted as the growth criterion for this example.
4. Create a status and growth accountability matrix (SGAM). Exhibit 31 shows the matrix used in this example.

35. See, for example, Auty and others (2008) and Betebenner (2009).

Exhibit 31.
Status and Growth Accountability Matrix

Status on Content Assessment		Growth on Content Assessment	
		Low Growth	**High Growth**
	Proficient or Above on Content Assessment	I Content Scale Score ≥ 350 and Student Growth Percentile < 75	II Content Scale Score ≥ 350 and Student Growth Percentile ≥ 75
	Not Proficient on Content Assessment	III Content Scale Score < 350 and Student Growth Percentile < 75	IV Content Scale Score < 350 and Student Growth Percentile ≥ 75

Exhibit reads: ELs in quadrant I scored proficient or above on the content assessment but exhibited growth of the content assessment that was lower than the growth recorded by the fastest growing 25 percent of ELs.

5. **Determine weights to calculate the percentage of students meeting the status and growth criteria.** In this example, if a student's status and growth fell within quadrants I, II, or IV, he received a weight of 1. If a student status and growth fell within quadrant III, he was assigned a weight of 0. Accountability percentages are calculated as the sum of students receiving a 1 divided by the total number of students. Exhibit 32 shows differences in outcomes obtained by applying this method to all third-grade students' results for mathematics and ELA.

Exhibit 32.
Comparison of Content Proficiency Outcomes With and Without the Status and Growth Accountability Matrix (SGAM) Method Applied, for All Grade 3 Students (N = 48,394), Education Agency 1

Method	Percent Proficient
Mathematics Proficiency (no method applied)	59.4%
3. Mathematics Proficiency using SGAM Method	61.6%
ELA Proficiency (no method applied)	29.9%
3. ELA Proficiency using SGAM Method	39.3%

Exhibit reads: For all grade 3 students, including both ELs and non-ELs, 59.4 percent scored proficient or above in mathematics without an adjustment. After applying the SGAM method, 61.6 percent were proficient or above (an increase of 2.2 percentage points).

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

The SGAM method yields higher proportions of students identified as making acceptable growth or attaining content proficiency. Since the results shown are for all students, Exhibit 33 disaggregates outcomes for ELs versus non-ELs.

Exhibit 33.
Comparison of Content Proficiency Outcomes With and Without the Status and Growth Accountability Matrix (SGAM) Method Applied, for EL and Non-EL Students in Grade 3, Education Agency 1

Group	N	Method	Percent Proficient
Non-EL	30293	Mathematics Proficiency (no method applied)	71.2%
Non-EL	30293	3. Mathematics Proficiency using SGAM Method	72.5%
Non-EL	30293	ELA Proficiency (no method applied)	44.0%
Non-EL	30293	3. ELA Proficiency using SGAM Method	50.5%
EL	18101	Mathematics Proficiency (no method applied)	39.3%
EL	18101	3. Mathematics Proficiency using SGAM Method	43.4%
EL	18101	ELA Proficiency (no method applied)	6.3%
EL	18101	3. ELA Proficiency using SGAM Method	20.7%

Exhibit reads: With no adjustment, 71.2 percent of grade 3 non-EL students were proficient in mathematics, and with the SGAM method applied, 72.5 percent were proficient (an increase of 1.3 percentage points).

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

As Exhibit 33 illustrates, 71.2 percent of grade 3 non-EL students were proficient in mathematics with no adjustment, and when the SGAM method is applied, 72.5 percent were so, an increase of 1.3 percentage points. For third-grade ELs, 39.3 percent were proficient in mathematics with no adjustment, and under SGAM, 43.4 percent were so, an increase of 4.1 percentage points. ELs benefit more than non-ELs in mathematics with this procedure. A substantially higher percentage increase for ELs occurs when the method is applied to ELA. EL percent proficient results increase by 14.4 percentage points (from 6.3 percent to 20.7 percent) when SGAM is applied, versus 6.5 percent for non-ELs.

Comparison of Method Outcomes

The methods illustrated above yield varying differences in third-grade outcomes at the Education Agency level. While some methods appear to have little effect on EL results at grade 3, greater effects might be observed when all tested grades within a school are calculated and combined. In fact, there are notable differences in outcomes at the school level, depending upon the salient characteristics of ELs. Given that these methods set out to take into account newer, low-ELP-level EL students' academic performance in rigorous, meaningful ways (e.g., by ELP level, time in the school system, growth in ELP or academics), the following exhibits show each method's effects on *schools'* third-grade current EL outcomes, disaggregated by differing densities of new ELs (i.e., ELs with three years or less in an EL program) for mathematics and ELA. In these exhibits, the term "Low" characterizes schools where less than 5.5 percent (the 25th percentile) of their third-grade ELs would be classified as new and "High" characterizes schools where 15.6 percent (the 75th percentile) or more of their third-grade ELs would be classified as new.

Exhibit 34.
Method Outcome Comparisons for ELs (N = 18,101) in Mathematics at Grade 3, by Density of New ELs in Schools, Education Agency 1

		Mean Percent Proficient in Mathematics			
		Schools Clustered by Density of New ELs			
Method		Low (N=115)	Moderate (N=230)	High (N=113)	All Schools (N=458)
No method applied	Mean	47%	40%	47%	43%
	Std	0.24	0.15	0.21	0.20
1.a. ELP Level Adjusted Scale Score Method	Mean	47%	40%	47%	43%
	Std	0.24	0.15	0.21	0.20
1.b. ELP Level Adjusted Count Method	Mean	48%	42%	56%	47%
	Std	0.25	0.16	0.30	0.23
3. Status and Growth Accountability Matrix Method	Mean	51%	44%	49%	47%
	Std	0.24	0.15	0.21	0.20

Exhibit reads: When no adjustment methods were applied 47 percent of third-grade ELs in schools with low densities of new ELs were proficient in mathematics.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

The third-grade sample is drawn from 458 schools in Education Agency 1. Across all schools, the average percent proficient in mathematics for third-grade ELs is 43 percent (see the first row of the "All Schools" column). Schools with low densities of new ELs show on average 47 percent of their EL students are proficient in mathematics. Schools with moderate densities of new ELs averaged 40 percent proficient, and schools with high densities of new ELs averaged 47 percent proficient. There are no outcome differences across schools between the unadjusted ("no method applied") and the adjusted scale score progressive benchmarking method. The adjusted count progressive benchmarking method shows a 1 percentage point increase among schools with low densities of new ELs, a 2 percentage point increase among schools with moderate densities, and a 9 percentage point increase among schools with high densities of new ELs. The SGAM method shows a 4 percentage point increase among schools with low and moderate densities and a 2 percentage point increase among schools with high densities of new ELs. Exhibit 35 compares outcomes of the various methods applied to ELA results.

Exhibit 35.
Method Outcome Comparisons for ELs (N = 18,101) in English or Language Arts at Grade 3, by Density of New ELs in Schools, Education Agency 1

Method		Mean Percent Proficient in English Language Arts			
		Schools Clustered by Density of New ELs			
		Low (N=115)	Moderate (N=230)	High (N=113)	All Schools (N=458)
No method applied	Mean	10%	7%	9%	8%
	Std	0.17	0.07	0.11	0.11
1.a. ELP Level Adjusted Scale Score Method	Mean	11%	8%	14%	10%
	Std	0.17	0.08	0.14	0.12
1.b. ELP Level Adjusted Count Method	Mean	12%	7%	12%	10%
	Std	0.23	0.08	0.16	0.15
2. ELP Indexed Progress Method	Mean	22%	17%	22%	20%
	Std	0.21	0.10	0.15	0.15
3. Status and Growth Accountability Matrix Method	Mean	23%	21%	22%	22%
	Std	0.19	0.11	0.16	0.15

Exhibit reads: When no adjustment methods were applied, 10 percent of third-grade ELs in schools with low densities of new ELs were proficient in ELA.

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Across all schools, 8 percent of third-grade EL students scored proficient in ELA. Among low- and high-density new EL schools, the percentage of ELs scoring proficient in ELA increased to 10 percent and 9 percent, respectively, while 7 percent were proficient in moderate-density schools. When the adjusted scale score progressive benchmarking method is applied, the percentage of ELs deemed proficient among low-, moderate-, and high-density new EL schools increased by 1, 1, and 5 percentage points, respectively. The adjusted count progressive benchmarking method increased outcomes by 2 and 3 percentage points for low- and high-density schools, respectively, but no change was observed among moderate density schools. The indexed progress method (utilized only on ELA), yielded increases of 12, 10, and 13 percentage points among schools with low-, moderate-, and high-densities of new ELs, respectively. This represents a much larger change in outcomes compared with the progressive benchmarking methods. Finally, the SGAM method yielded increases of 13, 14, and 13 percentage points among schools with low-, moderate-, and high-densities of new ELs, respectively.

Summary and Caveats

Of the different methods explored, the Status and Growth Accountability Matrix generated the greatest percentage point differences in the results of third-grade EL students. This model does not adjust for English language proficiency level, however. Also *all* students, not just ELs, were used to generate the student growth percentile values which affected school results, although the method did yield a greater change in outcomes for ELs than for non-ELs. Finally, the worked example above could not model whether EL student academic growth at the 75th percentile (the standard for being considered "high growth")—though quite rigorous—would be sufficient to ensure that such EL students are "on track" to attain academic proficiency in a reasonable time frame. The Indexed Progress method yielded the next highest range of percentage point differences, though this method was applied only to ELA results of the restricted subset of eligible EL students. It is unclear how this method would influence mathematics or ELA proficiency outcomes if some type of composite (e.g., ELP growth + mathematics achievement

or ELP growth + ELA achievement) were utilized. The Progressive Benchmarking methods generated the least changes in outcomes. Of the two Progressive Benchmarking methods explored, the Adjusted Counts method yielded greater differences, compared with the Adjusted Scale Scores method, although these were still very modest differences (1 to 2 percentage points) in the low- and moderate-density new-EL schools, and 3 to 5 percentage points in the high-density new-EL schools. The change in outcomes under these methods was likely modest because many of the eligible third-grade EL students in Education Agency 1 to which these methods were applied had "timed out" of the possibility to benefit from such adjustments due to their higher initial ELP level or lack of expected ELP progress by time in the school system.

As evidenced above, implementing any of these methods requires important decisions. For example, the 75th percentile was chosen as a guiding criterion. Such decisions are more policy related than empirical. The 75th percentile was chosen here to establish more rigorous expectations and to provide some degree of consistency for method comparisons. Also, an ambitious time frame was utilized for expected progress to the English-language proficient performance standard (i.e., four years of those initial ELP-level 1 ELs, with proportionally less time for those beginning at higher initial ELP levels). Other criteria could very well provide different scenarios and would certainly generate different results. As noted throughout this report, expert stakeholders must understand assumptions made within methods, how empirical data are used, and be presented with policy options in making decisions. There will be differential impacts resulting from the chosen method and criteria, and experts must be informed to support decisionmaking with a focus on ELs.

There are numerous caveats. First, the methods illustrated in this chapter are exploratory and meant to spur discussion and foster further research. They are by no means definitive and should not be viewed as such. Second, the outcomes generated were based only on one grade in one education agency. Other grades and different populations of students from other educational agencies might yield different results. Generating results for all grades tested was beyond the scope of this chapter, and may very well alter the outcomes for the education agency featured in the worked examples. Moreover, other education agencies with different characteristics may well have different findings. Any method explored should be applied to all grades for which data are available. Third, employing different criteria will likely result in different findings. A state may choose to examine these methods using optimal criteria. For example, such an approach might examine external criteria regarding school or teacher quality, and each method could then be compared using these schools. Analyses could be conducted to determine which method most often identifies schools with high-quality instruction. Fourth, the consequences of implementing these models must be considered. Not only should outcome data be generated, but careful consideration also should be paid to a method's degree of intuitive appeal, comprehension, and perceived fairness, particularly among key stakeholder groups. This should be explored before adopting any method. Fifth, substantial statistical and database capacity is required to implement the methods described in this chapter, particularly for the Status and Growth Accountability Matrix method. Finally, the ultimate goal behind these methods is to more accurately determine and represent how ELs are performing on state content assessments, assessments which in many cases are not designed for low-ELP-level ELs. The methods presented here yield results that are influenced by the ELP and content assessments employed by this education agency, for its EL population. Differences in assessments, performance standards, and EL students will all affect how these methods function and the results each generates.

Nevertheless, the empirical methods explored in this chapter can help policymakers begin to address key factors highlighted throughout this report that shape assessment and accountability for English Learners: Namely, an EL's level of English language proficiency fundamentally affects their academic performance on assessments conducted in English; it takes ELs time to develop levels of ELP needed to benefit from instruction in English and perform on these assessments; this time frame varies based on several factors,

including students' level of initial English proficiency, age and grade on entry, and the quality of instructional services they receive relative to their linguistic and academic needs; rigorous yet reasonable time-based expectations for ELs to learn English for academic performance can be set and monitored; and expectations for ELs' progress toward academic proficiency can incorporate and reflect those rigorous English-learning expectations.

While methods explored in this chapter are complex, they efficiently capture a complex reality facing EL students and their educators. As such, they offer a start to help state decisionmakers develop expectations that establish rigorous accountability and that address a fundamental need for fairness and realism. Such methods may also stimulate discussion and experimentation among education agencies to develop more nuanced ways of measuring and evaluating ELs' academic progress and proficiency in relationship to their initial ELP levels and time in the school system.

References

Abedi, J., ed. 2007. *English language proficiency assessment in the nation: Current status and future practice*. Davis, CA: University of California. Posted at: http://education.ucdavis.edu/sites/main/files/ELP_Report.pdf (accessed June 15, 2010).

Abedi, J. 2004. *The No Child Left Behind Act* and English-language learners: Assessment and accountability issues. *Educational Researcher*, 33 (1): 4–14.

American Educational Research Association, American Psychological Association, and National Council of Measurement in Education. 1999. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Auty, W., P. Bielawski, T. Deeter, G. Hirata, C. Hovanetz-Lassila, J. Rheim, P. Goldschmidt, K. O'Malley, R. Blank, and A. Williams. 2008. *Implementer's guide to growth models*. Washington D.C.: Council of Chief State School Officers.

Betebenner, D. 2009 (April 6). *Growth, standards and accountability*. National Center for the Improvement of Educational Assessment, Inc. (NCIEA): Dover, NH. Posted at: http://www.nciea.org/publications/growthandStandard_DB09.pdf.

Betebenner, D. 2008 (March 20). *Norm- and criterion-referenced student growth*. National Center for the Improvement of Educational Assessment, Inc. (NCIEA): Dover, NH. Posted at: http://www.nciea.org/publications/normative_criterion_growth_DB08.pdf.

Box, G. E. P., and N. R. Draper. 1987. *Empirical model-building and response surfaces*. New York: Wiley.

Butler, F. A., and M. Castellon-Wellington. 2000 and 2005. Students' concurrent performance on tests of English language proficiency and academic achievement. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives*. CSE Tech. Rep. No. 663, 47–77. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Cook, H. G., T. Boals, C. Wilmes, and M. Santos. 2008. *Issues in the development of annual measurable achievement objectives for WIDA consortium states*. WCER Working Paper No. 2008-2. Madison, WI: University of Wisconsin–Madison, Wisconsin Center for Education Research. Posted at: http://www.wcer.wisc.edu/publications/workingPapers/papers.php (accessed December 7, 2010).

Cook, H. G., E. Hicks, S. Lee, and R. Freshwater. 2009. *White paper on methods for establishing English language proficiency using state content and language proficiency assessments*. Unpublished manuscript.

Cook, H. G., and Y. Zhao. 2011 (April 8). *How English-language proficiency assessments manifest growth: An examination of language proficiency growth in a WIDA state*. Paper presented at the American Educational Research Association conference, New Orleans, LA.

Dietz, S. 2010. *How many schools have not made adequate yearly progress under the* No Child Left Behind Act? Policy report March 11, 2010. Washington, DC: Center on Education Policy. Posted at: http://www.cep-dc.org/index.cfm?fuseaction=document_ext.show DocumentByID&nodeID=1&DocumentID=303. (accessed July 14, 2010).

Education Trust. 2006. *ESEA: Myths versus realities. Answers to common questions about the No Child Left Behind Act*. Washington D.C.: Author. Posted at: http://www.edtrust.org/dc/publication/esea-myths-versus-realities (accessed October 10, 2009).

Federal Register, 76 (75), 21978–21984. April 19, 2011.

Federal Register, 73 (202), 61828–61844. October 17, 2008.

Francis, D. J., and M. O. Rivera. 2007. Principles underlying English language proficiency tests and academic accountability for ELLs. In *English language proficiency assessment in the nation: Current status and future practice,* ed. J. Abedi, 13–31. Davis, CA: University of California, Davis, School of Education.

Government Accountability Office. 2006. No Child Left Behind Act: *Assistance from education could help states better measure progress of students with limited English proficiency*. Report GAO-06-815 (July). Washington DC: Author.

Genesee, F., K. Lindholm-Leary, B. Saunders, and D. Christian. 2006. *Educating English language learners: A synthesis of research evidence.* New York: Cambridge University Press.

Haertel, E. 2002. Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice,* (21)1: 16–22.

Haertel, E. 2008. Standard setting. In *The future of test-based educational accountability,* eds. K. Ryan and L. Shepherd, 139–54. New York: Routledge.

Hakuta, K, Y., G. Butler, and D. Witt. 2000. *How long does it take English learners to attain proficiency?* The University of California Linguistic Minority Research Institute Policy Report 2000-1. Posted at: http://escholarship.org/uc/item/13w7m06g#page-1 (accessed November, 1 2010).

Hambleton, R. K., and M. J. Pitoniak. 2006. Setting performance standards. In *Educational measurement, 4th edition*, ed. R. L. Brennan, 433–70. Washington, DC: American Council on Education.

Ho, A. D. 2008. The problem with "proficiency:" Limitations of statistics and policy under *No Child Left Behind. Educational Researcher*, 37 (6): 351–60.

Kato, K., D. Albus, K. Liu, K. Guven, and M. Thurlow. 2004. *Relationships between a statewide language proficiency test and academic achievement assessments.* LEP Projects Report 4. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Posted at: http://education.umn.edu/NCEO/OnlinePubs/LEP4.html (accessed April 17, 2008).

Klein, J. P. and M. L. Moeschberger. 1997. *Survival Analysis: Techniques for censored and truncated data.* New York: Springer-Verlag New York, Inc.

Linn, R. L. 2008. Educational accountability systems. In *The future of test-based educational accountability*, eds. K. D. Ryan and L. A. Shepard, 3–24. New York: Routledge.

Linn, R. L. 2005. Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Educational Policy Analysis Archives*, 13 (33). Posted at: http://epaa.asu.edu/epaa/v13n33/ (accessed September 14, 2006).

Linn, R. L. 2003. Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives*, 11 (31). Posted at: http://epaa.asu.edu/epaa/v11n31/ (accessed April 15, 2010).

Linquanti, R., and C. George. 2007. Establishing and utilizing an NCLB Title III accountability system: California's approach and findings to date. In *English language proficiency assessment in the nation: Current status and future practice*, ed. J. Abedi, 105–18. Davis: University of California. Posted at: http://education.ucdavis.edu/sites/main/files/ELP_Report.pdf (accessed June 15, 2010).

Linquanti, R., E. Crane, and M. Huang. Forthcoming. *Examining growth in English language proficiency of California's English Learners*. Issues and Answers Report. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West.

Mehrens, W. A., and G. J. Cizek. 2001. Standard setting and the public good: Benefits accrued and anticipated. In *Setting performance standards: Concepts, methods, and perspectives*, ed. G. J. Cizek, 477–85. Mahwah, NJ: Lawrence Erlbaum.

National Council of La Raza. 2006. *Improving assessment and accountability for English language learners in the No Child Left Behind Act*. Issue Brief No. 16. Washington D.C.: Author.

Olsen, L. 2010. Reparable harm: Fulfilling the unkept promise of educational opportunity for long term English Learners. Long Beach, CA: Californians Together.

Parker, C. E., J. Louie, and L. O'Dwyer. 2009. *New measures of English language proficiency and their relationship to performance on large-scale content assessments*. Issues and Answers Report, REL 2009–No. 066. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Posted at: http://ies.ed.gov/ncee/edlabs (accessed December 2009).

Ryan, K. E., and L. A. Shepard, eds. 2008. *The future of test-based educational accountability*. New York: Routledge.

Stevens, R. A., F. A. Butler, and M. Castellon-Wellington. 2000. *Academic language and content assessment: Measuring the progress of English language learners (ELLs)*. CSE Tech. Rep. No. 552. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Taylor, J., M. Chinen, T. Chan, I. Brodziak de los Reyes, A. Boyle, C. Tanenbaum, and M. Petroccia. Forthcoming. *A Description of English Learner student achievement in six jurisdictions—National evaluation of Title III implementation*. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

The No Child Left Behind Act of 2001 (NCLB). Pub. L. 107–110, Jan 8, 2002. Stat.115. 1425–2094.

U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011. *Final report on the evaluation of the Growth Model Pilot Project*, Washington D.C. Posted at: http://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/index.html (accessed on September 12, 2011).

U.S. Department of Education. 2009. *Standards and assessments peer review guidance: Information and examples for meeting requirements of the* No Child Left Behind Act of 2001. Washington, DC: U.S. Department of Education. Posted at: http://www2.ed.gov/policy/elsec/guid/saaprguidance.pdf (accessed on May 12, 2011).

U.S. Department of Education, Office of Communications and Outreach. 2005. "Secretary Spellings Announces Growth Model Pilot, Addresses Chief State School Officers' Annual Policy Forum in Richmond," press release, November 18, 2005. Posted at: http://www2.ed.gov/news/pressreleases/2005/11/11182005.html (accessed on September 12, 2011).

Working Group on ELL Policy. 2011. *Improving educational outcomes for English language learners: Recommendations for* ESEA *reauthorization.* Policy Brief (March 25). Palo Alto, CA: Working Group on ELL Policy. Posted at: http://ellpolicy.org/wp-content/uploads/PolicyBrief.pdf (accessed on June 19, 2011).

Working Group on ELL Policy. 2010. *Improving educational outcomes for English-language learners: Recommendations for the reauthorization of the* Elementary and Secondary Education Act. Palo Alto, CA: Working Group on ELL Policy. Posted at: http://ellpolicy.org/wp-content/uploads/ESEAFinal.pdf (accessed May 26, 2010).

APPENDIX A:
DECISION CONSISTENCY METHOD

Appendix A. Decision Consistency Method

Adapted from several areas of measurement research (e.g., bias, reliability, and standard setting), this decision-theory approach examines the relationship between two sets of measures, in this case, academic content assessments and ELP assessments (Cook et al., 2009) (see Exhibit A.1 for a display of this relationship).

Exhibit A.1.
State ELP and Academic Content Assessment Decision Matrix

		State ELP Assessment Cut-Score (TBD)	
		Not Proficient (TBD)	**Proficient (TBD)**
State Content Assessment Proficient Cut Score (Given)	**Proficient**	Proficient on content Below language proficient (Quadrant I)*	Proficient on content Proficient in language (Quadrant II)
State Content Assessment Proficient Cut Score (Given)	**Not Proficient**	Below content proficient Below language proficient (Quadrant III)	Below content proficient Proficient in language (Quadrant IV)*

* Cells in gray are defined as inconsistent decisions.

The four cells characterize outcomes of decisions made with state content and ELP assessments. Cells in gray are defined as inconsistent decisions. Students classified in quadrant I scored proficient on the content assessment but less than proficient on the ELP assessment. Those classified in quadrant IV scored proficient on the ELP assessment but less than proficient on the given content assessment. The other cells are defined as consistent decisions based on results of both measures.

Assuming that a state academic content assessment's proficient performance standard is determined and fixed, this method allows users to identify the ELP assessment score band or value that maximizes the proportion of consistent decisions, thereby identifying a possible ELP performance standard for consideration. A decision consistency score or index value can be created for each individual score or score band as follows:

$$Decision\ consistency\ (DC) = \frac{quadrant\ II\ +\ quadrant\ III}{sum\ of\ quadrants\ I\ to\ IV}$$

Using the above formula, DC values could be plotted across the ELP score band or values creating a graphic view of the change in decision consistency. Steps in creating the graph would be as follows:

1. Determine a protocol for delineating ELP levels/score bands (e.g., how many bands, how to subdivide them) with a separate band established for state's current ELP performance standard.
2. Select grades and/or grade bands to compare.

Calculate a DC score for each ELP level or band and plot.

If the hypothesis holds regarding the decreasing relationship between English language proficiency and academic content performance after a certain ELP level is reached, then there should be an observable

decrease in the DC values on this graph. The point at which this decrease is observed is suggestive of where the ELP performance standard might be considered. Typically, this relationship varies by subject area and grade or grade span; so the method is expected to give not a single, definitive result but rather a performance range for consideration. An example follows.

Assume the information in Exhibit A.2 represents ELP and (ELA) assessment information from a state. This fictitious state has five ELP levels: Entering, Beginning, Intermediate, Advanced, and Bridging. To calculate decision consistency, the study team will follow the steps mentioned earlier. First, the five ELP levels are subdivided into ten bands. Each proficiency level band has a "low" or "high" designator. In this state, each composite proficiency level has a scale score range. The scale score point at the midpoint of each proficiency level range is then demarcated. Students below the midpoint are classified as "low." Students at or above that point, are classified as "high." The first column in Exhibit A.2 shows the ten composite ELP bands created. The next two columns present the numbers of students "Not Proficient" or "Proficient" on the state's ELA assessment for each language proficiency level band. For example in the Advanced High band, 331 students were not proficient on the state's ELA assessment, and 178 were proficient. The last column lists the decision consistency percentage.

Exhibit A.2.
Example Decision Consistency Table, Grade 5 English or Language Arts

Composite ELP Level Bands	English Language Arts Assessment		Decision Consistency Percentage
	Number Not Proficient	Number Proficient	
Entering Low	4	0*	37%
Entering High	44	2*	37%
Beginning Low	61	0*	39%
Beginning High	114	6*	41%
Intermediate Low	124	13*	46%
Intermediate High	327	57*	51%
Advanced Low	245	81*	63%
Advanced High	331*	178	70%
Bridging Low	134*	249	77%
Bridging High	64*	257	72%
Total	1,448	843	

Notes: Total n = 2,291. Data are hypothetical.
* Shading in the two English or Language Arts assessment columns illustrates values used to calculate the decision consistency percentage for Advanced High level.

Next, the decision consistency percentage is calculated for each proficiency level band. Keeping with the Advanced High as an example, students below the Advanced High band and not proficient on the ELA assessment are in quadrant III (i.e., classified below proficient on both assessments). Students at or above Advanced High and proficient on the ELA assessment are in quadrant II (classified above proficient on both assessments). Neither group is shaded in Exhibit A.2, consistent with Exhibit A.1. Note that the decision consistency approach is designed to support decisions about "language

proficiency." Claiming students at or above Advanced High are proficient is somewhat of a misnomer. It should more appropriately be termed, "if language proficiency were set at Advanced High or higher." Decision consistency at this point is 70 percent and is calculated as follows:

$$70\% = \frac{(4 + 44 + 61 + 114 + 124 + 327 + 245) + (178 + 249 + 257)}{2{,}291}.$$

That is, if Advanced High were the language proficiency level, 70 percent of students would be "consistently classified." Students in shaded areas would be classified inconsistently. Plotting decision consistency percentages for all bands yields the following graph (see Exhibit A.3).

Exhibit A.3.
Example ELP and English or Language Arts Decision Consistency Graph

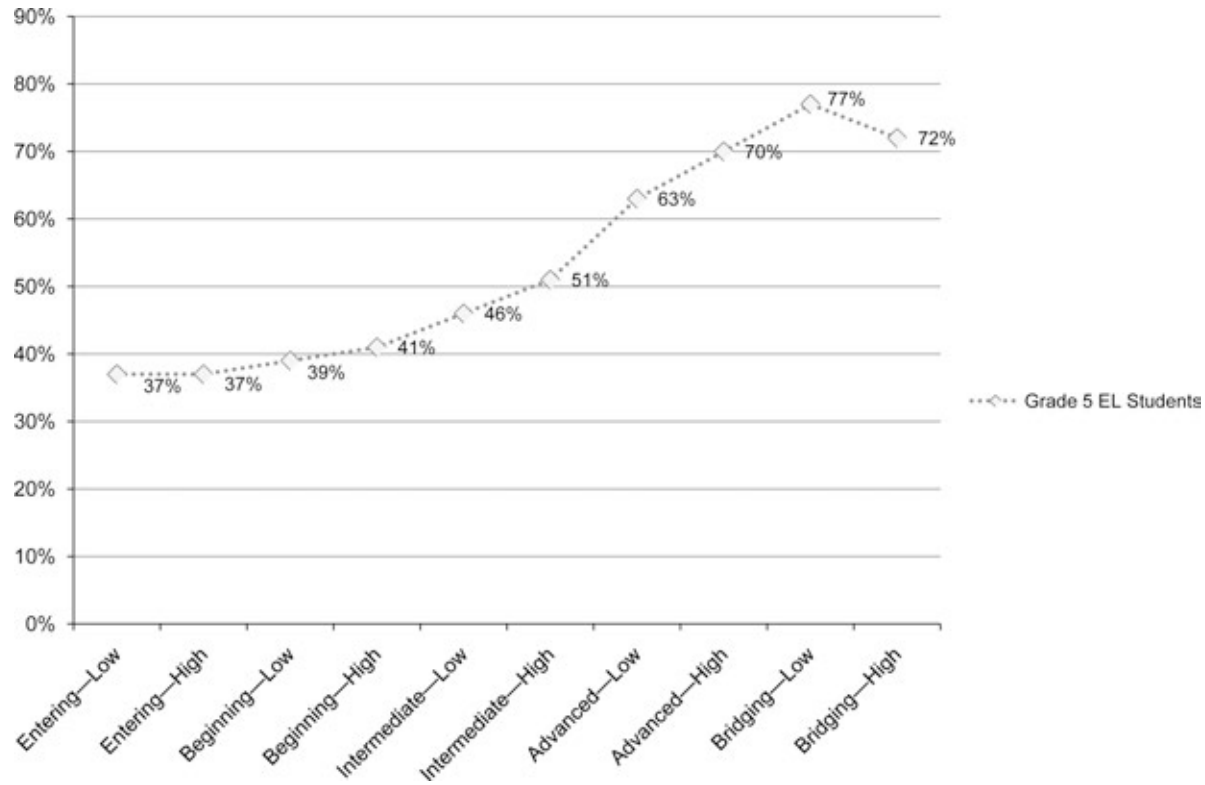


Exhibit reads: The cumulative rate of consistent decisions obtained for fifth-grade EL students in English language arts up through the bridging low ELP performance level is 77 percent.

Note: Exhibit A.2 presents the number of students with ELP and English language arts proficiency level scores and the percentage of consistent decisions.

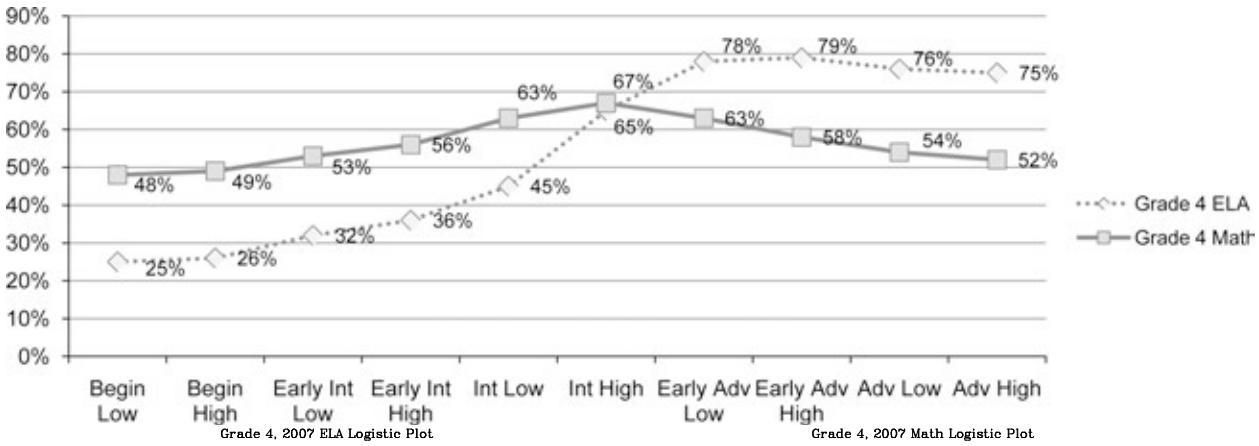
Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Note that between the Bridging Low and Bridging High bands, the values decrease. That is, where the English language proficient performance standard set at the Bridging High band, fewer consistent decisions would be made relative to the ELA assessment, than at the Bridging Low band. The band where decision consistency is at its highest is where deliberations about English-language proficiency should begin.

APPENDIX B:
EDUCATION AGENCY 1

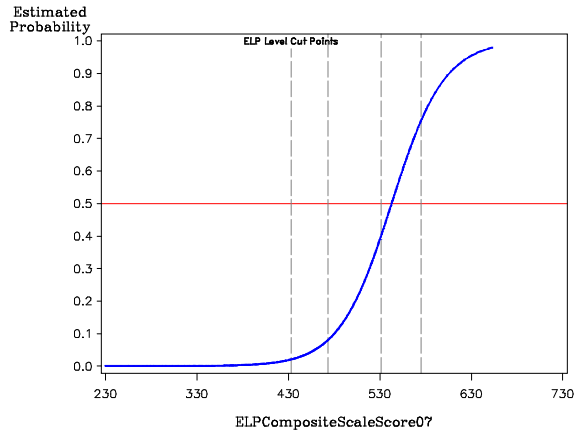
Appendix B. Education Agency 1

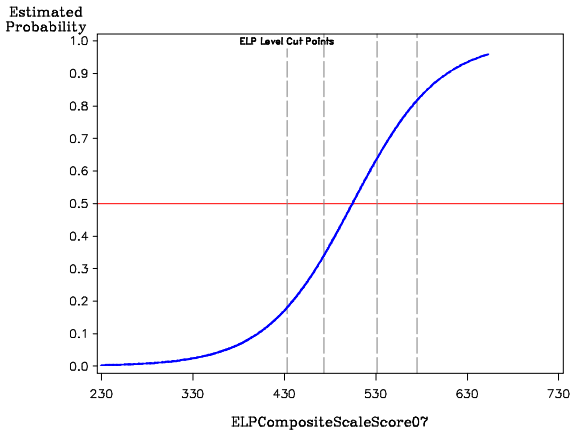
Exhibit B.1.
Education Agency 1, Grade 4: Decision Consistency Analysis, Logistic Plot, and Box Plot (2006–07)

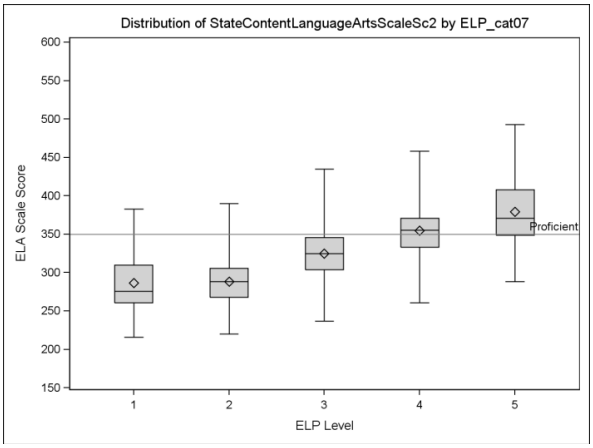


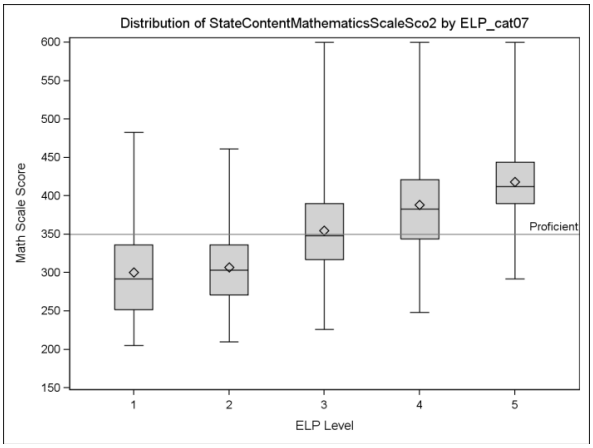
Grade 4, 2007 ELA Logistic Plot

Grade 4, 2007 Math Logistic Plot





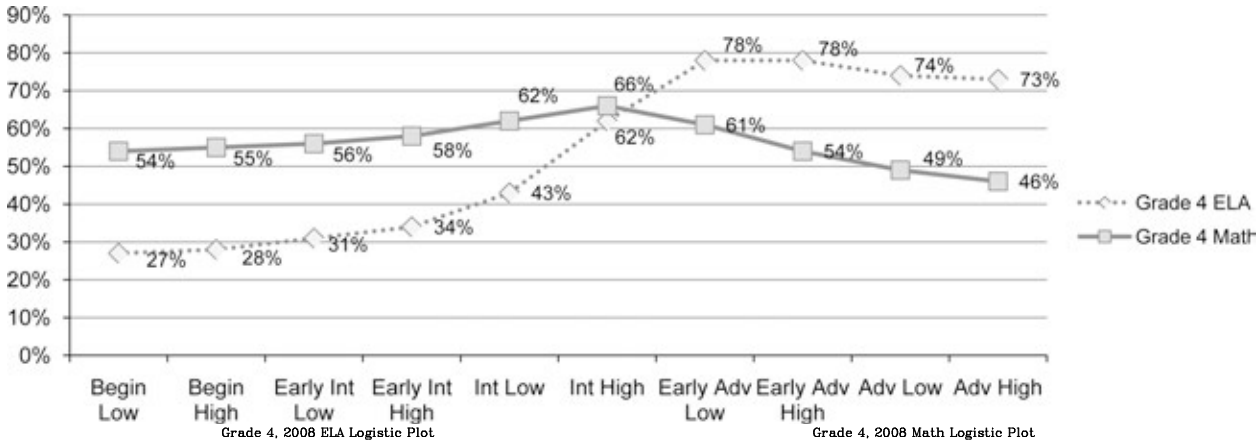




Note: The corresponding data tables for Exhibit B.1 are Exhibits B.3 and B.4 (for the line graphs), Exhibit B.7 (for the logistic plots), and Exhibit B.9 (for the box plots).

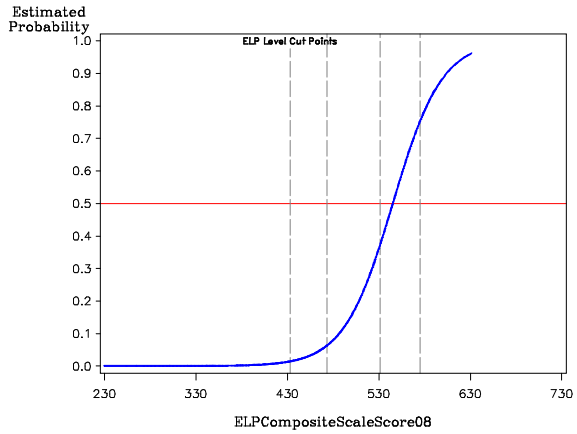
Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

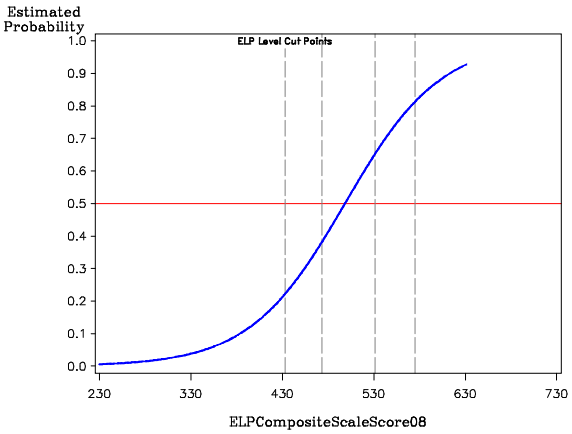
Exhibit B.2.
Education Agency 1, Grade 4: Decision Consistency Analysis, Logistic Plot, and Box Plot (2007–08)

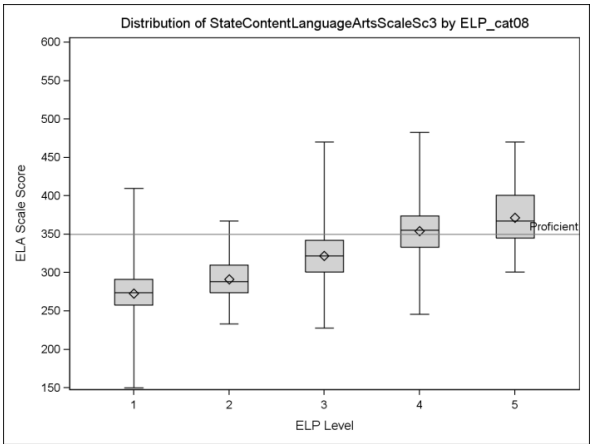


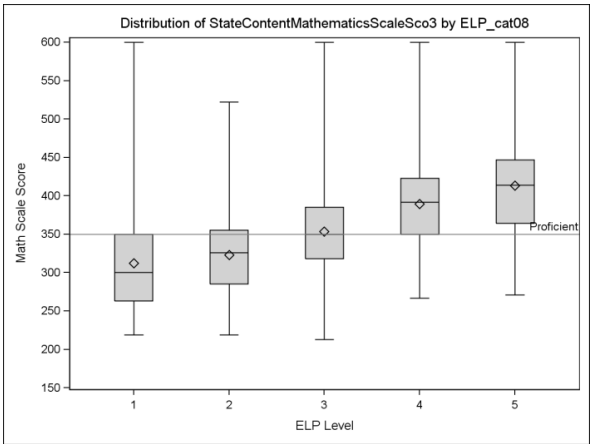
Grade 4, 2008 ELA Logistic Plot

Grade 4, 2008 Math Logistic Plot









Note: The corresponding data tables for Exhibit B.2 are Exhibits B.5 and B.6 (for the line graphs), Exhibit B.8 (for the logistic plots), and Exhibit B.10 (for the box plots).

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit B.3.
Education Agency 1, Grade 4: ELP and English or Language Arts
Decision Consistency Analysis (2006–07)

ELP Level	Number of Students With ELA Scores		Percent of Consistent Decisions
	Not Proficient ELA	Proficient ELA	
Beginning Low	31	0	25%
Beginning High	128	10	26%
Early Intermediate Low	84	0	32%
Early Intermediate High	211	5	36%
Intermediate Low	486	60	45%
Intermediate High	487	200	65%
Early Advanced Low	128	122	78%
Early Advanced High	48	96	79%
Advanced High	14	34	76%
Advanced Low	0	2	75%
N (%)	1,617 (75%)	529 (25%)	

Note: Total n = 2,146

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit B.4.
Education Agency 1, Grade 4: ELP and Mathematics
Decision Consistency Analysis (2006–07)

ELP Level	Number of Students With Math Scores		Percent of Consistent Decisions
	Not Proficient Math	Proficient Math	
Beginning Low	27	4	48%
Beginning High	110	28	49%
Early Intermediate Low	78	6	53%
Early Intermediate High	178	38	56%
Intermediate Low	318	226	63%
Intermediate High	297	389	67%
Early Advanced Low	79	171	63%
Early Advanced High	28	116	58%
Advanced High	5	43	54%
Advanced Low	0	2	52%
N (%)	1,120 (52%)	1,023 (48%)	

Note: Total n = 2,143

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit B.5.
Education Agency 1, Grade 4: ELP and English or Language Arts
Decision Consistency Analysis (2007–08)

ELP Level	Number of Students With ELA Scores		Percent of Consistent Decisions
	Not Proficient ELA	Proficient ELA	
Beginning Low	17	1	27%
Beginning High	52	1	28%
Early Intermediate Low	61	0	31%
Early Intermediate High	153	5	34%
Intermediate Low	372	37	43%
Intermediate High	418	142	62%
Early Advanced Low	140	132	78%
Early Advanced High	42	108	78%
Advanced High	26	47	74%
Advanced Low	0	0	73%
N (%)	1,281 (73%)	473 (27%)	

Note: Total n = 1,754

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit B.6.
Education Agency 1, Grade 4: ELP and Mathematics
Decision Consistency Analysis (2007–08)

ELP Level	Number of Students With Math Scores		Percent of Consistent Decisions
	Not Proficient Math	Proficient Math	
Beginning Low	12	4	54%
Beginning High	39	15	55%
Early Intermediate Low	49	13	56%
Early Intermediate High	111	49	58%
Intermediate Low	248	163	62%
Intermediate High	239	327	66%
Early Advanced Low	74	198	61%
Early Advanced High	24	128	54%
Advanced High	12	61	49%
Advanced Low	0	0	46%
N (%)	808 (46%)	958 (54%)	

Note: Total n = 1,766

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit B.7.
Education Agency 1, Grade 4: Logistic Regression Results on English or Language Arts and Mathematics Proficiency (2006–07)

Subject of Content Assessment	Parameter	Estimate	Standard Error	Wald $X^2(1)$	Pr > X^2
ELA	Intercept	–18.98	1.05	329.93	<.0001
	ELP Reading Scale Score	0.04	0.00	301.81	<.0001
	N	2,146			
Math	Intercept	–10.69	0.69	236.95	<.0001
	ELP Reading Scale Score	0.02	0.00	236.98	<.0001
	N	2,143			

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit B.8.
Education Agency 1, Grade 4: Logistic Regression Results on English or Language Arts and Mathematics Proficiency (2007–08)

Subject of Content Assessment	Parameter	Estimate	Standard Error	Wald $X^2(1)$	Pr > X^2
ELA	Intercept	–20.38	1.17	303.14	<.0001
	ELP Reading Scale Score	0.04	0.00	281.93	<.0001
	N	1,754			
Math	Intercept	–9.54	0.75	160.04	<.0001
	ELP Reading Scale Score	0.02	0.00	167.02	<.0001
	N	1,766			

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit B.9.
Education Agency 1, Grade 4: Descriptive Statistics Box Plot Analysis (2006–07)

Subject of Content Assessment	ELP Level	Mean	Standard Deviation	Minimum	Maximum	First Quartile	Median	Third Quartile
ELA	Beginning	286	36	216	383	261	275	310
	Early Int.	288	27	220	390	268	287	304
	Intermediate	325	32	237	435	304	325	346
	Early Adv.	355	32	261	458	333	355	371
	Advanced	378	45	288	493	349	371	403
Math	Beginning	300	56	205	483	252	292	332
	Early Int.	306	46	210	461	271	303	336
	Intermediate	355	55	226	600	321	351	390
	Early Adv.	389	60	248	600	344	383	421
	Advanced	418	58	292	600	390	412	444

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit B.10.
Education Agency 1, Grade 4: Descriptive Statistics Box Plot Analysis (2007–08)

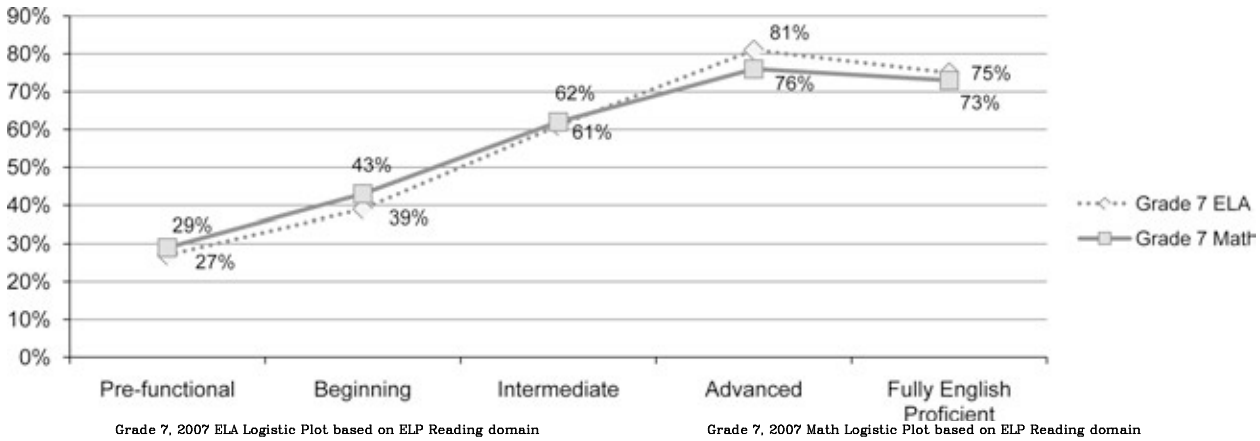
Subject of Content Assessment	ELP Level	Mean	Standard Deviation	Minimum	Maximum	First Quartile	Median	Third Quartile
ELA	Beginning	273	37	150	410	258	275	292
	Early Int.	291	26	233	367	274	288	310
	Intermediate	322	32	228	470	298	322	342
	Early Adv.	354	33	246	483	333	355	374
	Advanced	370	36	301	470	345	364	397
Math	Beginning	315	68	219	600	263	300	350
	Early Int.	324	49	219	522	289	326	355
	Intermediate	353	55	213	600	318	350	385
	Early Adv.	389	57	267	600	350	389	414
	Advanced	412	62	271	600	364	414	447

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

APPENDIX C:
EDUCATION AGENCY 2

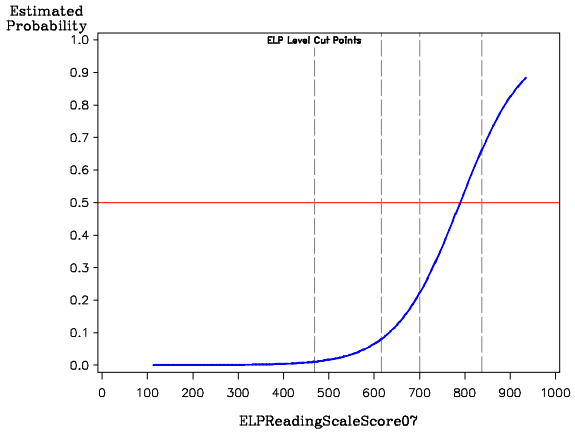
Appendix C. Education Agency 2

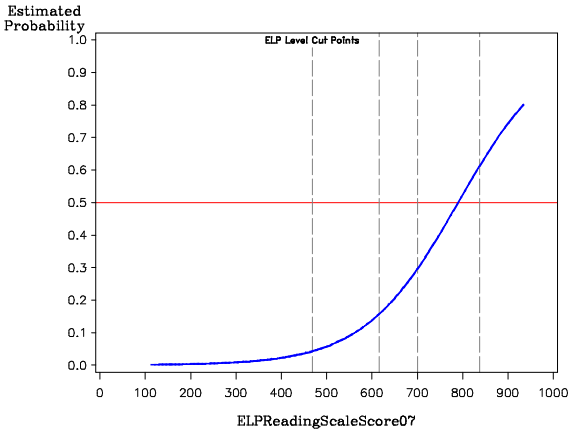
Exhibit C.1.
Education Agency 2, Grade 7: Decision Consistency Analysis, Logistic Plot, and Box Plot (2006–07)

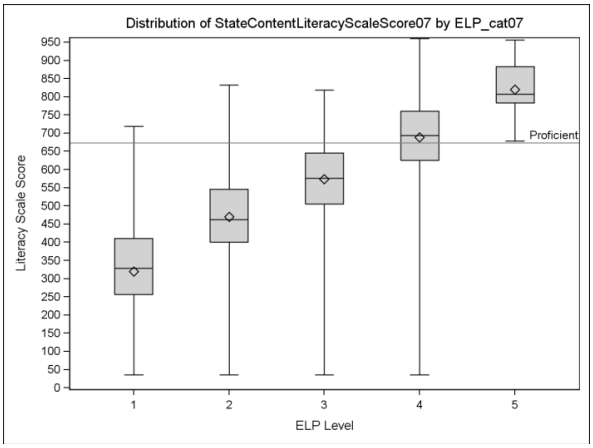


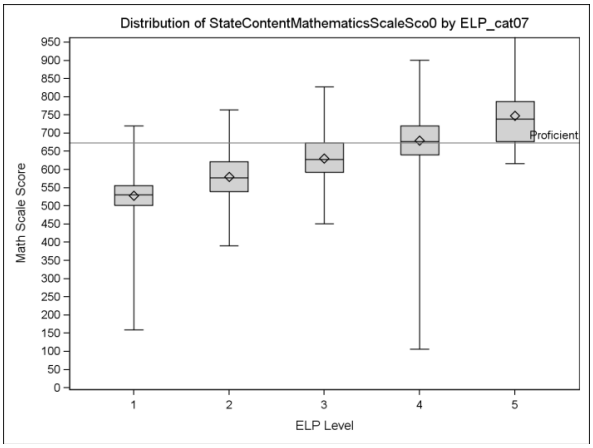
Grade 7, 2007 ELA Logistic Plot based on ELP Reading domain

Grade 7, 2007 Math Logistic Plot based on ELP Reading domain





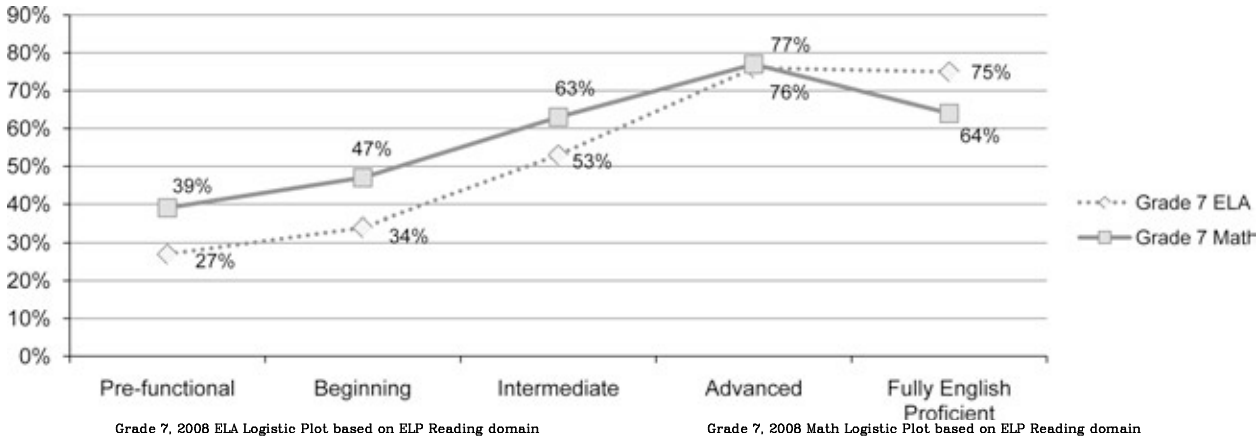




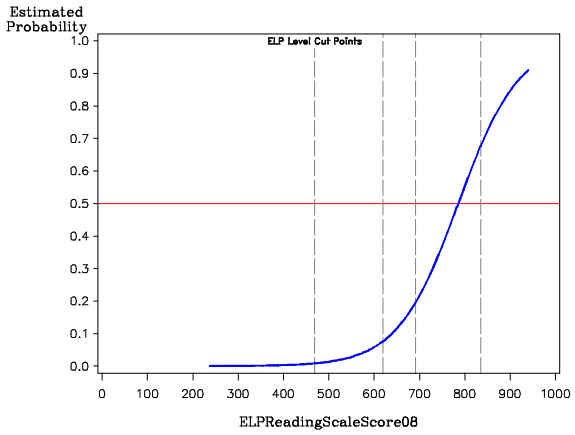
Note: The corresponding data tables for Exhibit C.1 are Exhibits C.3 and C.4 (for the line graphs), Exhibit C.7 (for the logistic plots), and Exhibit C.9 (for the box plots).

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

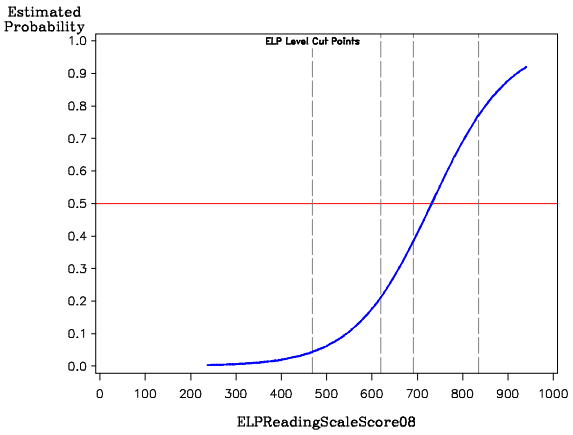
Exhibit C.2.
Education Agency 2, Grade 7: Decision Consistency Analysis, Logistic Plot, and Box Plot (2007–08)

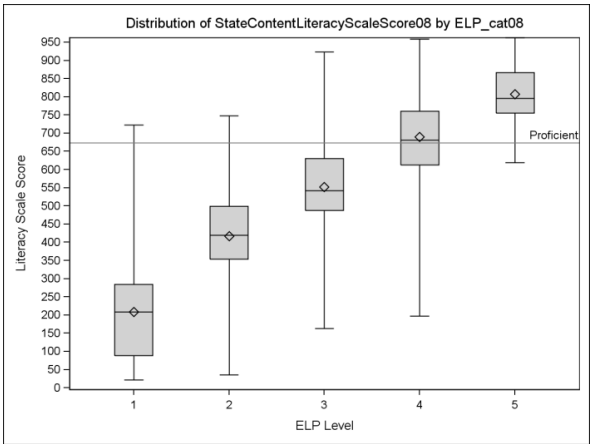


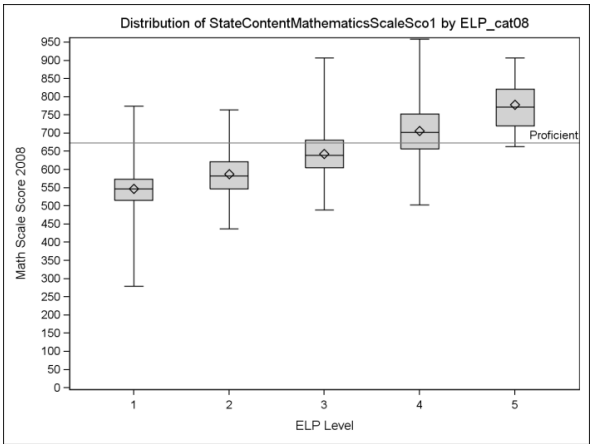
Grade 7, 2008 ELA Logistic Plot based on ELP Reading domain



Grade 7, 2008 Math Logistic Plot based on ELP Reading domain







Note: The corresponding data tables for Exhibit C.2 are Exhibits C.5 and C.6 (for the line graphs), Exhibit C.8 (for the logistic plots); and Exhibit C.10 (for the box plots).

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit C.3.
Education Agency 2, Grade 7: ELP and English or Language Arts
Decision Consistency Analysis (2006–07)

ELP Level	Number of Students With ELA Scores		Percent of Consistent Decisions
	Not Proficient ELA	Proficient ELA	
Pre-functional	154	1	27%
Beginning	291	8	39%
Intermediate	319	63	61%
Advanced	171	242	81%
Fully English Proficient	0	30	75%
N (%)	935 (73%)	344 (27%)	

Note: Total n = 1,279

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit C.4.
Education Agency 2, Grade 7: ELP and Mathematics
Decision Consistency Analysis (2006–07)

ELP Level	Number of Students With Math Scores		Percent of Consistent Decisions
	Not Proficient Math	Proficient Math	
Pre-functional	181	1	29%
Beginning	280	23	43%
Intermediate	283	99	62%
Advanced	184	230	76%
Fully English Proficient	5	25	73%
N (%)	933 (71%)	378 (29%)	

Note: Total n = 1,311

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit C.5.
Education Agency 2, Grade 7: ELP and English or Language Arts
Decision Consistency Analysis (2007–08)

ELP Level	Number of Students With ELA Scores		Percent of Consistent Decisions
	Not Proficient ELA	Proficient ELA	
Pre-functional	109	1	27%
Beginning	283	3	34%
Intermediate	410	59	53%
Advanced	277	300	77%
Fully English Proficient	3	35	75%
N (%)	1,082 (73%)	398 (27%)	

Note: Total n = 1,480

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit C.6.
Education Agency 2, Grade 7: ELP and Mathematics
Decision Consistency Analysis (2007–08)

ELP Level	Number of Students With Math Scores		Percent of Consistent Decisions
	Not Proficient Math	Proficient Math	
Pre-functional	125	4	39%
Beginning	267	26	47%
Intermediate	335	135	63%
Advanced	198	380	76%
Fully English Proficient	1	37	64%
N (%)	926 (61%)	582 (39%)	

Note: Total n = 1,508

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit C.7.
Education Agency 2, Grade 7: Logistic Regression Results on English or Language Arts and Mathematics Proficiency (2006–07)

Subject of Content Assessment	Parameter	Estimate	Standard Error	Wald $X^2(1)$	Pr > X^2
ELA	Intercept	–11.08	0.64	299.52	<.0001
	ELP Reading Scale Score	0.01	0.00	277.86	<.0001
	N	1,288			
Math	Intercept	–7.61	0.45	291.55	<.0001
	ELP Reading Scale Score	0.01	0.00	255.37	<.0001
	N	1,321			

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit C.8.
Education Agency 2, Grade 7: Logistic Regression Results on English or Language Arts and Mathematics Proficiency (2007–08)

Subject of Content Assessment	Parameter	Estimate	Standard Error	Wald $X^2(1)$	Pr > X^2
ELA	Intercept	–11.77	0.64	335.50	<.0001
	ELP Reading Scale Score	0.02	0.00	303.88	<.0001
	N	1,484			
Math	Intercept	–8.56	0.47	330.45	<.0001
	ELP Reading Scale Score	0.01	0.00	312.57	<.0001
	N	1,512			

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit C.9.
Education Agency 2, Grade 7: Descriptive Statistics Box Plot Analysis (2006–07)

Subject of Content Assessment	ELP Level	Mean	Standard Deviation	Minimum	Maximum	First Quartile	Median	Third Quartile
ELA	Pre-functional	319	132	35	719	256	328	411
	Beginning	470	106	35	832	400	463	546
	Intermediate	573	102	35	819	505	576	645
	Advanced	689	104	35	960	625	694	761
	Fully Eng. Prof.	820	81	679	956	783	807	883
Math	Pre-functional	529	59	159	720	502	531	556
	Beginning	580	62	391	764	540	578	622
	Intermediate	631	62	451	827	592	628	673
	Advanced	680	70	106	901	640	677	720
	Fully Eng. Prof.	748	89	616	969	677	738.5	787

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit C.10.
Education Agency 2, Grade 7: Descriptive Statistics Box Plot Analysis (2007–08)

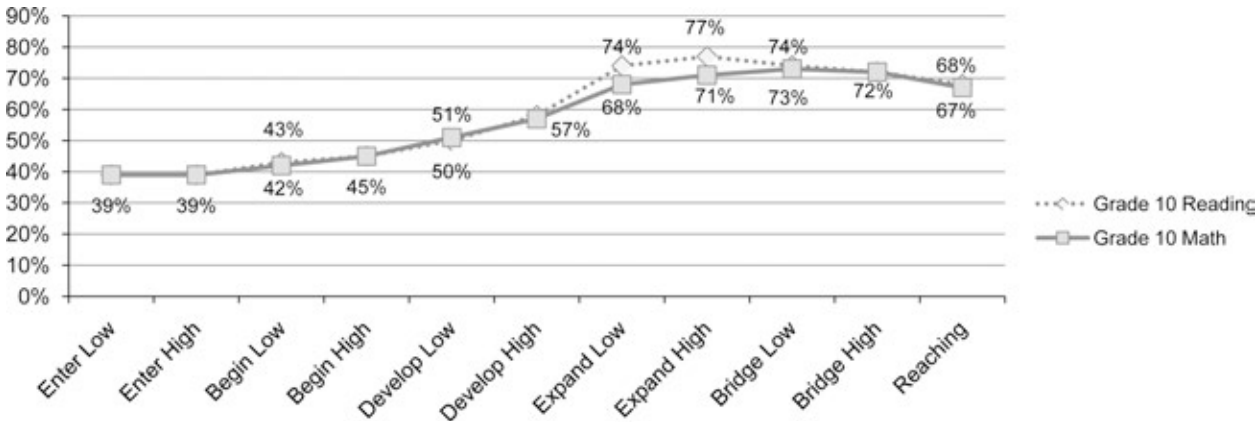
Subject of Content Assessment	ELP Level	Mean	Standard Deviation	Minimum	Maximum	First Quartile	Median	Third Quartile
ELA	Pre-functional	209	135	22	723	89	208.5	284
	Beginning	416	125	36	748	354	419	499
	Intermediate	552	108	163	923	488	542	630
	Advanced	690	112	197	959	613	681	761
	Fully Eng. Prof.	808	92	619	962	755	796	867
Math	Pre-functional	548	60	279	775	515	547	574
	Beginning	587	57	437	764	547	583	622
	Intermediate	643	59	489	907	605	639	681
	Advanced	706	66	503	959	657	703	753
	Fully Eng. Prof.	779	65	663	907	720	772	821

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

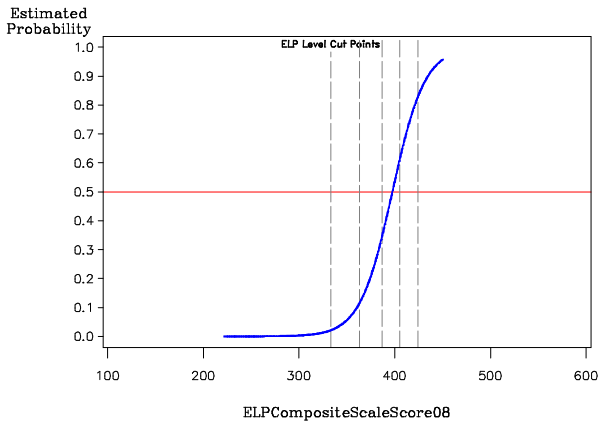
APPENDIX D:
EDUCATION AGENCY 3

Appendix D. Education Agency 3

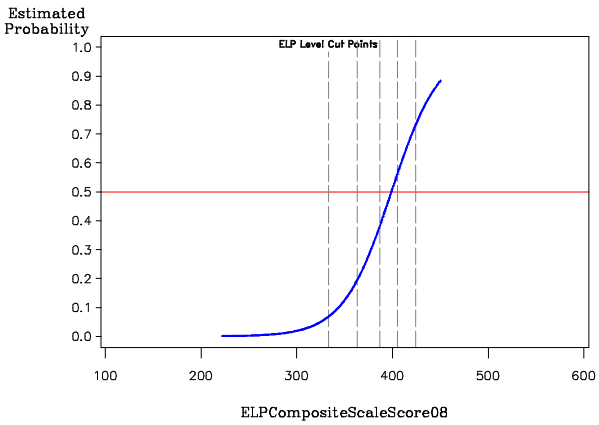
Exhibit D.1.
Education Agency 3, Grade 10: Decision Consistency Analysis, Logistic Plot, and Box Plot (2008–09)

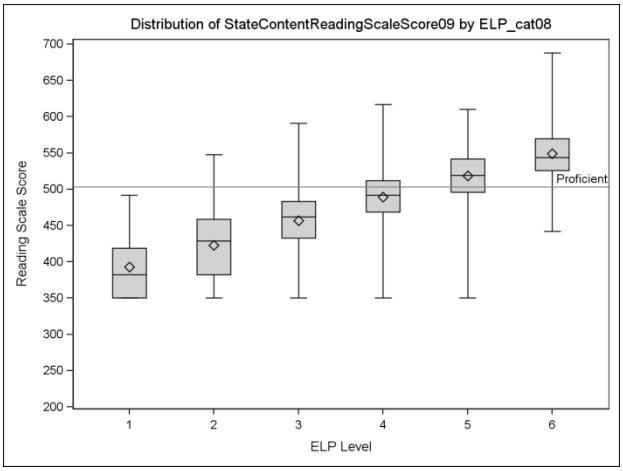


Grade 10, 2009 Reading Proficient Logistic Plot



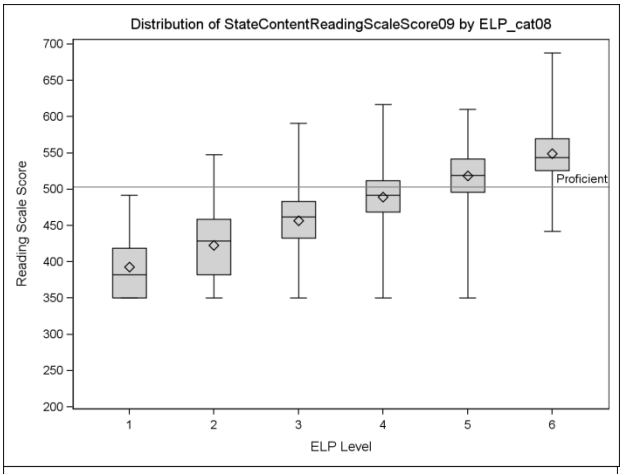
Grade 10, 2009 Mathematics Proficient Logistic Plot

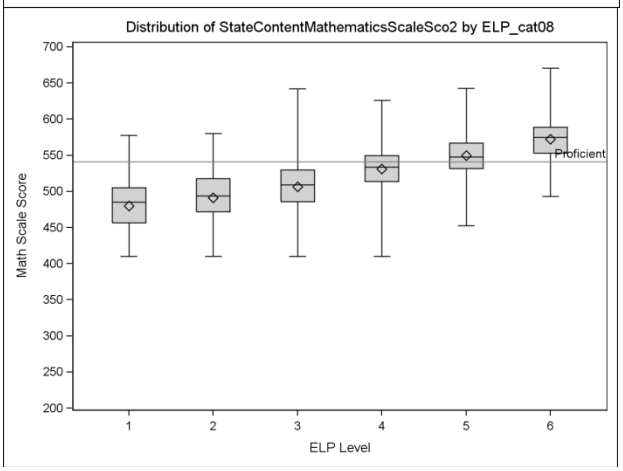




continued next page

Exhibit D.1. (continued)
Education Agency 3, Grade 10: Decision Consistency Analysis, Logistic Plot, and Box Plot (2008–09)

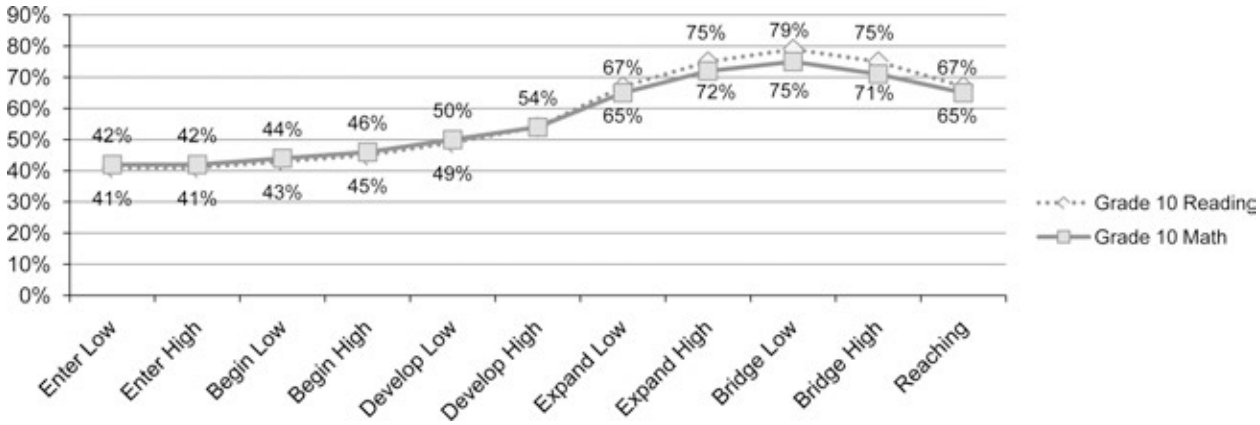




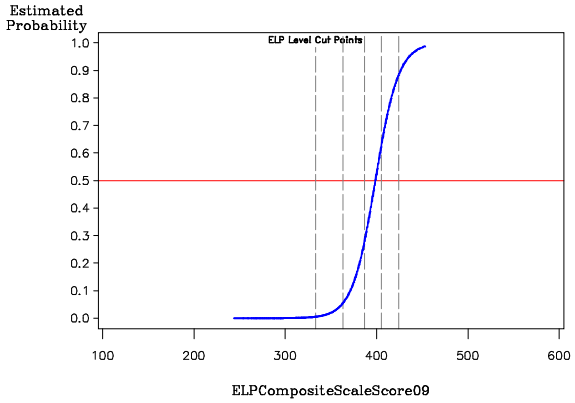
Note: The corresponding data tables for Exhibit D.1 are Exhibits D.3 and D.4 (for the line graphs), Exhibit D.7 (for the logistic plots), and Exhibit D.9 (for the box plots).

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

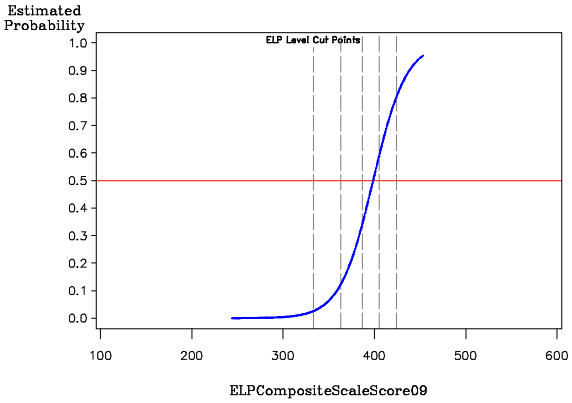
Exhibit D.2.
Education Agency 3, Grade 10: Decision Consistency Analysis, Logistic Plot, and Box Plot (2009–10)

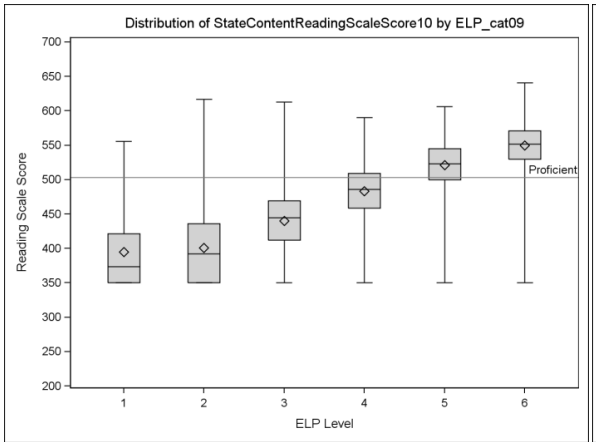


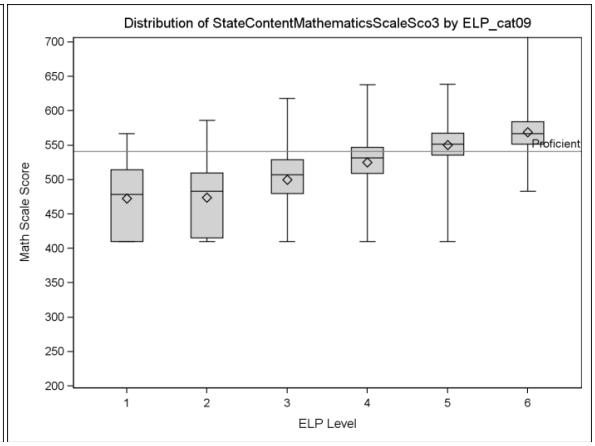
Grade 10, 2010 Reading Proficient Logistic Plot



Grade 10, 2010 Mathematics Proficient Logistic Plot







Note: The corresponding data tables for Exhibit D.2 are Exhibits D.5 and D.6 (for the line graphs), Exhibit D.8 (for the logistic plots), and Exhibit D.10 (for the box plots).

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit D.3.
Education Agency 3, Grade 10: ELP and Reading
Decision Consistency Analysis (2008–09)

ELP Level	Number of Students With Reading Scores		Percent of Consistent Decisions
	Not Proficient Reading	Proficient Reading	
Entering Low	3	0	39%
Entering High	69	0	39%
Beginning Low	51	2	42%
Beginning High	128	8	45%
Developing Low	179	12	50%
Developing High	309	47	58%
Expanding Low	177	80	70%
Expanding High	242	173	74%
Bridging Low	112	188	77%
Bridging High	58	185	74%
Reaching	16	173	68%
N (%)	1,344 (61%)	868 (39%)	

Note: Total n = 2,212

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit D.4.
Education Agency 3, Grade 10: ELP and Mathematics
Decision Consistency Analysis (2008–09)

ELP Level	Number of Students With Math Scores		Percent of Consistent Decisions
	Not Proficient Math	Proficient Math	
Entering Low	4	0	39%
Entering High	99	7	39%
Beginning Low	65	9	43%
Beginning High	152	19	45%
Developing Low	169	21	51%
Developing High	300	63	57%
Expanding Low	167	92	68%
Expanding High	232	182	71%
Bridging Low	136	164	73%
Bridging High	66	180	72%
Reaching	30	159	67%
N (%)	1,420 (61%)	896 (39%)	

Note: Total n = 2,316

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit D.5.
Education Agency 3, Grade 10: ELP and Reading
Decision Consistency Analysis (2009–10)

ELP Level	Number of Students With Reading Scores		Percent of Consistent Decisions
	Not Proficient Reading	Proficient Reading	
Entering Low	3	2	41%
Entering High	45	2	41%
Beginning Low	62	0	43%
Beginning High	115	6	45%
Developing Low	132	6	49%
Developing High	356	29	54%
Expanding Low	264	63	67%
Expanding High	307	203	75%
Bridging Low	137	247	79%
Bridging High	62	260	75%
Reaching	17	218	67%
N (%)	1,500 (59%)	1,036 (41%)	

Note: Total n = 2,536

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit D.6.
Education Agency 3, Grade 10: ELP and Mathematics
Decision Consistency Analysis (2009–10)

ELP Level	Number of Students With Math Scores		Percent of Consistent Decisions
	Not Proficient Math	Proficient Math	
Entering Low	3	2	42%
Entering High	49	2	42%
Beginning Low	70	3	44%
Beginning High	119	12	46%
Developing Low	119	22	50%
Developing High	331	53	54%
Expanding Low	247	80	65%
Expanding High	302	208	72%
Bridging Low	141	244	75%
Bridging High	86	236	71%
Reaching	27	207	65%
N (%)	1,494 (58%)	1,069 (42%)	

Note: Total n = 2,563

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit D.7.
Education Agency 3, Grade 10: Logistic Regression Results on Reading and Mathematics Proficiency (2008–09)

Subject of Content Assessment	Parameter	Estimate	Standard Error	Wald $X^2(1)$	Pr > X^2
Reading	Intercept	–23.54	1.15	416.25	<.0001
	ELP Composite Scale Score	0.06	0.00	408.69	<.0001
	N	2,236			
Math	Intercept	–15.81	0.87	332.57	<.0001
	ELP Composite Scale Score	0.04	0.00	320.92	<.0001
	N	2,340			

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit D.8.
Education Agency 3, Grade 10: Logistic Regression Results on Reading and Mathematics Proficiency (2009–10)

Subject of Content Assessment	Parameter	Estimate	Standard Error	Wald $X^2(1)$	Pr > X^2
Reading	Intercept	–31.78	1.38	530.09	<.0001
	ELP Composite Scale Score	0.08	0.00	525.14	<.0001
	N	2,551			
Math	Intercept	–22.05	1.06	429.24	<.0001
	ELP Composite Scale Score	0.05	0.00	422.85	<.0001
	N	2,579			

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit D.9.
Education Agency 3, Grade 10: Descriptive Statistics Box Plot Analysis (2008–09)

Subject of Content Assessment	ELP Level	Mean	Standard Deviation	Minimum	Maximum	First Quartile	Median	Third Quartile
Reading	Entering	393	42	350	492	350	383	419
	Beginning	423	49	350	548	382	429	459
	Developing	456	42	350	591	433	462	483
	Expanding	489	38	350	617	469	492	512
	Bridging	518	34	350	610	496	519	542
	Reaching	549	37	442	688	526	544	570
Math	Entering	480	40	410	578	457	485	505
	Beginning	491	41	410	580	472	494	518
	Developing	507	37	410	642	486	509	530
	Expanding	531	31	410	626	514	534	550
	Bridging	550	26	453	643	532	548	567
	Reaching	572	30	493	671	553	575	589

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit D.10.
Education Agency 3, Grade 10: Descriptive Statistics Box Plot Analysis (2009–10)

Subject of Content Assessment	ELP Level	Mean	Standard Deviation	Minimum	Maximum	First Quartile	Median	Third Quartile
Reading	Entering	395	55	350	556	350	374	422
	Beginning	401	50	350	617	350	392	436
	Developing	440	46	350	613	412	445	469
	Expanding	483	41	350	590	459	486	509
	Bridging	521	35	350	606	500	523	545
	Reaching	550	35	350	641	530	552	571
Math	Entering	473	48	410	567	410	479	515
	Beginning	474	48	410	586	416	483	510
	Developing	500	43	410	618	480	507	529
	Expanding	525	35	410	638	509	532	547
	Bridging	551	28	410	639	536	552	568
	Reaching	569	28	483	750	552	567	584

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

APPENDIX E:
EVENT HISTORY ANALYSIS

Appendix E. Event History Analysis

The survival function ($\hat{S}(t)$) is calculated as follow:

$$\hat{S}(t) = \prod_{i|t_i \leq t} \frac{n_i - d_i}{n_i}$$

Where n_i is the number of EL students that have the potential of becoming proficient at t_i minus the number of ELs becoming proficient and number of censored students. Censoring is discussed further below. d_i is the number of ELs who become proficient at time t_i. The product, $\hat{S}(t)$, is the overall probability of not becoming proficient at time t or less. In this particular of study, the focus is on the probability of becoming proficient at time t; therefore, we subtracted $\hat{S}(t)$ from 1.

Exhibit E.1 presents the number and probability of ELs identified during 2003–04 of becoming proficient in Grades K–2 and Initial ELP Level 1. For example, in Year 0 for grades K–2 with an ELP level 1, in order to obtain the probability of becoming proficient, the following steps were used. First, the probability of not becoming proficiency in Year 1 was obtained using $\hat{S}(t=1)$ or (7,728–809)/7,728, which equals 0.895. Then, the total number of students for Year 1 is determined by subtracting the number of student becoming proficient and the number of students censored from the total number of students in year 0 (7,728–809–0), which equals 6,919. The probability of not becoming proficient in year 1 was obtained using $\hat{S}(t=2)$ or (6,919–784)/6,919, which equals 0.886. These two numbers were then multiplied together to obtain the estimated probability of not becoming proficient through both years 0 and 1, which equals 0.79. In order to determine the probability of becoming proficient through year 1, this number (0.79) is subtracted from 1, giving rise to 0.21.

Exhibit E.1.
Number and Probability of ELs Identified During 2003–04 Becoming Proficient, in Grades K–2 and Initial ELP Level 1, Education Agency 1

Years	Total Number of Students	Number of Students Becoming Proficient	Number of Students Censored	Probability of Becoming Proficient
0	7,728	809	0	0.10
1	6,919	784	464	0.21
2	5,671	430	416	0.27
3	4,825	837	342	0.39
4	3,646	0	3,646	0.39

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit E.2.
Censored Adjustment 1—Underestimate
Number and Probability of ELs Identified During 2003–04 Becoming Proficient, in Grades K–5, by ELP Level, Education Agency 1

Years	Total Number of Students	Number of Students Becoming Proficient	Number of Students Censored	Probability of Becoming Proficient	Standard Error	95% Confidence Interval	
Grades K–2 ELP Level 1							
1	7,728	809	0	0.10	0.00	0.10	0.11
2	6,919	784	464	0.21	0.00	0.20	0.22
3	5,671	430	416	0.27	0.01	0.26	0.28
4	4,825	837	342	0.39	0.01	0.38	0.41
5	3,646	0	3,646	0.39	0.01	0.38	0.41
Grades K–2 ELP Level 2							
1	7,603	2,184	0	0.29	0.01	0.28	0.30
2	5,419	1,070	420	0.43	0.01	0.42	0.44
3	3,929	382	281	0.48	0.01	0.47	0.50
4	3,266	757	231	0.60	0.01	0.59	0.61
5	2,278	0	2,278	0.60	0.01	0.59	0.61
Grades K–2 ELP Level 3							
1	10,045	5,271	0	0.52	0.01	0.52	0.53
2	4,774	1,618	329	0.69	0.00	0.68	0.69
3	2,827	464	226	0.74	0.00	0.73	0.75
4	2,137	668	162	0.82	0.00	0.81	0.83
5	1,307	0	1,307	0.82	0.00	0.81	0.83
Grades 3–5 ELP Level 1							
1	1,043	60	0	0.06	0.01	0.05	0.07
2	983	91	110	0.14	0.01	0.12	0.17
3	782	112	139	0.27	0.01	0.24	0.30
4	531	96	80	0.40	0.02	0.37	0.43
5	355	0	355	0.40	0.02	0.37	0.43
Grades 3–5 ELP Level 2							
1	235	75	0	0.32	0.03	0.26	0.38
2	160	18	23	0.40	0.03	0.34	0.46
3	119	26	29	0.53	0.03	0.46	0.60
4	64	14	15	0.63	0.04	0.56	0.70
5	35	0	35	0.63	0.04	0.56	0.70
Grades 3–5 ELP Level 3							
1	335	215	0	0.64	0.03	0.59	0.69
2	120	23	19	0.71	0.02	0.66	0.76
3	78	14	25	0.76	0.02	0.71	0.81
4	39	11	10	0.83	0.02	0.78	0.87
5	18	0	18	0.83	0.02	0.78	0.87

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit E.3.
Censored Adjustment 2—Overestimate
Number and Probability of ELs Identified During 2003–04 Becoming Proficient, in Grades K–2, by ELP Level, Education Agency 1

Years	Total Number of Students	Number of Students Becoming Proficient	Number of Students Censored	Probability of Becoming Proficient	Standard Error	95% Confidence Interval	
Grades K–2 ELP Level 1							
1	7,728	809	0	0.10	0.00	0.10	0.11
2	6,919	784	0	0.21	0.00	0.20	0.22
3	6,135	430	0	0.26	0.01	0.25	0.27
4	5,705	837	0	0.37	0.01	0.36	0.38
5	4,868	0	0	0.37	0.01	0.36	0.38
6	4,868	0	0	0.37	0.01	0.36	0.38
7	4,868	0	4,868	0.37	0.01	0.36	0.38
Grades K–2 ELP Level 2							
1	7,603	2,184	0	0.29	0.01	0.28	0.30
2	5,419	1,070	0	0.43	0.01	0.42	0.44
3	4,349	382	0	0.48	0.01	0.47	0.49
4	3,967	757	0	0.58	0.01	0.57	0.59
5	3,210	0	0	0.58	0.01	0.57	0.59
6	3,210	0	3,210	0.58	0.01	0.57	0.59
Grades K–2 ELP Level 3							
1	10,045	5,271	0	0.52	0.01	0.52	0.53
2	4,774	1,618	0	0.69	0.00	0.68	0.69
3	3,156	464	0	0.73	0.00	0.72	0.74
4	2,692	668	0	0.80	0.00	0.79	0.81
5	2,024	0	2,024	0.80	0.00	0.79	0.81

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

Exhibit E.4.
Censored Adjustment 2—Overestimate
Number and Probability of ELs Identified During 2003–04 Becoming Proficient, in Grades 3–5, by ELP Level, Education Agency 1

Years	Total Number of Students	Number of Students Becoming Proficient	Number of Students Censored	Probability of Becoming Proficient	Standard Error	95% Confidence Interval	
Grades K–2 ELP Level 1							
1	1,043	60	0	0.06	0.01	0.05	0.07
2	983	91	0	0.14	0.01	0.12	0.17
3	892	112	0	0.25	0.01	0.23	0.28
4	780	96	0	0.34	0.01	0.32	0.37
5	684	0	0	0.34	0.01	0.32	0.37
6	684	0	0	0.34	0.01	0.32	0.37
7	684	0	684	0.34	0.01	0.32	0.37
Grades K–2 ELP Level 2							
1	235	75	0	0.32	0.03	0.26	0.38
2	160	18	0	0.40	0.03	0.34	0.46
3	142	26	0	0.51	0.03	0.44	0.57
4	116	14	0	0.57	0.03	0.50	0.63
5	102	0	0	0.57	0.03	0.50	0.63
6	102	0	102	0.57	0.03	0.50	0.63
Grades K–2 ELP Level 3							
1	335	215	0	0.64	0.03	0.59	0.69
2	120	23	0	0.71	0.02	0.66	0.76
3	97	14	0	0.75	0.02	0.71	0.80
4	83	11	0	0.79	0.02	0.74	0.83
5	72	0	72	0.79	0.02	0.74	0.83

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

APPENDIX F:
EDUCATION AGENCY 1

Appendix F. Education Agency 1

Exhibit F.1.
Education Agency 1, Grade 3: Descriptive Statistics Box Plot Analysis (2007–08)

Subject of Content Assessment	ELP Level	Mean	Standard Deviation	Minimum	Maximum	First Quartile	Median	Third Quartile
ELA	Beginning	248	31.9	166	447	229	242	266
	Early Int.	265	31.3	156	413	242	262	285
	Intermediate	299	33.3	182	464	278	300	322
	Early Adv.	333	34.3	219	487	311	334	356
	Advanced	362	40.9	202	600	334	356	390
Math	Beginning	282	64.4	150	600	236	273	310
	Early Int.	302	56.7	150	600	264	296	336
	Intermediate	349	60.9	162	600	306	346	384
	Early Adv.	398	66.6	200	600	352	391	437
	Advanced	439	72.3	211	600	391	437	486

Source: National Evaluation of Title III Implementation student-level longitudinal achievement data sets.

The Department of Education's mission is to promote student achievement and preparation for global competitiveness by fostering educational excellence and ensuring equal access.

www.ed.gov