edCount LLC
because all students count

edCount Case Study:
Evaluating the Validity of English Language Proficiency Assessments

From 2009 to 2011, edCount played a leadership role in the Evaluating the Validity of English language proficiency Assessments (EVEA) project, a federally funded initiative to develop an argument-based approach to validity evaluation for English language proficiency assessments (ELPAs). The EVEA project brought together a consortium of five states – Washington, Oregon, Montana, Indiana, and Idaho – with a team of researchers and a panel of experts. The project partners worked together to develop a comprehensive and coherent framework for considering the meaning and usefulness of scores from English language proficiency assessments. The project's many outputs – including the generic framework for ELPA validity evaluation – are now publicly available to any state or audience, via the EVEA project website (www.eveaproject.com).

Introduction

The Evaluating the Validity of English language proficiency Assessments project (EVEA) sought to address two needs at once – one at the policy level, and one at the practice level.

The policy problem:
English learners (ELs) are a fast-growing population of students in U.S. schools. Between 2000 and 2010, growth rates for EL populations exceeded 250 percent in eight states (Alabama, Arkansas, Delaware, Kansas, Kentucky, Mississippi, Ohio, and South Carolina; National Center for Education Statistics 2001; National Center for Education Statistics 2011). Currently, ELs represent approximately 10 percent of the total U.S. student population, or 4.7 million students (National Center for Education Statistics 2011). And, although two-thirds of the nation's ELs are concentrated within five states (California, Texas, Florida, New York, and Illinois; Capps, Fix, Murray, Ost, Passel, and Herwantoro 2005), ELs can be found in all 50 states and the District of Columbia, representing anywhere from one to 29 percent of total student enrollment.

The EL population is incredibly diverse and heterogeneous. ELs may be any age (the law specifically identifies them as between the ages of 3 and 21), they may speak any language (e.g., Vietnamese, Haitian Creole, Spanish, Tagalog, Chinese-Mandarin, Russian, etc.), and come from any of hundreds of countries, including the United States. It may surprise some people to know that the majority of ELs are actually born in the United States (Editorial Projects in Education, 2009).

Students who are identified as ELs (decided on an individual basis and solely based on language needs, not ethnicity, nationality, etc.) are eligible to receive specialized support, instruction, or accommodations both to help them learn English, and to ensure that they may continue to keep up in their other subjects while they learn.

Since 2002, NCLB has required all states to test all of their K-12 ELs annually using an English language proficiency assessment (ELPA). In addition to counting towards Title I and Title III accountability measures, student performance on this assessment is typically the primary (and sometimes sole) factor used in decisions about whether ELs are ready to be reclassified and stop receiving special language support[1]. As such, the ELPA is the one assessment under NCLB that has a high-stakes outcome for individual children: a proficient score could mean that a child stops participating in English as a second language classes, or no longer receives certain accommodations like the use of a glossary or dictionary on classroom tests throughout the year. Losing these services prematurely could be severely detrimental for the student's continued growth and progress, both in mastering English and in learning academic content.

Given these stakes, it is critically important that those who use ELPA tests are extremely confident that the scores from these tests are valid and reliable. If administrators and teachers cannot be sure that student scores are reliable (meaning that students would be likely to score in the same performance range if re-tested) and valid (meaning the scores do indeed provide an accurate and sufficient representation of the relevant language skills a student needs to survive without language support), this could mean that students who need language supports are not receiving them.

For other high-stakes assessments under NCLB, the U.S. Department of Education requires states to collect evidence showing that their assessment system can validly support the various uses to which states put the scores. This evidence is reviewed by expert peers, who may require states to adjust their systems or collect additional evidence to ensure that their assessments are appropriate given the state's intended uses. Currently, there is no such peer review system for ELPAs, despite their high-stakes status in the lives of individual children.

The practice problem:

It is possible, and perhaps even likely, that many practitioners at the state education agency (SEA) level are aware of these problems and would appreciate the opportunity to investigate their own ELPA systems further and evaluate their validity. Unfortunately, these practitioners may face many challenges that render them unable to follow through on these intentions.

[1] In reality, few, if any, states rely solely on ELPA performance for reclassification decisions. Some states have requirements about minimum scores on specific subtests (e.g., a student must be proficient or above on reading specifically, in addition to having an overall proficient composite score across all test), or may require that students score proficient or above for two consecutive administrations before they exit services. Other

First, money is often a limiting factor. Title III is not a highly-funded program compared to other federal programs, and state budgets generally do not have leftover dollars for non-mandated activities. Second, time is in short supply. SEA personnel often are juggling a number of high-priority, time-consuming responsibilities, and do not necessarily have time to commission or lead additional research activities, even if they would like to see such research done. Third, expertise may be hard to come by. Evaluating an ELPA system requires strong expertise in policy, measurement, research, data analysis, and language acquisition. It is not common for one individual to have expertise in all of these areas. Even if an SEA has a group of qualified staff, coordinating this sort of research effort may still be overwhelmingly challenging given their individual responsibilities and priorities.

Many states have chosen to minimize these barriers by participating in assessment consortia for their ELPAs, hiring an outside vendor to design, administer, and score their ELPA, as well as provide support to personnel and carry out any necessary safeguards to prove and ensure test reliability and quality. Such consortia also provide states with a community of other states with which to communicate and share ideas and best practices, as well as a team of experts who are available to answer questions and provide support.

Not all states have gone this route, however, and those that do not participate in assessment consortia may face real challenges in ensuring that their ELPAs are well-designed and well-functioning. EVEA sought to target states in this second group, and provide them with support to engage in activities, alongside other state partners, that they might not have the time, money, or expertise to do on their own.

edCount worked with Washington State Office for Public Instruction (OPI) to design a project that would address both of these problem areas at once. Together, the WA OPI and edCount gathered a group of states, experts, and partner organizations to collaborate on the EVEA project. These partners included:

- The state education agencies from Washington, Oregon, Montana, Indiana and Idaho;
- The National Center for the Improvement of Educational Assessment (NCIEA), a non-profit devoted to improving educational practices in assessment accountability;
- The Graduate School of Education and Information Sciences (GSE&IS) at the University of California, Los Angeles;
- The Pacific Institute for Research and Evaluation (PIRE), a non-profit research institution that served as the external evaluator of the project's activities;
- Synergy Enterprises, Inc. (SEI), a woman-owned small business that designed the project's private and public websites; and
- A panel of nine pre-eminent experts from the fields of assessment, validity theory, and second language acquisition.

Supported by an Enhanced Assessment Grant from the Office of Elementary and Secondary Education at the U.S. Department of Education, EVEA's purpose was to develop an argument-based approach to validity evaluation for ELPAs. This approach, which is meant to be adaptable to any state's system, offers a comprehensive and coherent framework for considering the meaning and usefulness of scores from ELPAs.

The project's goals included:

- Developing a common argument about how an ELPA is theorized to function within a larger system of education and assessment,
- Developing and piloting research instruments and protocols that states could use to gather information about their ELPA, and
- Gathering resources and information for states about language acquisition, policy relating to ELs and ELPAs, and the validity evaluation process.

The EVEA team met all of these goals, and has made the products available via the project website: www.eveaproject.com.

In addition, the team was able to support two of its principal investigators, Marianne Perie (NCIEA) and Alison Bailey (UCLA/CRESST), in the authorship of four white papers on various topics relating to the assessment of ELs. Each partner state also collaborated with a dedicated research partner to create a theory of action about how its ELPA system functions, and a validity evaluation plan outlining steps and studies the state could use to evaluate this theory of action and collect evidence to support the system's validity.

EVEA Project's Theory of Action
(Visit the *Interactive Theory of Action for this Case Study* online: http://edcount.com/index.php/case-studies/evaluating-validity-of-el-assessments. See the Appendix for static images from the interactive TOA.)

Our theory of action for this case study speaks not to how we conducted the project, but rather to the issues and claims that project partners identified as critical to ELPA validity. To see a specialized theory of action about English language proficiency assessments, readers are encouraged to view the common interpretive argument on the EVEA website, at http://www.eveaproject.com/cia.aspx.

Instructional Contexts
English language proficiency standards and instruction are fundamental components of a successful ELPA system; if these are weak, then assessment outcomes may be meaningless even if the ELPA is well-aligned, well-designed, and faithfully administered and interpreted. Examining these components closely was beyond the scope of the EVEA project itself, but all partners agreed that it would be important for a state to ensure the quality of its standards and instruction as a precursor condition for producing valid ELPA scores.

Many partners agreed that **preparing teachers with information, preparation, resources, and support** was likely to be an area in which many states might need to focus effort to ensure consistent conditions throughout the state. If all teachers, including content teachers, do not understand their responsibility to include and accommodate ELs, it is unlikely that these students will receive the instruction and support they need and deserve.

Assessment Design

Appropriate identification practices emerged as a particularly important – and vulnerable – component of an ELPA system. States must be confident that students are being appropriately identified according to consistent, meaningful and defensible criteria and protocols; the EVEA partners realized that this can sometimes be a difficult process to track and control. The instruments used for this process should also be validated and scrutinized, as they essentially serve as gatekeepers, determining which students go on to receive services and which don't.

The EVEA partners also noted that ELPA scores should appropriately represent student abilities in the four linguistic domains (reading, writing, speaking, listening) and comprehension in order to **represent adequate sampling of knowledge and skill expectations.** When extraneous factors systematically affect test scores, the resulting "construct-irrelevant variance" in test scores may give misleading or inaccurate impressions of how students are actually doing. Construct-irrelevant variance on an ELPA might stem from administration inconsistencies, scoring inconsistencies on teacher-scored subtests for speaking and listening, or even test design flaws. Some of these questions also pertain to whether **test design and delivery conditions allow students to demonstrate their knowledge and skills.** Similarly, some states were interested in ascertaining whether their ELPAs (including all domain subtests) were truly accessible to ELs with disabilities, and allowed these students to demonstrate what they know and can do.

In addition to worrying about scoring inconsistencies, states might also wish to ascertain that their **performance levels accurately differentiate students according to meaningful differences in ability.** While many state partners expressed that they were confident that they knew which students were truly beginners, and which are proficient, parsing different performance levels in between these can be challenging and difficult to do on a meaningful and consistent basis. Many of the project partners agreed that setting meaningful standards and performance levels for an ELPA is a challenging but important component to ensuring meaningful scores.

Score and Interpretation Uses
For most states, ELPA scores serve two purposes. At the student level, ELPA scores should **provide valid and reliable information that allows stakeholders to track student progress** towards attaining proficiency in English. Specifically, states should be confident that changes in test scores for students truly represent changes in skill or ability; if test scores fluctuate for other reasons (such as design errors, scoring errors, systematic item biases), this would undermine the state's ability to use scores as an indicator of student progress and performance.

At the program level, ELPA scores should also **provide valid and reliable information to support accountability decisions and program evaluation**. According to NCLB and the civil rights act, districts must evaluate their programs for ELs regularly, and are obligated to modify and improve programs that are failing to achieve their intended outcomes (usually these outcomes are that students learn English, and do not fall behind in content classes). If student scores do not show growth or progress over time, this might be an indicator the district needs to revise its program to make it more effective. In order to make such interpretations with confidence, however, the state should be confident first that scores are true reflections of student ability, and program quality.

Finally, while the EVEA partners were all committed to getting and providing teachers with any information that might be useful to inform their instruction of ELs, they were skeptical that ELPA scores can **provide valid and reliable information that teachers can use to build aligned instruction.** Due to the nature of the test and the lag time in receiving scores, the partners doubted that the information teachers finally receive from this test would be unlikely to be able to play any major role in their instructional design or decision-making, at least on a day-to-day basis. On a more general level, the partners did agree that it might at least be helpful for teachers to receive their ELs' previous year's ELPA scores and performance levels at the beginning of the year, so they may have a rough sense of where their different ELs are in the second language acquisition process.

System Goals
While the ultimate goal for ELs is that they will **leave high school ready for college and careers**, ELPAs focus on the more intermediate goal that ELs will attain the proficiency in English necessary to allow them to exit services and participate in mainstream English-only classes without special support. In order to attain this goal, the **students must achieve increasingly higher outcomes in English language proficiency**, which can be facilitated by ensuring that **students get greater exposure to high quality academic instruction** that is cognizant of their linguistic needs, and helps them to learn both English and content.

Lessons Learned

Many of the EVEA project participants were ultimately surprised to realize how little may be known or regulated about ELPAs and the larger systems into which they are embedded. The state partners were able to reflect on their own systems with a high degree of scrutiny thanks to the support that EVEA offered them in terms of time, money, and expertise, and were very grateful to be able to do so based on the valuable insights they gained from the project. Lacking this opportunity, however, and lacking a direct federal mandate to explore these issues, it is unclear whether these states would have been able to accomplish any of these activities given the significant demands they already face.

Based on these and other findings, we believe that a mandated federal peer review for ELPAs would be highly advisable, *but*, that states may need additional resources, support, and funding to meaningfully accomplish this.

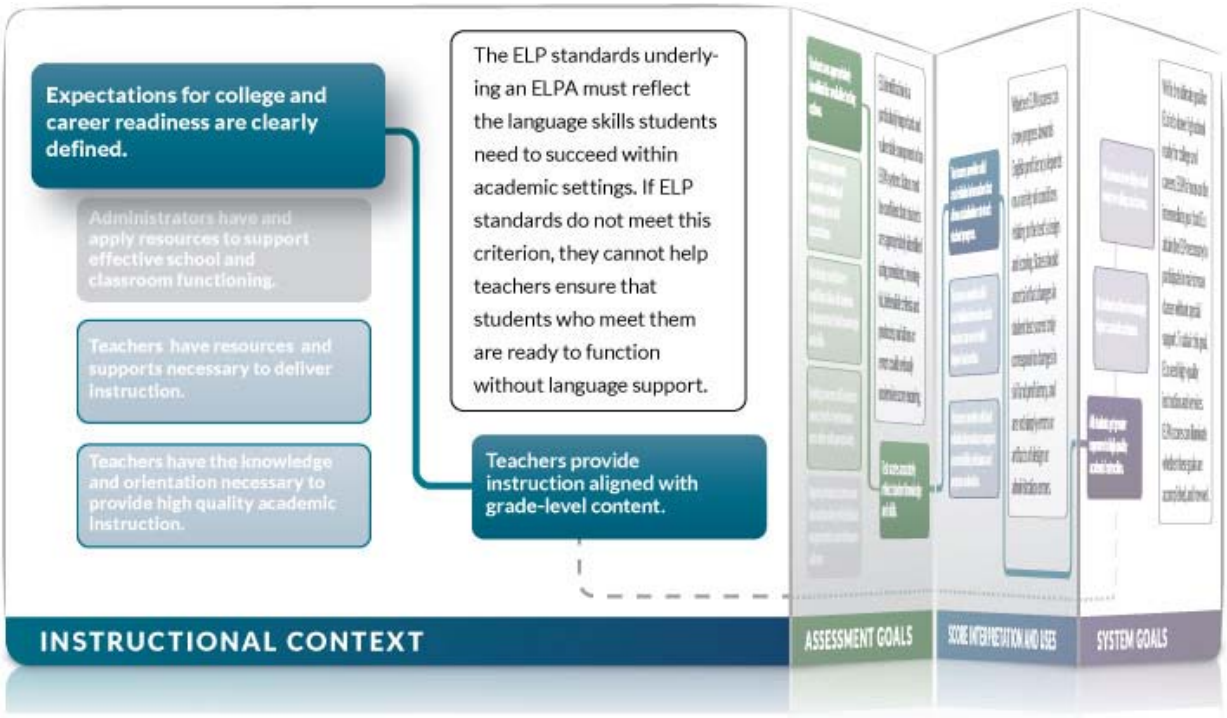
As of September 2011, the U.S. Department of Education has funded a consortium of 28 states, led by Wisconsin, to develop a next-generation English language proficiency assessment. We believe this consortium could stand to benefit significantly from the EVEA project's findings as it embarks on this work. The holistic framework we propose for understanding and interpreting ELPA scores could provide valuable insights that could shape certain components of the new ELPA's design, administration, and scoring.

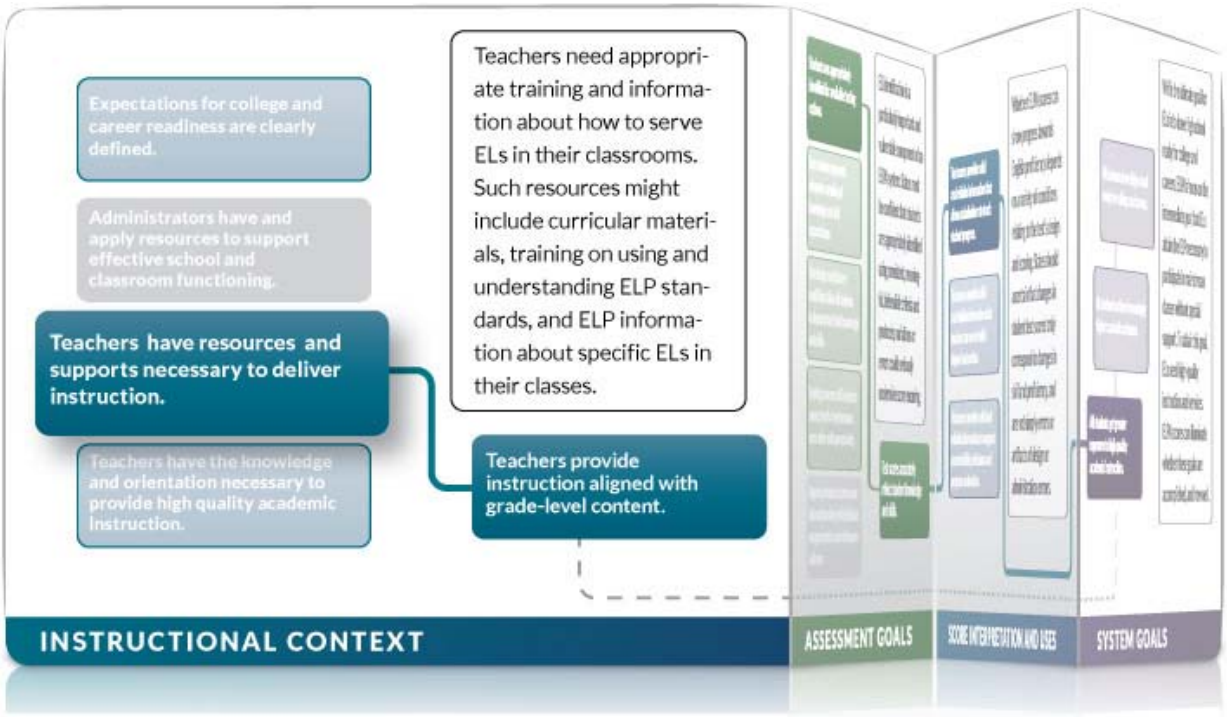
Conclusion

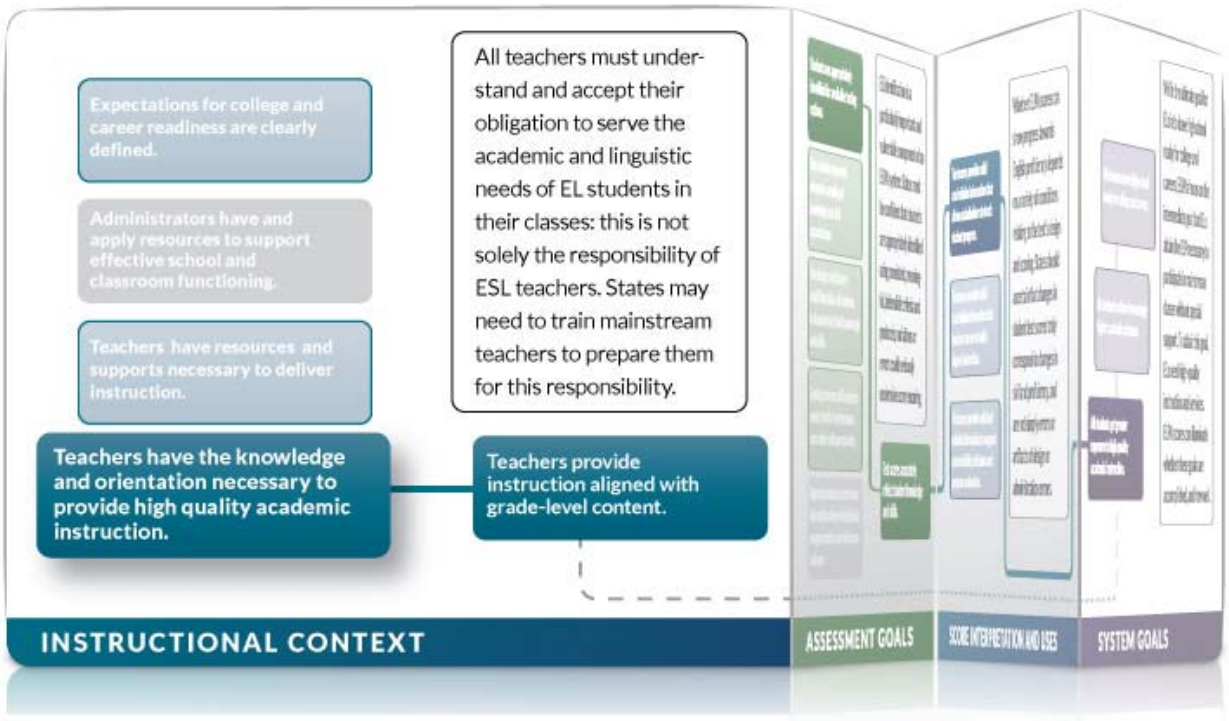
Although the EVEA project has officially ended, edCount remains committed to the issues raised by this work, and to finding venues to support further dissemination, and exploration of the themes and ideas developed here. We continue to work with many of the EVEA partners on other projects, and to keep our eyes open for additional funding opportunities that could support further pursuit of this work. Some of edCount's other current work, including a research project in which we are conducting site visits and literature research about designing, implementing, and evaluating language instruction educational programs (LIEPs), also relates directly to some of the issues and questions raised by our work on EVEA.

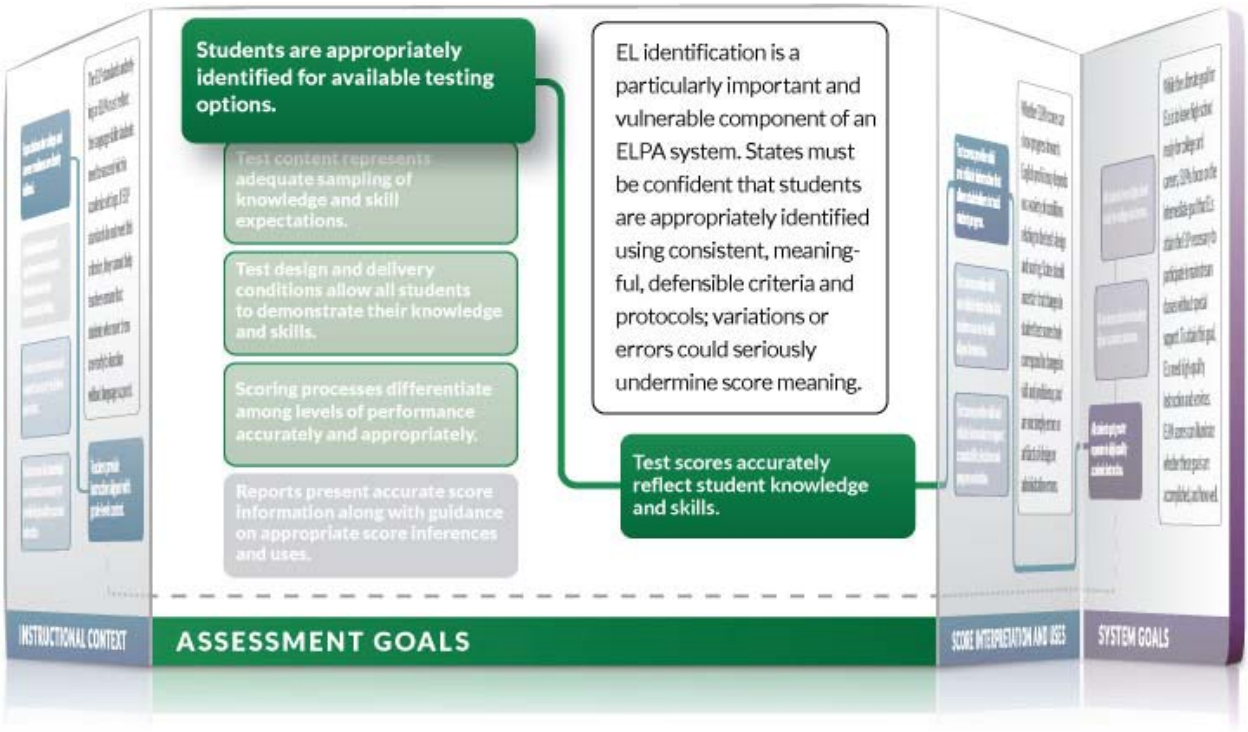
We believe that successfully supporting and including English learners is one of the most critical education issues of the current era, and deserves significant attention and dedication from states, researchers, universities, teachers, communities, and individuals. We remain committed to seeking practical, high-quality solutions in support of this population.

Case Study Appendix: Theory of Action Slides for EVEA
(Visit the *Interactive* Theory of Action for this case study online:
http://edcount.com/index.php/case-studies/evaluating-validity-of-el-assessments)

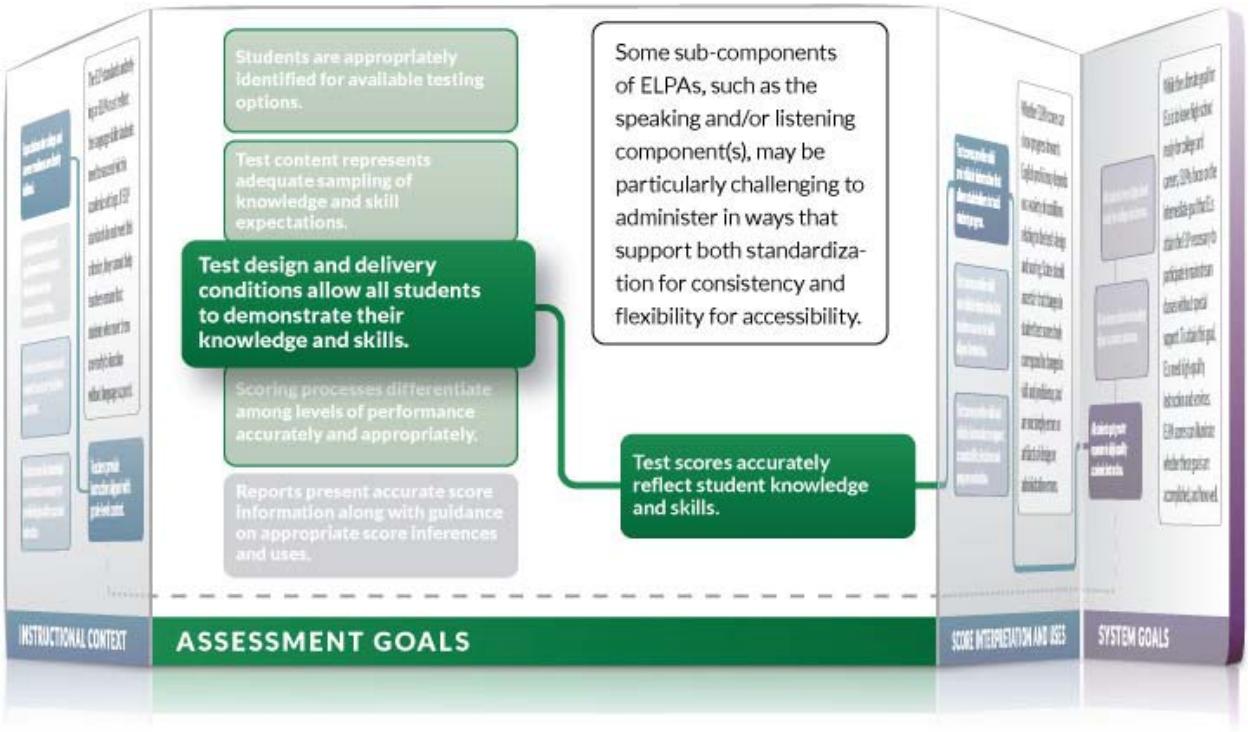


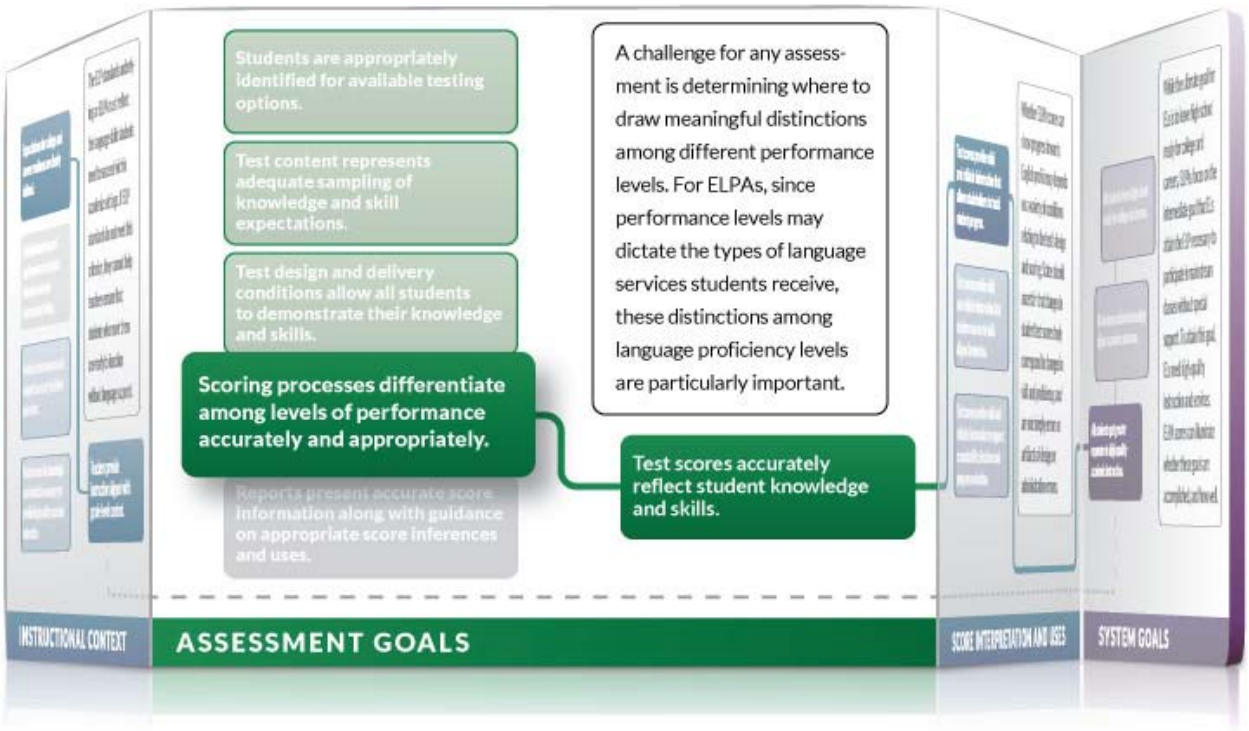


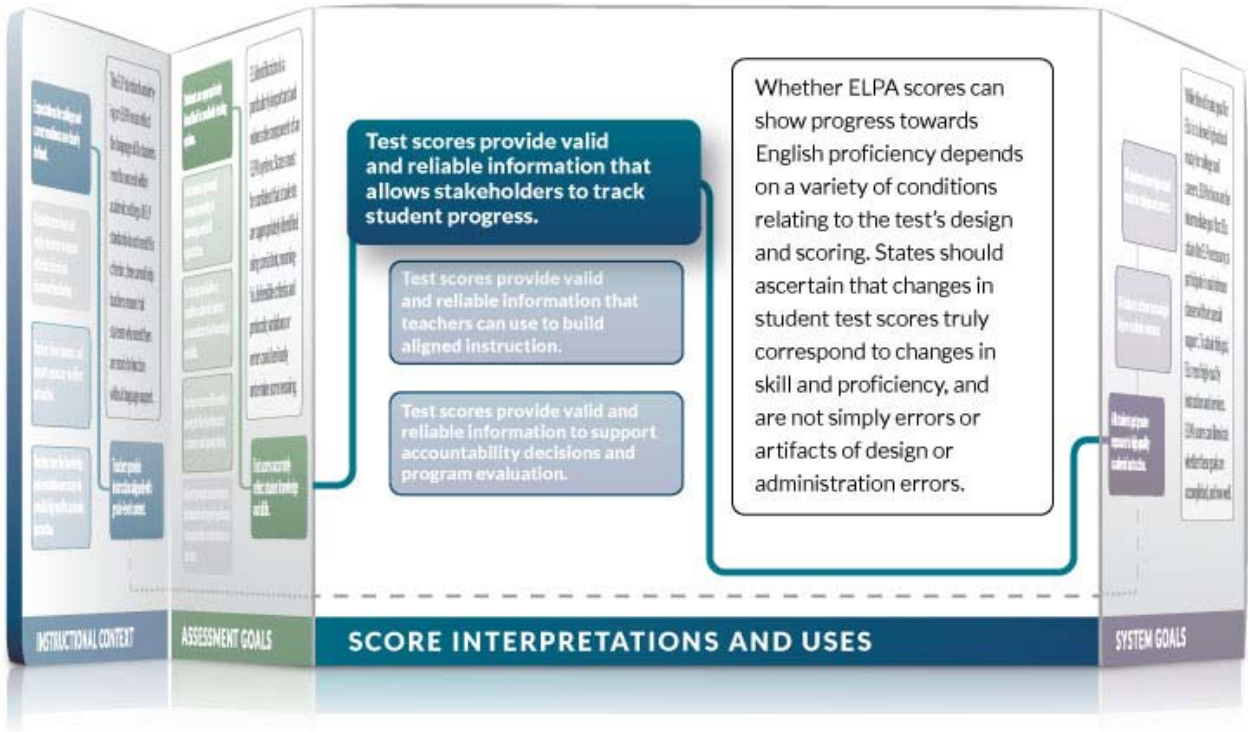

Expectations for college and career readiness are clearly defined.
Administrators have and apply resources to support effective school and classroom functioning.
Teachers have resources and supports necessary to deliver instruction.
All teachers must understand and accept their obligation to serve the academic and linguistic needs of EL students in their classes: this is not solely the responsibility of ESL teachers. States may need to train mainstream teachers to prepare them for this responsibility.
Teachers have the knowledge and orientation necessary to provide high quality academic instruction.
Teachers provide instruction aligned with grade-level content.
INSTRUCTIONAL CONTEXT
ASSESSMENT GOALS
SCORE INTERPRETATION AND USES
SYSTEM GOALS

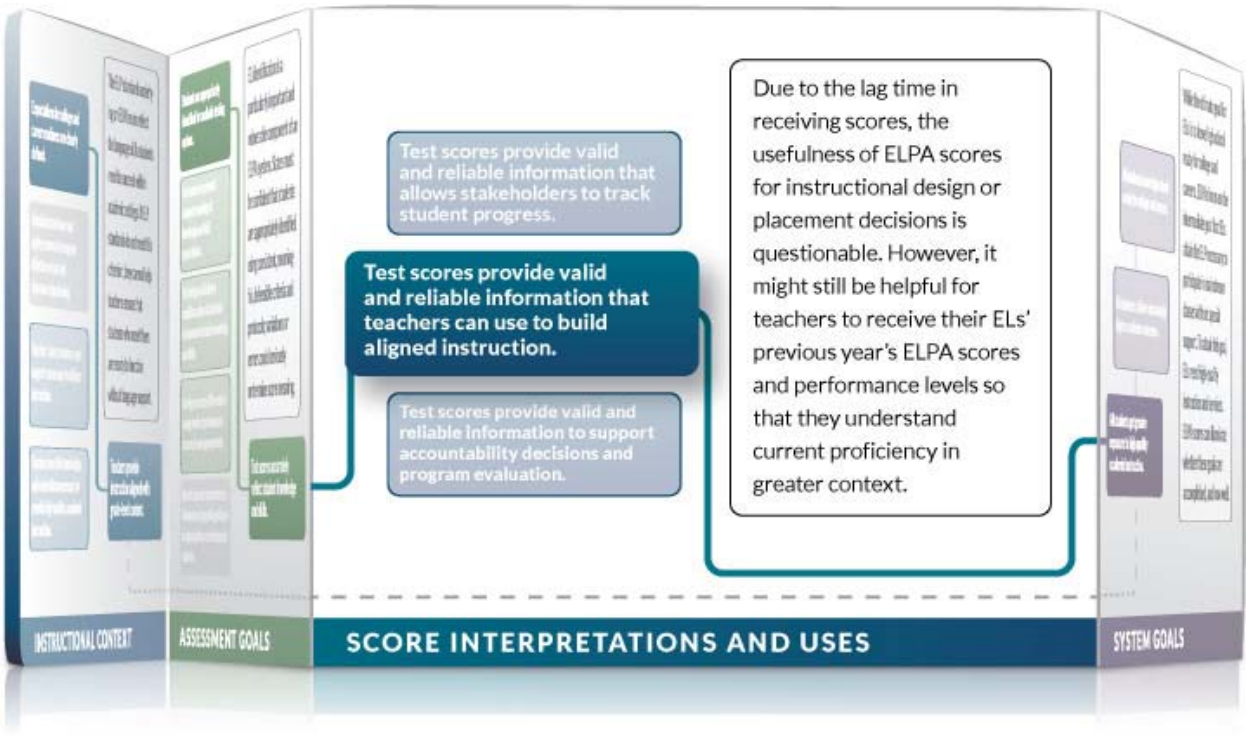

Students are appropriately identified for available testing options.
Test content represents adequate sampling of knowledge and skill expectations.
Test design and delivery conditions allow all students to demonstrate their knowledge and skills.
Scoring processes differentiate among levels of performance accurately and appropriately.
Reports present accurate score information along with guidance on appropriate score inferences and uses.
EL identification is a particularly important and vulnerable component of an ELPA system. States must be confident that students are appropriately identified using consistent, meaningful, defensible criteria and protocols; variations or errors could seriously undermine score meaning.
Test scores accurately reflect student knowledge and skills.
INSTRUCTIONAL CONTEXT
ASSESSMENT GOALS
SCORE INTERPRETATION AND USES
SYSTEM GOALS

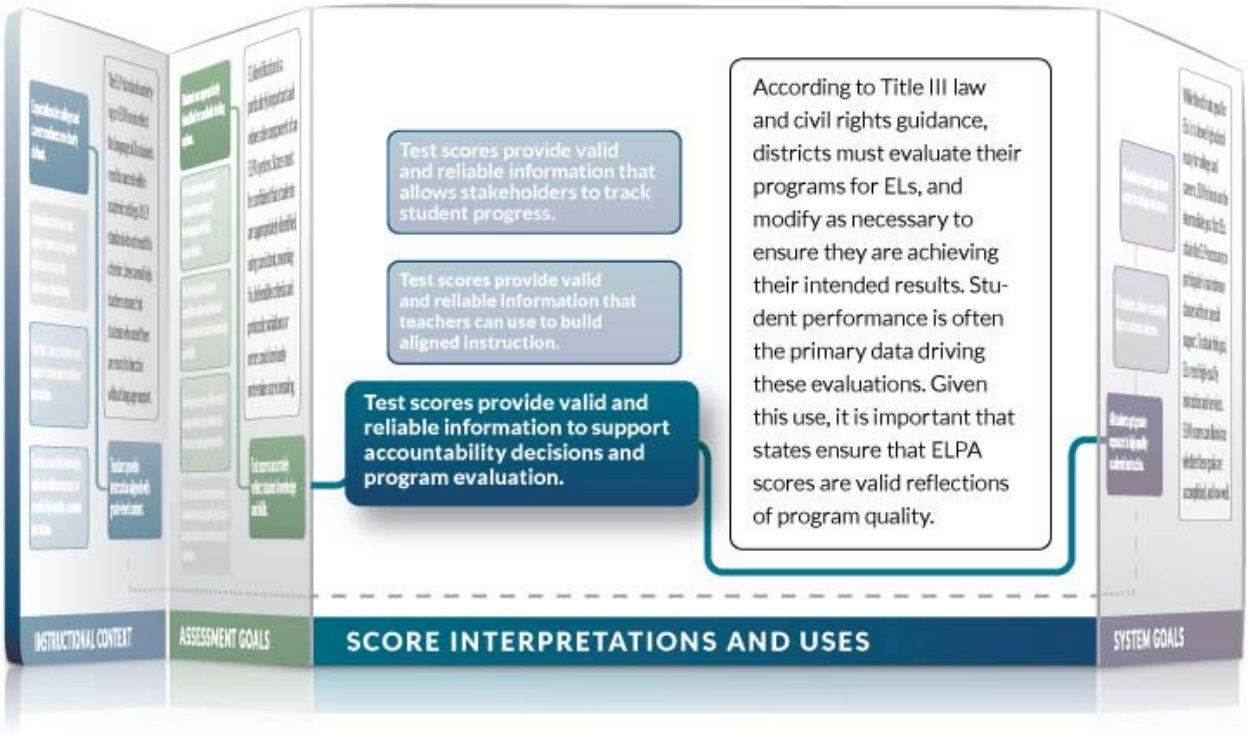

Students are appropriately identified for available testing options.
Test content represents adequate sampling of knowledge and skill expectations.
Test design and delivery conditions allow all students to demonstrate their knowledge and skills.
Scoring processes differentiate among levels of performance accurately and appropriately.
Reports present accurate score information along with guidance on appropriate score inferences and uses.
By law, ELPAs must yield scores for four language domains (reading, writing, speaking, listening) and comprehension. To support the intended uses of these scores, administrators must collect evidence that the assessment does in fact measure these domains as intended.
Test scores accurately reflect student knowledge and skills.
INSTRUCTIONAL CONTEXT
ASSESSMENT GOALS
SCORE INTERPRETATION AND USES
SYSTEM GOALS

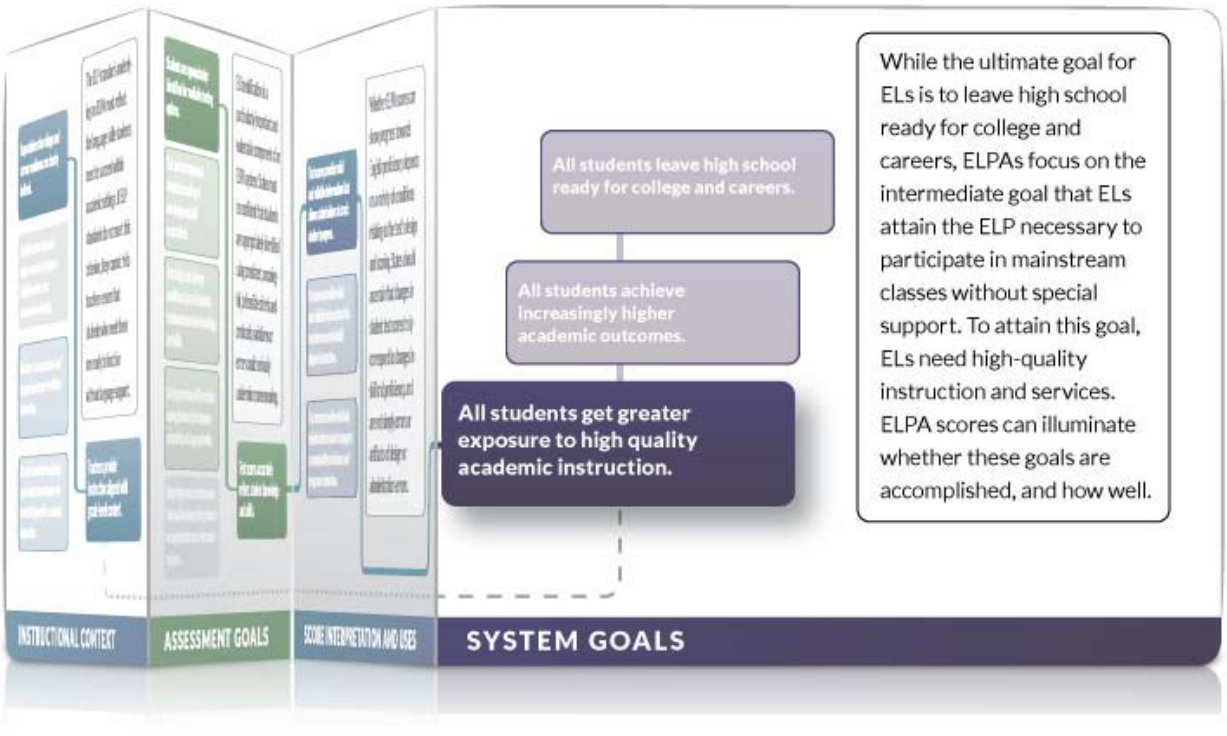

Students are appropriately identified for available testing options.
Test content represents adequate sampling of knowledge and skill expectations.
Test design and delivery conditions allow all students to demonstrate their knowledge and skills.
Scoring processes differentiate among levels of performance accurately and appropriately.
Reports present accurate score information along with guidance on appropriate score inferences and uses.
Some sub-components of ELPAs, such as the speaking and/or listening component(s), may be particularly challenging to administer in ways that support both standardization for consistency and flexibility for accessibility.
Test scores accurately reflect student knowledge and skills.
INSTRUCTIONAL CONTEXT
ASSESSMENT GOALS
SCORE INTERPRETATION AND USES
SYSTEM GOALS

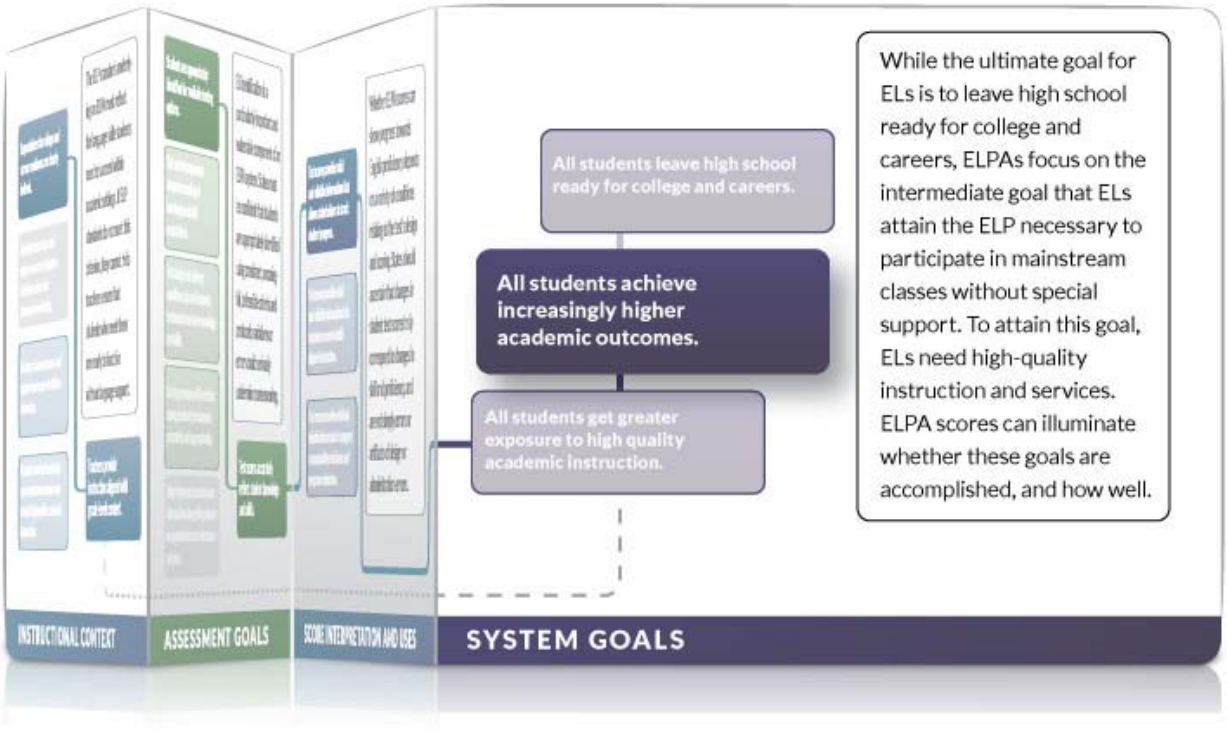

Students are appropriately identified for available testing options.
Test content represents adequate sampling of knowledge and skill expectations.
Test design and delivery conditions allow all students to demonstrate their knowledge and skills.
Scoring processes differentiate among levels of performance accurately and appropriately.
Reports present accurate score information along with guidance on appropriate score inferences and uses.
A challenge for any assessment is determining where to draw meaningful distinctions among different performance levels. For ELPAs, since performance levels may dictate the types of language services students receive, these distinctions among language proficiency levels are particularly important.
Test scores accurately reflect student knowledge and skills.
INSTRUCTIONAL CONTEXT
ASSESSMENT GOALS
SCORE INTERPRETATION AND USES
SYSTEM GOALS

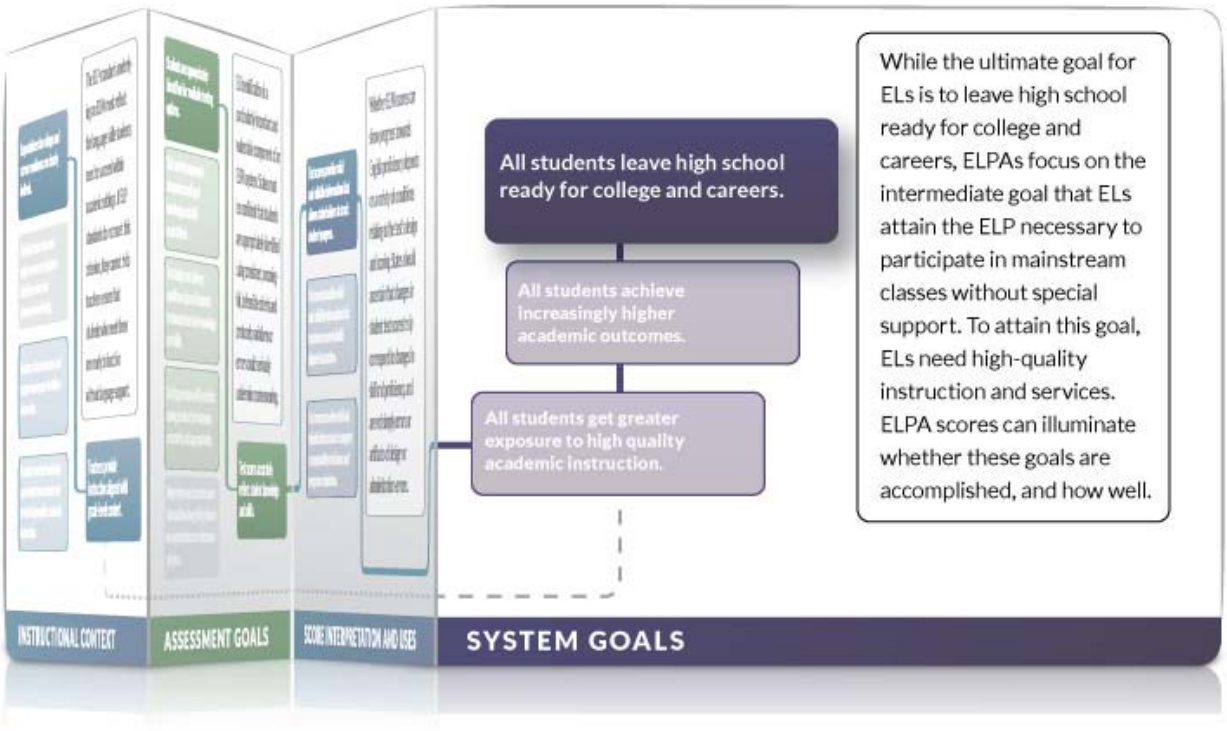

Test scores provide valid and reliable information that allows stakeholders to track student progress.
Test scores provide valid and reliable information that teachers can use to build aligned instruction.
Test scores provide valid and reliable information to support accountability decisions and program evaluation.
Whether ELPA scores can show progress towards English proficiency depends on a variety of conditions relating to the test's design and scoring. States should ascertain that changes in student test scores truly correspond to changes in skill and proficiency, and are not simply errors or artifacts of design or administration errors.
INSTRUCTIONAL CONTEXT
ASSESSMENT GOALS
SCORE INTERPRETATIONS AND USES
SYSTEM GOALS


Test scores provide valid and reliable information that allows stakeholders to track student progress.
Test scores provide valid and reliable information that teachers can use to build aligned instruction.
Test scores provide valid and reliable information to support accountability decisions and program evaluation.
Due to the lag time in receiving scores, the usefulness of ELPA scores for instructional design or placement decisions is questionable. However, it might still be helpful for teachers to receive their ELs' previous year's ELPA scores and performance levels so that they understand current proficiency in greater context.
INSTRUCTIONAL CONTEXT
ASSESSMENT GOALS
SCORE INTERPRETATIONS AND USES
SYSTEM GOALS


Test scores provide valid and reliable information that allows stakeholders to track student progress.
Test scores provide valid and reliable information that teachers can use to build aligned instruction.
Test scores provide valid and reliable information to support accountability decisions and program evaluation.
According to Title III law and civil rights guidance, districts must evaluate their programs for ELs, and modify as necessary to ensure they are achieving their intended results. Student performance is often the primary data driving these evaluations. Given this use, it is important that states ensure that ELPA scores are valid reflections of program quality.
INSTRUCTIONAL CONTEXT
ASSESSMENT GOALS
SCORE INTERPRETATIONS AND USES
SYSTEM GOALS


All students leave high school ready for college and careers.
All students achieve increasingly higher academic outcomes.
All students get greater exposure to high quality academic instruction.
While the ultimate goal for ELs is to leave high school ready for college and careers, ELPAs focus on the intermediate goal that ELs attain the ELP necessary to participate in mainstream classes without special support. To attain this goal, ELs need high-quality instruction and services. ELPA scores can illuminate whether these goals are accomplished, and how well.
INSTRUCTIONAL CONTEXT
ASSESSMENT GOALS
SCORE INTERPRETATION AND USES
SYSTEM GOALS


All students leave high school ready for college and careers.
All students achieve increasingly higher academic outcomes.
All students get greater exposure to high quality academic instruction.
While the ultimate goal for ELs is to leave high school ready for college and careers, ELPAs focus on the intermediate goal that ELs attain the ELP necessary to participate in mainstream classes without special support. To attain this goal, ELs need high-quality instruction and services. ELPA scores can illuminate whether these goals are accomplished, and how well.
INSTRUCTIONAL CONTEXT
ASSESSMENT GOALS
SCORE INTERPRETATION AND USES
SYSTEM GOALS


All students leave high school ready for college and careers.
All students achieve increasingly higher academic outcomes.
All students get greater exposure to high quality academic instruction.
While the ultimate goal for ELs is to leave high school ready for college and careers, ELPAs focus on the intermediate goal that ELs attain the ELP necessary to participate in mainstream classes without special support. To attain this goal, ELs need high-quality instruction and services. ELPA scores can illuminate whether these goals are accomplished, and how well.
INSTRUCTIONAL CONTEXT
ASSESSMENT GOALS
SCORE INTERPRETATION AND USES
SYSTEM GOALS