Measuring the Power of Learning.™

Next-Generation Summative English Language Proficiency Assessments for English Learners: Priorities for Policy and Research

ETS RR-16-08

Mikyung Kim Wolf • Danielle Guzman-Orth • Maurice Cogan Hauck

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Next-Generation Summative English Language Proficiency Assessments for English Learners: Priorities for Policy and Research

Mikyung Kim Wolf, Danielle Guzman-Orth, & Maurice Cogan Hauck

Educational Testing Service, Princeton, NJ

This paper is the third in a series concerning English language proficiency (ELP) assessments for K–12 English learners (ELs). The series, produced from Educational Testing Service (ETS), is intended to provide theory- and evidence-based principles and recommendations for improving next-generation ELP assessment systems, policies, and practices and to stimulate discussion on better serving K–12 EL students. The first paper articulated a high-level conceptualization of next-generation ELP assessment systems (Hauck, Wolf, & Mislevy, 2016). The second paper addressed accessibility issues in the context of ELP assessments for ELs and ELs with disabilities (Guzman-Orth, Laitusis, Thurlow, & Christensen, 2016). The fourth paper dealt with one of the major uses of ELP assessments—the initial identification and classification of ELs (Lopez, Pooler, & Linquanti, 2016). The present paper focuses on summative ELP assessments that states use for accountability purposes. Based on the conversation that took place at the K–12 ELP Assessment Working Meeting on summative ELP assessments, we identify critical policy and research issues. We also highlight the challenges associated with measuring ELP in a diverse group of EL students and recommend ways to embrace and support EL students through shared goals and active communication across federal agencies, state agencies, local agencies, researchers, and practitioners.

Keywords English-language proficiency assessments; K-12 English learners; standards; academic language; reclassification

doi:10.1002/ets2.12091

On June 11, 2014, the K–12 Center at Educational Testing Service (ETS), in partnership with the Houston Independent School District, convened a one-day meeting titled The English Language Proficiency Assessment Research Working Meeting: Summative Assessments (the Working Meeting, henceforth). The focus of the working meeting was on the annual summative English language proficiency (ELP) assessments that states administer to their English learners (ELs) toward the end of each school year. These summative ELP assessments, also known as Title III assessments, have several important uses. The scores are reported to the federal government to satisfy state accountability provisions under the Elementary and Secondary Education Act (ESEA), and they impact federal funding to state and local education agencies. In addition, the assessments inform EL reclassification decisions (i.e., determining when an EL student exits from EL status), which have significant impact on the type of instructional programs that EL students receive.

The goals of the meeting were threefold: (a) to share information regarding the current status and challenges in the development of next-generation summative ELP assessments for EL students; (b) to discuss valid, research-based uses of summative ELP assessments; and (c) to make policy and research recommendations for the improvement of summative ELP assessments and their uses. The meeting participants included researchers, language assessment experts, state and district leaders, policy organization leaders, and representatives of consortia including Assessment Services Supporting ELs through Technology Systems (ASSETS), English Language Proficiency Assessment for the 21st Century (ELPA21), the Partnership for Assessment of Readiness for College and Careers (PARCC), and Smarter Balanced Assessment Consortium (Smarter Balanced; see Appendix A for a list of attendees).

During the Working Meeting, invited speakers made presentations on key issues and relevant research around summative ELP assessments. After each presentation on the specific topic, whole- and small-group discussions were convened (see Appendix B for the Working Meeting agenda and the topics presented). The leaders of each small group reported to the whole group and submitted written summaries of their discussions. The written summaries of each small group discussion revealed that a wide range of topics were discussed in addition to the topics listed in the agenda. For example,

Corresponding author: M. K. Wolf, E-mail: mkwolf@ets.org

at some small-group tables topics such as understanding the developmental characteristics of EL students from various backgrounds, assessment delivery modes (i.e., paper vs. computer), and qualities of instructional programs were discussed as issues that impact the use of summative ELP assessments. While Working Meeting generated rich discussion and raised a number of areas to further examine to improve summative ELP assessments, the time available in a one-day meeting did not provide an opportunity for the meeting participants to evaluate the various topics raised and distill them into an agreed-upon set of policy and research recommendations.

In this paper, we aim to report on the most pressing issues related to summative ELP assessment that emerged from the presentations and discussions at the meeting. This paper is intended to inform national and state leaders with respect to opportunities, challenges, and strategic areas of focus related to the improvement of summative ELP assessments and their uses. We have attempted to identify key policy areas that need federal and/or state policy makers' attention on the use of summative ELP assessments to improve EL education overall. These policy recommendations are based on what the authors have understood to be the most significant issues raised from the presentations and recurring topics discussed across the large- and small-group discussions at the Working Meeting. Recommendations in this paper that are directly gleaned from presentations and discussions at the Working Meeting are so indicated, as are recommendations reflecting the judgment and interpretation of the authors. We also reviewed additional, existing studies on the given topics. We then identified the relevant research areas that should be addressed by the field in the near future to inform the recommended policy areas and stimulate advances in summative ELP assessment.

We begin this paper with relevant background information on the current status of summative ELP assessment to contextualize the call for improved policy and research. We then summarize the presentations and discussions of the Working Meeting on the major topics set by the meeting agenda. Drawing from these discussions and research findings, we discuss policy issues that are ripe for improvement or further investigation in order to maximize the usefulness of summative ELP assessments. This paper is not intended to summarize all the discussions that took place at the meeting, but instead to emphasize a set of major, common issues that the participants raised with respect to annual summative ELP assessments and policies. For instance, screening assessments (those used to identify ELs initially upon their enrollment in school) were not a focus of the Working Meeting. Thus, unless noted, ELP assessments in this paper refer to annual summative ELP assessments. Similarly, formative assessment, despite its importance in improving educational outcomes for ELs, was not a focus of the Working Meeting and is outside of the scope of this paper. We conclude the paper with a set of high-priority research areas to support improvement in summative ELP assessment design, use, and policy.

Contextual Background

The Working Meeting was held at a time of far-ranging change, challenge, and opportunity in the education and assessment of ELs. The current educational reform efforts entail the implementation of new academic standards, and next-generation assessments aligned with these standards, for all students including ELs. With an emphasis on equipping students with the skills and knowledge needed for college and career readiness, almost all states have adopted college and career readiness (CCR) standards, such as the Common Core State Standards (CCSS) and Next-Generation Science Standards (NGSS). Some states that have not adopted these standards have instead modified their own existing standards to reflect an increased focus on CCR. Next-generation assessments aligned with CCR standards are already in place in some states. For instance, in many states, in the 2014–2015 school year, the PARCC and Smarter Balanced consortia will roll out assessments that are aligned with the CCSS for students in Grades 3–8 and one high school grade.

Both the CCSS and individual states' CCR standards feature more rigorous academic expectations and higher language demands than existing states' standards, such as close reading of complex texts, academic discussions, and presentations and construction of arguments from evidence in English language arts (ELA), mathematics, and science, to name a few. This explicit demand for more sophisticated academic language use signals important changes in expectations for the education of ELs and creates an urgent need for new ELP standards and assessments. It is of paramount importance to assist EL students in their development of ELP necessary to meet new CCR standards.

The Working Meeting took place at a point at which states across the country are at different stages in the process of developing and implementing new ELP standards and next-generation assessments based on them. Two consortia, ASSETS and ELPA21, are now working toward implementing next-generation ELP assessments in the 2015–2016 school year. ASSETS' summative ELP assessment will be based on the World-Class Instructional Design and Assessment (WIDA)

consortium's English language development (ELD) standards. The WIDA ELD Standards were originally developed in 2004 and have undergone revision since then. The 2012 WIDA ELD standards were enhanced to make more explicit connections with the language demands in the CCSS. Reflecting a somewhat different approach to a similar goal, ELPA21's ELP standards were newly developed after the release of the CCSS and the NGSS with the intent of making direct correspondence between the new ELP standards and the new content standards. Additionally, some states that are not in either of these consortia have developed their own new ELP standards taking CCR standards into consideration and are working on new ELP assessments (e.g., California and New York).

Next-generation ELP assessments do not differ from earlier summative ELP assessments in their primary use—to meet the federal requirement for all schools to report annually on EL students' progress in and attainment of ELP for accountability purposes. Next-generation assessments will also continue to be used to designate EL students (i.e., to determine students' level of ELP and to exit students from EL designation). Since the passage of the No Child Left Behind Act (NCLB, 2002), current summative ELP assessments have advanced significantly in terms of their systematic measurement of four language domains (listening, reading, speaking, and writing) and their focus on academic ELP.

However, important limitations of existing ELP assessments have also come to light. Heavy emphasis on the accountability uses of these ELP assessments has left local administrators and teachers wanting more timely and useful information to guide and inform instructional programs and support services for EL students. Even though score reports for the current ELP assessments have been designed to meet federal accountability requirements, their usefulness could be enhanced if they were to provide more detailed information about students' language skills. Part of the limited usefulness of current assessment results stems from long scoring turn-around times and poor access to score reports (Hauck, Wolf, & Mislevy, 2016). Another major criticism of current ELP assessments relates to their lack of comparability in terms of the constructs measured and the cut-off scores used to determine students' levels of language proficiency. When the current generation of ELP assessments was developed, relatively little research was available to operationalize the concept of academic language proficiency needed in K–12 school contexts (Bailey & Huang, 2011; Wolf & Farnsworth, 2014). Thus, current ELP assessments differ in their approaches in defining and measuring ELP. This has an important implication for making decisions about EL designation.[1]

In the era of CCR standards and next-generation assessments, we face both significant challenges and significant opportunities to improve the usefulness of new ELP assessments to further support EL students' learning. The next generation of ELP assessments is being developed on the basis of new ELP standards that converge in their correspondence to CCR standards (Linquanti & Hakuta, 2012). These new ELP assessments are capitalizing on advances in technology to measure students' ELP more efficiently and accurately and to improve timely access to assessment results. For example, technology-based test delivery allows for the development of more authentic, multimedia-enhanced items and tasks that elicit students' ELP in situations that more closely resemble real-life communication. In addition, automated scoring and easy data digitization can increase the efficiency of access to test data. To achieve this vision of generating more useful ELP assessments to support the education of EL students, it is important to draw upon what we have learned from the uses and outcomes of current assessments, so as to establish a common understanding of the most promising avenues for enhancing their quality and usefulness.

Key Areas to Consider in the Development and Use of New English Language Proficiency Assessments

In this section, we describe the key issues that emerged from the presentations and discussions at the Working Meeting. The major topics addressed by the invited speakers included the current status and challenges of new ELP assessment development, developments in the definition of the ELP construct given new CCR and ELP standards, the definition of an *English proficient* performance standard to determine an EL's status and the relationship between ELP and academic performance in content areas. For the purposes of this paper, we have reorganized these topics into the following four areas: ELP construct, score reporting, relationship between ELP and content assessments, and EL reclassification. Challenges in summative ELP assessment are discussed in relation to each of these four areas. As noted earlier, these areas are limited to those that were included on the agenda for the Working Meeting. Before the summary and discussion of these four areas, we discuss the value of summative ELP assessment to provide contexts for selecting those areas as pressing issues.

The Value of Summative English Language Proficiency Assessments

It is important to understand why standardized summative ELP assessments are necessary and what value they add to the effective education of EL students who are already being tested in various ways. Putting aside the federal reporting requirements, we maintain that knowing individual EL students' level of ELP constitutes an essential piece of information. For schools that have the capacity to offer ELD classes at different levels, this information is essential for student placement. It is also critical for states to have standardized ELP assessments to serve as objective and common measures of students' year-to-year progress in ELP development. A common measure is also critical for the standardization of EL designation (i.e., ELP level classification and reclassification) across schools.

Standardized summative ELP assessment results can also add critical value to understanding EL students' academic performance. EL students must participate in state content assessments in ELA, mathematics, and science for accountability purposes. The current accountability system, however, does not take into account EL students' ELP assessment results when reporting their content assessment performance, a practice that limits the interpretability of those content assessment results. For students who have just begun to develop ELP, content assessment results are likely to confound content knowledge and ELP (Abedi, 2006). That is, inferences made about students' content knowledge and skills from content assessment results without an understanding of test takers' ELP levels are very likely to be inadequate. Given the critical informational roles that ELP assessments must play, it is essential that they encompass the entire range of ELP skills, from foundational to higher order, to assess students at a wide range of ELP levels. In this sense, ELP assessments have unique value and features that set them apart from all content assessments, including ELA assessments.

Conceptualization and Operationalization of the English Language Proficiency Construct

It is critical to understand the construct to be measured by next-generation ELP assessments if assessment results are to support valid inferences about students' ELP and appropriate decisions based on those results. As briefly mentioned earlier, current-generation ELP assessments were built upon a relatively weak understanding of school language and academic language proficiency. In recent years, considerably more research has been conducted to conceptualize the language used in schools (i.e., school language), a development which should help to test designers better understand how to operationalize academic language proficiency for the ELP assessment (e.g., Bailey, 2007; Gottlieb, Katz, & Ernst-Slavit, 2009; Scarcella, 2003; Schleppegrell, 2004; Snow & Uccelli, 2009; Valdes & Wong Fillmore, 2011). School language can be conceptualized broadly as covering both the language used for learning academic content and the language used to communicate with peers and school staff outside of instructional time. The literature has mainly focused on the characteristics of academic language uses across the instructional content areas. The various ELP standards to which older ELP assessments have been aligned to have required substantial change in order to correspond to the more rigorous CCR standards. In order for ELP standards to be used in tandem with new CCR standards, ELP standards must include the types of language skills that students need to benefit from instruction in content areas and perform the tasks manifested in these CCR standards. Thus, for those who develop or adopt new summative ELP assessments, it is of the utmost importance to revisit and clearly define the ELP construct.

During the working meeting, the representatives of the ASSETS and ELPA21 consortia presented the underlying conceptualization of the ELP construct embodied in their respective standards. As described earlier, ASSETS is developing its summative ELP assessments based on the WIDA ELD Standards, which were revised to correspond to the CCR standards. In their presentation, Gottlieb and Wilmes (2014) explained that the conception of the ELP construct for the ASSETS assessment focused primarily on the use of academic language across various content areas, such as language arts, mathematics, science, and social studies (see Figure 1). To operationalize this construct for assessment development, the features of academic language are organized further into word/expression, sentence, and discourse levels. As noted in Figure 1, sociocultural context is emphasized to indicate that the context in which language is used is important to determine the characteristics of academic language use.

Representing the ELPA21 consortium, Ho (2014) presented ELPA21's newly developed ELP standards, which have the explicit aim of supporting the development of the language skills that EL students need to achieve CCR standards. In a manner similar to that underlying WIDA's ELD standards, Ho illustrated the conceptualization of the ELP construct as academic language used across content areas. Although WIDA's ELD standards include a sample or exemplar task specific to each disciplinary area, ELPA21's ELP standards focus on common language tasks that underlie multiple disciplinary

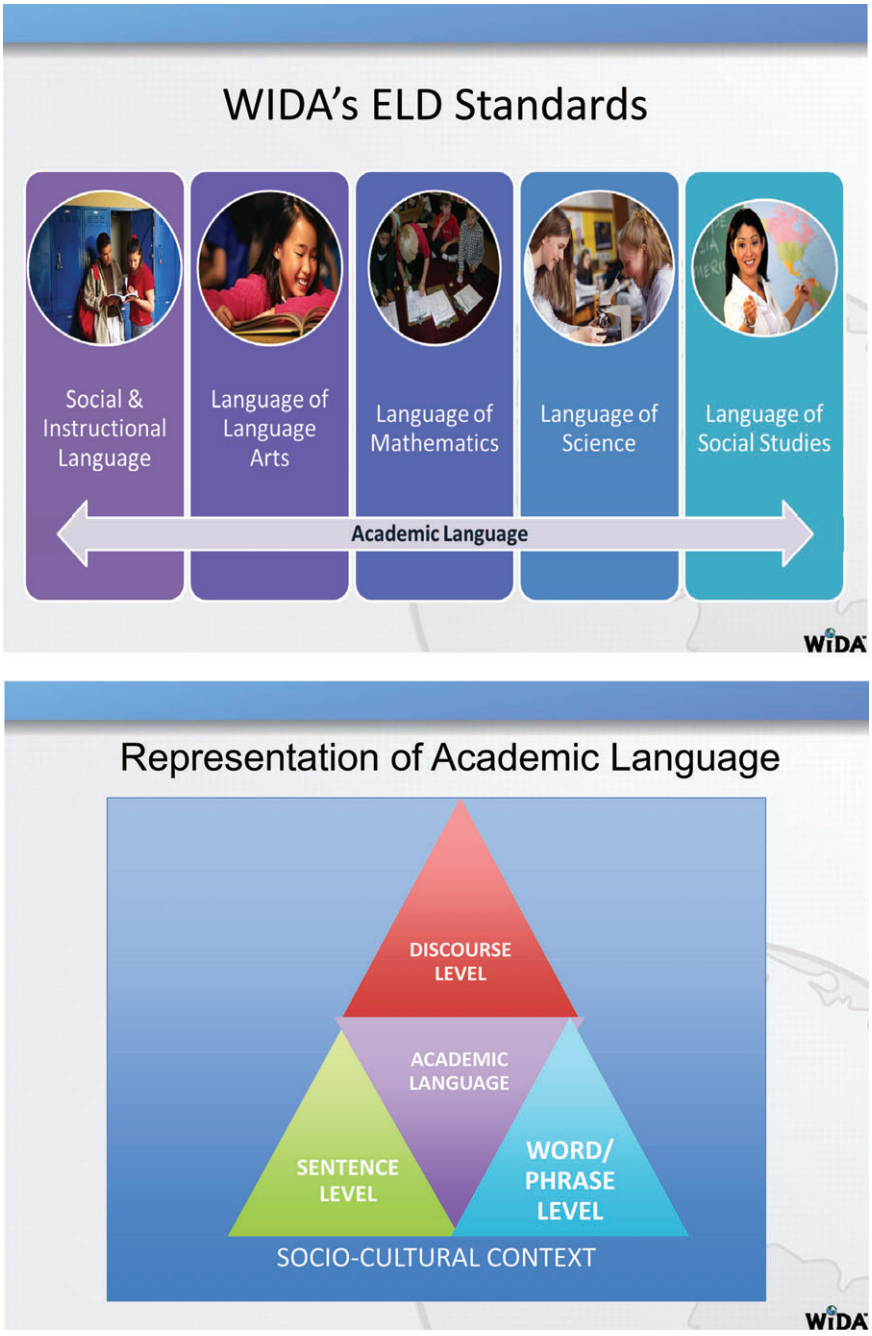

Figure 1 Illustrations of the World-Class Instructional Design and Assessment's approach to conceptualize the academic language construct for English language proficiency assessments. From *World-Class Instructional Design and Assessment English Language Development Standards and Assessment System* by M. Gottlieb and C. Wilmes, 2014, presentation given at the English Language Proficiency Assessment Research Working Meeting: Summative Assessments by the K–12 Center at ETS, Houston, TX.

areas (e.g., analyze and critique the arguments of others orally and in writing). Figure 2 illustrates the ELPA21 approach of identifying language skills across various disciplinary areas and determining the language skills to be measured in the summative ELP assessment.

It is interesting to note that ELPA21's ELP standards are not broken into the traditional four language domains of listening, reading, speaking, and writing. Rather, the standards include descriptions of integrated language skills, a feature which appears to focus on more complex uses of language in order to correspond to the more rigorous language demands

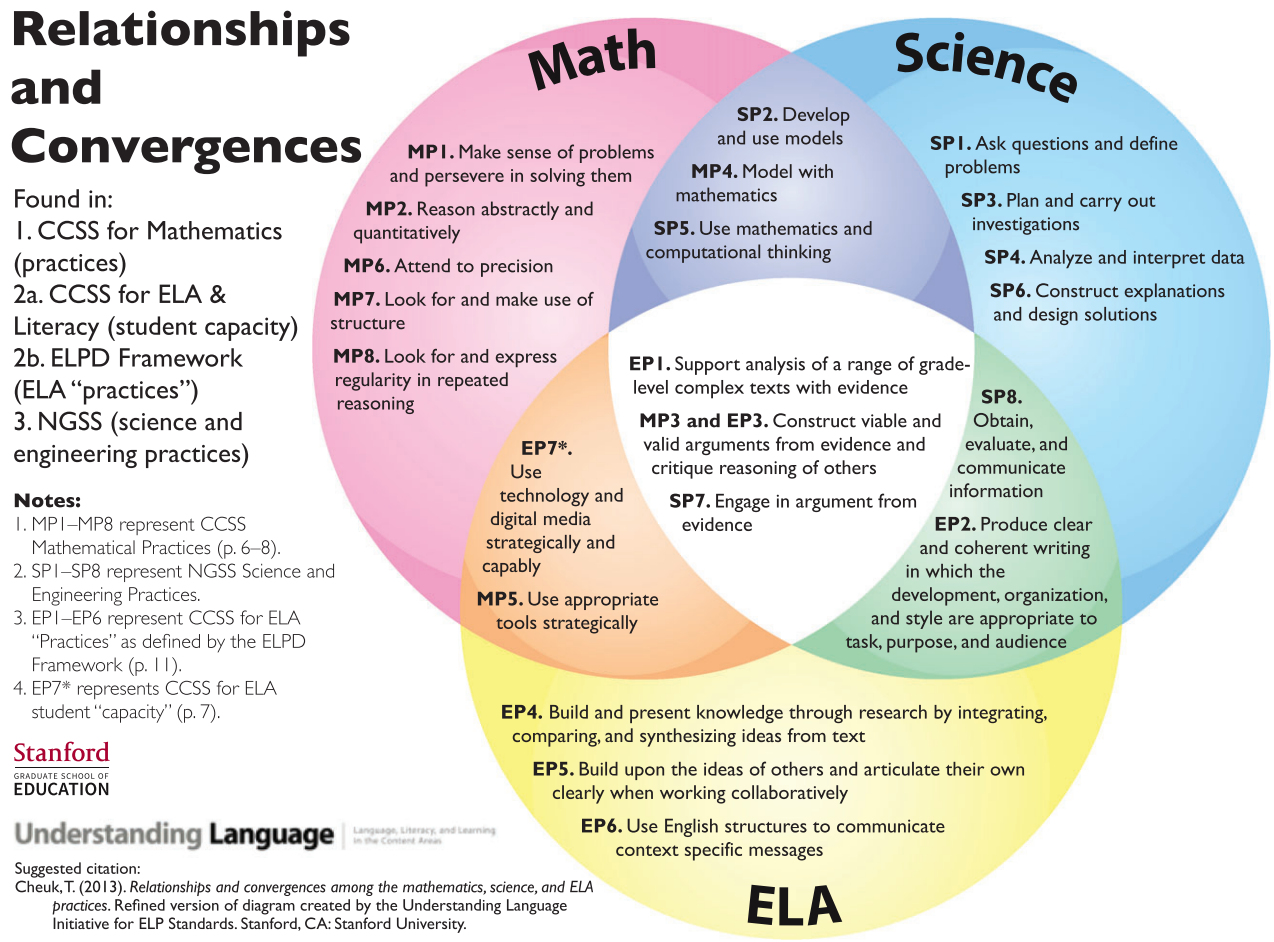

Figure 2 English Language Proficiency Assessment for the 21st Century's approach to identify common and specific language demands across various disciplinary areas. From *Relationships and Convergences Among the Mathematics, Science, and English Language Arts Practices* [Refined version of diagram created by the Understanding Language Initiative for ELP standards] by T. Cheuk, 2013, Stanford, CA: Stanford University.

embodied in CCR standards. This approach to conceptualizing academic language proficiency with reference to CCR standards is also reflected in California's new ELD standards (California Department of Education, 2012). California's ELD standards also emphasize the communicative use of language. Instead of the four language domains, California's new ELD standards are organized by three communication modes: collaborative, interpretive, and productive modes. These changes reflect the position that ELP should be taught and developed in highly contextualized situations (i.e., language use in various content learning environments). As will be discussed later, this changing view of the construct is in some degree of tension with the established practice of reporting scores by individual language domain.

Considering that the two ELP assessment consortia and California serve approximately 63% of the US EL population, the impact of these changes will be significant in the assessment of EL students. It is also noteworthy that the different approaches toward operationalizing the ELP construct and academic language by the two consortia might perpetuate differences in EL designation (e.g., different EL classifications depending on the types of ELP assessments), as well as in the types of information that result from the assessments.

Lopez and Wolf (2014) presented one way of conceptualizing and defining the construct of ELP assessments, given the various existing ELP standards and CCR standards. They argued that it is critical to identify in the standards, and in the literature, the key tasks that students have to perform and then to further explicate all the language skills needed to perform those tasks. The *Framework for English Language Proficiency Development Standards Corresponding to the Common Core State Standards and the Next-Generation Science Standards* (ELPD Framework; Council of Chief State School Officers [CCSSO], 2012) offers useful examples of key tasks involved in the language uses of the CCR standards. The approach

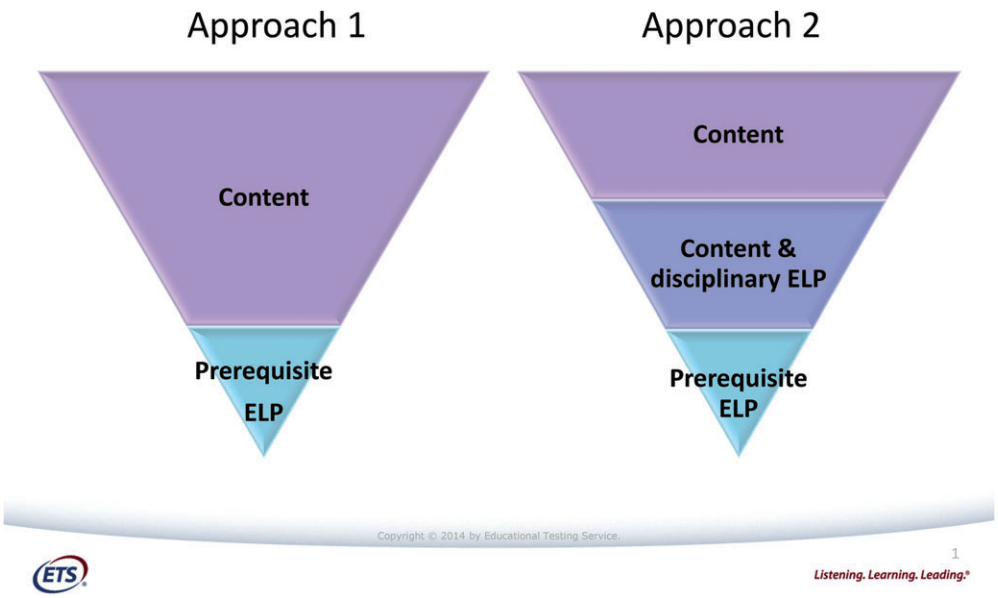

Figure 3 An illustration of varied approaches to operationalizing the English language proficiency construct. From *Conceptualizing and Operationalizing the Construct of Standards-Based English Language Proficiency Assessments* by A. Lopez and M. Wolf, 2014, presentation given at the English Language Proficiency Assessment Research Working Meeting: Summative Assessments by the K–12 Center at ETS, Houston, TX.

of identifying key tasks was reflected in ELPA21's ELP standards. Lopez and Wolf further proposed that organizing the language tasks, language functions, and specific linguistic resources associated with the tasks under target language use domains is one way of operationalizing the ELP construct systematically for the development of ELP assessments (see Wolf, Everson et al., 2014, for more details). Based on the target language use domains defined by Bailey and Heritage (2008), Lopez and Wolf offered social/interpersonal, school navigational, general-academic, and discipline-specific language use domains as examples.

Another important consideration in defining the ELP construct is the degree of overlap between ELP and content assessments. This is particularly the case for the relationship between ELP and ELA assessments. As ELA includes students' academic language use, it is natural that some language functions may be measured in both assessments (e.g., inferring meaning from context, analyzing the organization of arguments). This overlap is evidenced in many ELA and ELP standards (see Wolf, Wang, Huang, & Blood, 2014). Lopez and Wolf (2014) discussed two possible approaches to handling overlap in the development of ELP assessments (see Figure 3). In Approach 1 shown in Figure 3, ELP assessments are made to focus primarily on measuring foundational and prerequisite knowledge and skills for grade-level tasks. For instance, knowing the meanings of words, decoding text, and using understanding of sentence structure to interpret the meanings of the sentences are necessary to perform any tasks across content areas. In this approach, the purpose of ELP assessments is to measure students' prerequisite ELP. This approach is also intended to minimize the burden of double-testing EL students.

In Approach 2, ELP assessments encompass both prerequisite language knowledge and skills *and* more complex and sophisticated language knowledge and skills specific to the content areas, including ELA. For example, the ELP construct in Approach 2 may include knowing discipline-specific vocabulary and understanding specific language structures to explain mathematical procedures. Thus, there is overlap between the constructs of ELP and content assessments in this approach. Lopez and Wolf (2014) pointed out that states and consortia need to articulate their approach to ELP construct definition so that stakeholders can make adequate inferences about students' ELP on the basis of assessment results.

Summative English Language Proficiency Assessment Reporting

As described earlier, summative ELP assessment results have various intended uses, including as a tool for fulfillment of the federal requirement that states monitor annual progress and attainment of EL students' ELP (i.e., Title III Annual

Table 1 Examples of Variation in Composite Score Weighting for Sample Summative English Language Proficiency Assessments

ELP Assessment Name	State/Consortium	Listening (%)	Speaking (%)	Reading (%)	Writing (%)
CELDT (K – 1)	California	45	45	5	5
CELDT (2 – 12)	California	25	25	25	25
ACCESS	WIDA	15	15	35	35
TELPAS	Texas	10	10	50	30
KELPA (K – 1)	Kansas	35	35	15	15
KELPA (2)	Kansas	25	25	25	25
KELPA (3 – 5)	Kansas	25	15	30	30
KELPA (6 – 12)	Kansas	30	10	30	30

Note. ACCESS = Assessing Comprehension and Communication in English State-to-State for English Language Learners; CELDT = California English Language Development Test; ELP = English language proficiency; KELPA = Kansas English Language Proficiency Assessment; TELPAS = Texas English Language Proficiency Assessment System; WIDA = World-Class Instructional Design and Assessment. Adapted from *Score Reports for English Proficiency Assessments: Current Practices and Future Directions* by M. Faulkner-Bond, M. Shin, X. Wang, A. Zenisky, & E. Moyer, 2013, paper presented at the annual conference of the National Council on Measurement in Education, San Francisco, CA.

Measurable Achievement Objectives 1 and 2, NCLB, 2002) and as a key criterion used by schools to determine EL students' readiness to exit EL status (i.e., reclassification). To comply with federal reporting requirements, summative ELP assessments have reported scores in the four language domains of listening, reading, speaking and writing, as well as comprehension and overall composite scores to reflect students' ELP.

During the Working Meeting, three high-level issues were raised regarding the score reporting for summative ELP assessments. First, a comparability issue was mentioned in that there is variation across the states in how the four language domains are weighted to create composite scores of overall ELP. Second, a question was raised as to whether the traditional practice of reporting four domain scores is in harmony with current views — reflected in new ELP standards — of ELP as communicative and integrated. Third, possible ways to provide more useful information in summative ELP assessment reports were discussed, with the aim of offering more actionable information for both content-area and language teachers in next-generation assessment score reports. Below we provide details related to each issue presented on at the meeting.

Cook (2014) and Thompson (2014) pointed out that states have employed a range of distinct weighting schemes to determine overall ELP scores, complicating the interpretation of students' overall proficiency across states. The researchers argued that the different weighting schemes have a practical impact on the meaning of the label *English proficient* and on reclassification rates. Table 1 presents a few examples of variation in the weighting schemes of the four domain scores across states. The weights often differ across the grade levels within a single assessment. For instance, literacy skills (i.e., reading and writing) weights tend to be less in the primary grades and to increase in the upper grades, although this is not consistent across all states. (See Faulkner-Bond et al., 2013, for a review of various summative ELP assessments reporting across states.)

As indicated in Table 1, interpretation of overall ELP scores may vary depending on how the assessment domains are weighted. For example, most people would assume that a fourth grade student classified as *advanced* based on overall score on California English Language Development Test (CELDT) and another fourth grade student with the same classification as advanced on Texas English Language Proficiency Assessment System (TELPAS) would have similar ELP ability profiles. However, upon close examination, CELDT weighting in Grades 2 – 12 is evenly distributed across the four domains, while the TELPAS is weighted toward literacy (80% of the test). A score of advanced, then, can have different meanings on these two tests, different implications for EL placement, and quite possibly different subsequent consequences for reclassification.

Adding to this complexity, another challenge for new ELP assessments arises from the mismatch between the view of English proficiency embodied in existing federal accountability requirements and that of new ELP standards. NCLB requirements led states to create ELP standards and assessments reporting on students' English proficiency including comprehension, listening, reading, speaking, and writing. The requirements were interpreted by the states and test publishers as a need to report on students' skills in four domains. One advantage of this approach is that it emphasizes the importance of assessing all skills, not simply those that are easiest to measure. However, recent literature evidences a conceptual shift away from four discrete language skills (listening, reading, speaking, and writing) and toward an integrated

conception, also evident in new ELP standards (CCSSO, 2012; van Lier & Walqui, 2012). As described earlier, ELPA21's ELP standards and California's ELD standards have adopted a more complex conception of ELP to prepare EL students to meet CCR standards. If states are to continue current reporting practices, ELP assessments aligned to these new ELP standards may face technical and practical challenges to reporting the traditional four domain scores.

At the Working Meeting, Ho (2014) suggested that these reporting challenges can be viewed as opportunities to create a *better* reporting system for summative ELP assessments that goes beyond merely fulfilling accountability requirements. That is, if summative ELP assessments and score reports are (a) devised to provide more descriptive and actionable information about students' ELP, (b) are tightly aligned with the standards, and (c) can be provided in a more timely manner, then the assessments and score reports will be more useful to teachers in their instructional program planning and professional development efforts to support EL students' learning. Rivera (2014) further stressed the importance of making a systematic connection between the ELP and content assessment reports so that both language and content teachers can make use of the results effectively in planning instruction, as well as to increase the accuracy of the accountability system for EL students.

Relationships Between English Language Proficiency and Academic Performance

Understanding EL students' ELP levels and their impact on performance in content learning and assessment has been a key topic related to maximizing the usefulness of ELP assessment results. Since ELP assessment results are used to make inferences about students' readiness to participate meaningfully in academic contexts without specialized supports, a positive relationship between ELP and content assessment performance is considered to be important evidence to validate the use of ELP assessments. During the Working Meeting, research on this area was presented, demonstrating a positive correlation between EL students' levels of ELP and overall performance on content assessments. More interestingly, the presenters suggested that investigating the relationship between ELP and content assessments can help to define EL's language proficiency in a manner that will better inform EL classification decisions.

Cook (2014) and Linquanti (2014) described several methods for investigating the ELP performance range in which EL students' ELP is no longer a major hindrance to demonstrating content proficiency on academic content assessments. Use of these methods would provide empirical evidence for policymakers to use in determining the ELP level at which EL students are deemed English proficient and ready for reclassification. For instance, Cook, Linquanti, Chinen, and Jung (2012) utilized decision consistency, logistic regression, and box plot methodologies with large datasets from three states with different ELP and content assessments. In each state they found convergent evidence of a *sweet-spot* range on the ELP assessment in relation to EL students' content proficiency in ELA and mathematics assessments. Notably, they found that this range was higher in relation to performance in ELA (typically *early advanced* or higher ELP levels) than in mathematics (typically *intermediate* to *high intermediate* ELP levels). They argued that relating these available data from ELP and academic content assessments could provide empirical evidence to support policymakers in selecting the *neighborhood* for performance standard setting on ELP assessments.

In examining the relationship between ELP and content assessment performance, Francis (2014) pointed out that, despite strong evidence of a relationship between ELP and content assessment performance, proficiency on ELP assessments does not necessarily lead to proficiency on content assessments. To illustrate this point, Francis presented the results

Table 2 Percentage of Variance in English Language Arts and Mathematics Achievement Scores That Are Accounted for by English Language Proficiency Assessment Scores

	English Language Arts Grade					Mathematics Grade				
Source	4	5	6	7	8	4	5	6	7	8
District	0.37	0.54	0.56	0.57	0.58	0.21	0.16	0.26	0.20	0.18
School	0.39	0.41	0.38	0.74	0.69	0.34	0.33	0.16	0.28	0.26
Student	0.27	0.35	0.34	0.44	0.36	0.20	0.21	0.18	0.19	0.16

Note. From *Conceptualizing the Relationship Between English Language Proficiency and Academic Performance* by D. Francis, 2014, presentation given at the English Language Proficiency Assessment Research Working Meeting: Summative Assessments by the K–12 Center at ETS, Houston, TX.

of his empirical study on the analysis of variance in EL students' content assessment performance. Table 2 shows the percentage of variance in content assessment scores that are accounted for by students' ELP scores at the district, school, and student level across Grades 4 – 8 from one state's dataset. Overall, the table suggests that roughly 27 – 44% of the variance in students' ELA scores and 16 – 21% of students' mathematics scores are accounted for by their ELP performance. This suggests that while ELP is a valid predictor of student performance on ELA and mathematics content assessments, ELP is not the *only* plausible predictor of student performance.

This could be partly because the summative ELP assessments do not represent a complete range of English language skills. Francis also suggested that this variability in performance on content assessments might be attributable to other background characteristics. Moreover, he argued that the language demands embedded in both ELP and content assessments could obscure the relationship between ELP and content assessment performance, but that the relationship may be more meaningful when using statistical modeling that allows for unobservable (latent) classes to emerge (i.e., latent class modeling). Investigating theoretical unobservable latent groups as well as empirically observable groups for heterogeneous subgrouping of ELs could yield better informed estimates of what ELP looks like (and where the associated cut points should be) for ELs. Francis suggested that investigating use of instructional interventions and test accommodations for ELs (i.e., examining the extent to which the intervention or linguistic accommodation, such as translation, read aloud, etc., are no longer useful) may also help to create a better understanding of proficiency cut points. This has potential to impact the nature of the underlying relationship between language and content assessment performance (i.e., instead of relying on existing academic assessment proficiency cut points to determine the relationship in student language and content assessment performance, a more nuanced understanding of the pattern between language and content performance may emerge). In fact, both Cook (2014) and Francis (2014) asserted that failure to analyze the language complexity within content assessment tasks obscures understanding of how language is tied to the assessment of content, making it difficult to make meaningful inferences about what ELs know and are able to do.

Collectively, these studies represent a range of methods to consider in developing and implementing new ELP assessments. Further, they underscore the importance of examining the existing data to inform assessment development and development of assessment policies for appropriate uses of the assessment results.

English Learner Reclassification

Another dimension of the relationship between ELP and content assessment performance was explored in Linquanti's elaboration of the concept of *English proficient* in the context of ELP and academic assessments used for reclassification purposes (Linquanti, 2014). Reclassification is a high-stakes decision because it results in the removal of specialized support services that EL students are legally entitled to as a protected class under federal law. Even though EL students are expected to leave this category eventually, a sizable group of ELs do not meet the criteria for exit and remain ELs for an extended period of time (long-term ELs). This phenomenon underlines the need for further investigation into the process of reclassification and the criteria used to inform such decisions, which would provide adequate technical guidance and ensure EL students are well served.

Investigating the validity of current reclassification criteria and examining the factors contributing to retaining ELs in the designation is crucial. Linquanti (2014) noted that many states use multiple criteria for reclassifying EL students, and these criteria can include ELP assessment scores, content assessment scores, local criteria such as grades or GPA and teacher/parent/guardian recommendations (see Linquanti & Cook, 2013; Wolf et al., 2008). Discussing the variability in reclassification criteria and the validity challenges this raises, Linquanti noted that many instruments used for reclassification are neither designed nor intended to measure language proficiency. In his presentation, he posed questions to encourage the audience to reflect on the concept of *English proficient*, what such proficiency means for current and next-generation ELP assessments, and how educators might use multiple sources of ELP evidence.

Although the challenges to defining proficiency for ELs are clear, Linquanti (2014) suggested that these challenges can motivate development of more empirical tools and methodologies to determine how English proficiency can be captured in next-generation ELP assessments and how local evidence-gathering tools can help educators recognize and foster ELP. This redefinition is especially critical as the new standards signal further integration of language uses and content-area practices; as academic language uses may be assessed as part of academic content constructs (e.g., discipline-specific language uses); and as language functions (e.g., explain, justify, argue) are fully embedded in the content CCR standards

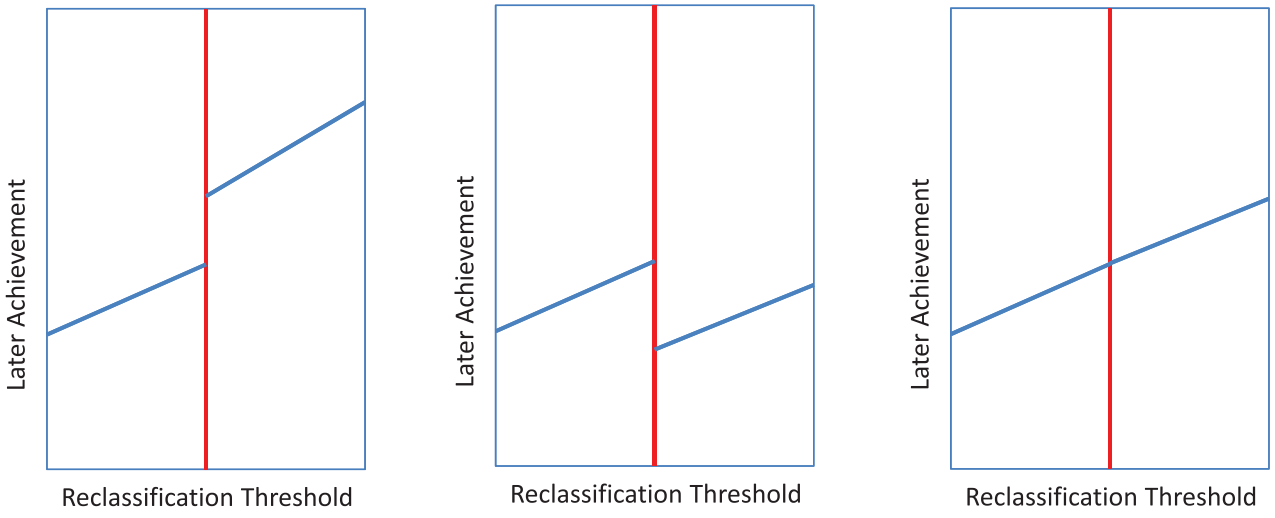

Figure 4 Possible outcomes that may occur when states set their English language proficient assessment proficient performance standard. From *Illustrating Empirical Relationships Between English Language Proficiency and Academic Performance* by K. Thompson, 2014, presentation given at the English Language Proficiency Assessment Research Working Meeting: Summative Assessments by the K–12 Center at ETS, Houston, TX.

(CCSSO, 2012). Opportunities now exist to refine evidence sources of ELP, to better conceptualize and support appropriate reclassification of current EL students as English proficient.

As a concrete example, Thompson (2014) presented empirical studies based on a longitudinal dataset of EL students' ELP and content assessment scores to explore how domain weights and test format and structure may influence how language domains correlate to content-area performance, and also how the cut point demarcating the English proficient performance standard can influence subsequent performance after reclassification. Her investigation into one state's process of rescaling their cut points for ELP levels provides unique insight into the importance of cut points, not only for the various proficiency levels within domains, but also for the implications the cut point has on ELs' ability to attain the English proficient performance—and possibly exit from the EL designation. Using state data from pre- and post-2007 (the year the state assessment was rescaled), Thompson found evidence to suggest that rescaling the ELP assessment made attaining ELP more difficult for ELs. However, doing so also improved the effect of reclassifying on later academic outcomes for high school students, who had not done as well under the old threshold, leading to the interpretation that they had previously been reclassified too soon.

As illustrated in Figure 4, Thompson conceptualized this cut point optimization as key to finding a proper reclassification threshold. The purpose is to avoid a simple *positive effect* (i.e., that reclassified EL students just beyond the ELP threshold perform on content assessments well beyond their counterparts just at the threshold), as well as a *negative effect* (i.e., that reclassified EL students just beyond the threshold perform on content assessments well below their counterparts just at the threshold). The goal is to determine the point of neither positive nor negative effect (i.e., that EL students just before and just after the ELP threshold perform equally). Thompson noted that the *no effect* result around the threshold is the proper sweet-spot range, as it suggests that students at the threshold do not demonstrate appreciably different performance. As with other presenters, Thompson noted that by such methods, reclassification itself does not act as a gatekeeper to rigorous content instruction opportunities.

Recommendations for Policy and Research Related to Summative English Language Proficiency Assessments

We have described the major issues discussed at the Working Meeting concerning existing ELP assessments, the development of new ELP assessments, and their uses in accountability and EL designation purposes. These issues reveal a number of policy areas that offer the potential for improvement to support the current reform efforts. In this section, we discuss the areas for which policy issues warrant attention, followed by a set of pressing research areas to inform those policy issues.

Policy Recommendations to Support the Meaningful Use of English Language Proficiency Assessments

Recommendation 1: Federal and State Education Agencies Should Provide a Clear Definition of English Learners and Guidelines for English Learner Designation

The current federal definition of an EL (the definition of *limited English proficient* in NCLB) includes the description of having difficulties in speaking, reading, writing, or understanding the English language that may deny the individual the ability to meet the state's proficient level of achievement on state assessments, to successfully achieve in classrooms where English is the language of instruction, or the opportunity to participate fully in society (NCLB, 2002). This definition has been operationalized with considerable variation in practice for EL designation. Varied standards and ELP assessments with differing constructs, as well as various weighting schemes for determining overall ELP, have contributed to this variation in defining the term *EL* in practice. Further, this federal definition is often interpreted to include students' academic performance as one of the criteria to define an EL, a situation which has lead to the use of a range of criteria, including content assessment results, in many states. The lack of clear guidelines on how to use these various criteria has also created variability in classifying EL students' ELP levels and exiting students from EL status, even across schools within the same state.

At present, there is a national discussion underway exploring how states and consortia might move toward a more common definition of EL (e.g., Linquanti & Cook, 2013). With the fast movement of current educational reforms in the form of CCR standards, new ELP standards and new assessments, it is possible that a more consistent concept of what English proficient means will emerge. The more streamlined correspondence between content and ELP assessments, largely due to the efforts of the consortia, and using common empirical methods, will help increase the comparability of EL classification across states. However, it is important to recognize that criteria other than results from ELP assessments will be used to make decisions about EL designation (e.g., observation, student interviews, and parental/guardian input). Allowing for local flexibility is essential, considering the different structure of decision-making processes, resources, or students' needs across local educational agencies. Yet, this flexibility should take place in principled ways.

Therefore, federal and state education agencies must offer clear guidelines about what EL definition local districts should adopt (e.g., the existing federal definition or the consortia's definition) and how to use ELP assessment results and other criteria to make EL designation decisions. In establishing such guidelines, federal policy guidelines for the reporting of EL students' ELP development should be revisited to provide explicit guidance on what to report from ELP assessments (i.e., discrete four domain scores, integrated language skill scores, and weighting schemes of different language skills for overall proficiency scores). Such policy actions would support efforts to establish a more common definition of the status of EL.

Recommendation 2: Federal and State Policies Should Provide Guidance on Systematic Approaches to Linking Content Assessment Results and English Language Proficiency Assessment Results

As noted in previous sections, research has yielded congruent results that there is a strong, positive relationship between students' ELP and academic performance in content areas based on the data from ELP and content assessments (e.g., Cook et al., 2012; Francis & Rivera, 2007; Parker, Louie, & O'Dwyer, 2009; Thompson, 2012). However, EL students' ELP levels are not systematically taken into consideration when reporting their content assessment results. At the policy level, the lack of specification or guidance on the connection between Title I (Improving the Academic Achievement of the Disadvantaged) and Title III (Language Instruction for Limited English Proficient and Immigrant Students) has contributed to an inadequate picture of EL students' performance in accountability reports (Hakuta, 2011; Hopkins, Thomspon, Linquanti, Hakuta, & August, 2013). It has also contributed to the isolation of Title III, making the ELP assessment results less known to the content teachers who teach EL students.

Understanding EL students' ELP levels and their areas of strength and needs in ELP through Title III facilitates better interpretations about EL students' academic performance on content assessments for the Title I requirement. More importantly, a better understanding aids educators in making appropriate instructional program decisions to meet the students' needs.

Thus, we call for policy at the federal or state level to require a systematic link between ELP and content assessment results. Federal guidance must be established, requiring the reporting of EL students' academic content assessment performance accompanied by their ELP levels. The findings from empirical studies suggest there is a clear, positive relationship between EL students' ELP levels and content assessment performance. This relationship confirms the importance of reporting EL students' content performance as a function of ELP levels (e.g., Cook et al., 2012). Given the lack of progress of many ELs at intermediate levels of ELP, it is also critical to report on EL students' ELP progress as a function of initial ELP level and time in the state school system (Hopkins et al., 2013; Working Group on ELL Policy, 2010). One of the next steps to support this action is to have a statewide database system that integrates content and ELP assessment data with the students' background information. It is also crucial to keep track of reclassified EL students' prior EL status and date of exit to allow for longitudinal consequential validity studies. Such an integrated database system will facilitate the generation of comprehensive score reports for EL students for both accountability and instructional services. It will also raise the awareness of content teachers about their EL students' language skills, creating opportunities for shared understanding and collaboration among language and content teachers. These changes would support more effective uses of content and ELP assessment results in order to advance effective EL education.

Recommendation 3: States Should Provide Guidance and Professional Support to Local Education Agencies for the Use of English Language Proficiency and Content Assessments to Support English Learners Instruction

To support the implementation of the next generation of content and ELP assessments based on CCR standards, professional development and support is more important than ever (Hakuta, 2014). Teachers need to understand the expectations and changes in the new CCR and ELP standards as well as in new assessments to provide appropriate instruction for EL students. To make appropriate interpretations of assessment results and make valid academic decisions about EL students based on these assessment results, teachers must also understand the construct of the assessments and polices around the assessment uses. Despite the critical importance of professional support for the success of any educational reform, relatively little attention has been paid to this area at the policy level.

As for policy changes to support improvements in EL instruction and assessments, systematic and ongoing professional development around the use of ELP assessments should be in place for both content and language teachers. Clear documents to explain new policies, standards, and assessments are essential to support professional development. An effective channel to communicate the guidelines should also be established at the state, district, and school levels. In particular, professional development to increase teachers' assessment literacy is crucial considering the current shifts in ELP constructs and the various ways of defining ELP levels across different assessments. Teachers and other stakeholders who use ELP assessment results should have a clear understanding of the ELP construct and assessment scores in order to view them in light of the standards and their instruction across content areas. Further, educators (teachers and administrators) must receive professional development regarding the appropriate and valid uses of assessment results to make informed decisions about EL placement and reclassification.

While funding for and consistent implementation of professional development activities are often a challenge, setting appropriate polices and funding to allow for systematic, continuous professional development should be a high priority for states, because, in practice, teachers are the ones who implement new reform efforts.

Recommendation 4: The Federal Department of Education Should Monitor the Validity and Effectiveness of the Use of English Language Proficiency Assessments

ELP assessment results involve relatively high-stakes decisions for program evaluation, funding eligibility, and resource allocation, as well as individual students' academic paths. Hence, it is crucial to ensure that the quality of the assessments and the uses of the assessment results are valid. The NCLB requires states to submit evidence to demonstrate the technical qualities of the states' standards and assessment systems for adequate accountability. To evaluate collective evidence,

a peer review program has been utilized with the federal guidelines as in the *Standards and Assessments Peer Review Guidance* (U.S. Department of Education, 2009). This guidance lays out important types of evidence to collect for technical quality and alignment of content assessments.

Federal guidance for states to examine and demonstrate the qualities of their ELP assessments is also needed and is currently under development. Many of the elements listed in *Standards and Assessments Peer Review Guidance* (U.S. Department of Education, 2009) are in fact applicable to ELP assessments. However, ELP assessments also require unique qualities, or different priorities in examining their qualities, compared to content assessments. For instance, alignment tools used for content assessments cannot be directly applied to ELP assessments considering the central aspect of linguistic complexity in addition to the cognitive aspect (Cook, 2005). The provision of evidence for rigorous standard setting to determine ELP levels is also crucial for ELP assessments because the levels determined from ELP assessments influence EL students' reclassification. Further, states should continuously monitor the uses of ELP assessments, including their consequences, as part of important validation efforts. Particularly, the use and consequences of the assessments for EL designation is a critical area where much guidance is needed. As in the current *Standards and Assessments Peer Review Guidance*, it will be useful to have examples of the evidence to be collected in the areas of ELP assessment uses and consequences. Such guidance will facilitate states' active investigation and documentation of empirical evidence to support or refine their accountability assessment systems, including those for ELP assessments.

Recommendation 5: Federal and State Education Agencies Should Offer Sustained Support for Collaboration Among State Stakeholders, Practitioners, and Researchers

States have been charged with various tasks to ensure that they have appropriate assessment systems for accountability. The development of summative ELP assessments, the appropriate uses of the assessment results, and the provision of validity evidence entail considerable amounts of specific resources and expertise. It may be unrealistic to expect states alone to comply with all these requirements. While we call for federal support through the provision of guidance for states, we also recommend that federal support include funding specifically for states to collaborate with researchers to investigate the validity of ELP assessment uses. A systematic connection among state stakeholders, researchers, and practitioners will be beneficial for improving the assessment system to support EL education. The sustainability of this effort can be achieved with federal-level guidance and policy enactment.

Research Areas Needed to Support Policy Recommendations

The policy recommendations discussed above demand much empirical research to establish appropriate guidance and policies around the use of ELP assessments. Based on the discussion at the Working Meeting and policy areas recommended above, we select the following research areas as the most pressing and relevant to help with implementing policy guidance.

Validation of Criteria Used in English Learner Designation

Much empirical research is needed to produce refined understanding and guidance on best practices in determining ELP levels and reclassifying ELs. As discussed earlier, schools use various combinations of multiple criteria to determine the exit of EL students from EL designation. The primary criterion, results of ELP assessments, requires in-depth investigation. For instance, investigation of the constructs, alignment with standards, composite scores, and standard setting to determine various levels of proficiency for ELP assessments is critical.

In particular, research to help determine the appropriate ELP level at which to reclassify EL students is urgently needed. While defining what it means to be proficient in English in school contexts can be done conceptually, based on the theories and principles, such definitions must be accompanied by empirical data to validate the conceptual definitions and inform the setting of proficiency levels from ELP assessments. To fulfill this need, some specific research areas for further examination include the relationship between students' ELP levels and their academic performance in content areas, the academic performance of reclassified EL students over time, and the longitudinal trajectories of current and reclassified EL students' performance in both ELP and content assessments.

Thus far, only a handful of studies have been completed, providing important insights into determining the levels for EL designation (Cook et al., 2012; Parker et al., 2009; Thompson, 2012). Further research should be conducted in order to

generalize and strengthen the findings so that they may inform appropriate policy and practical decisions with confidence. Research in this area also needs to be expanded to examine the consequences of EL reclassification to ensure students do not remain in the EL status longer than needed or exit prematurely despite the need for ongoing instructional support.

School Language Characteristics and English Language Proficiency Developmental Stages

The field has made considerable progress in understanding the concept of school language, including both academic and social language. Although there are competing conceptualizations of what constitutes academic language, a general consensus has emerged that academic language needs to be explicitly taught to EL students and assessed through summative ELP assessments (DiCerbo, Anstrom, Baker, & Rivera, 2014). As discussed throughout this paper, the CCR standards and new ELP standards require certain language skills for students to participate successfully in content-area learning. Moreover, disciplinary literacy at the secondary level in history/social studies, science, and technology is specifically explicated in the CCSS. While these disciplinary literacy skills will be embedded in content assessment for all students, it is unclear what language skills need to be assessed in ELP assessments. For instance, should we measure discipline-specific language in summative ELP assessments? If so, how? How can the line between general academic language and discipline-specific language be defined? More research is needed to identify the language characteristics specific to and common across the various content areas, as well as the foundational language skills with which EL students should be equipped in school settings.

With the new types of language demands expected of students in new CCR standards, further empirical research is necessary to help educators and stakeholders better understand the characteristics of the school language EL students need and the ways to operationalize this school language in ELP assessments. This line of research will be beneficial not only for assessment design and development but also for the instruction of ELP to ensure that students are provided opportunities to develop appropriate language skills. The recent national initiative called Understanding Language[2] is an example of research that delves into understanding the language and literacy skills contained in the CCSS and NGSS.

Another important area of research is concerned with collecting empirical evidence to validate the ELP levels described in standards and assessments. A close and systematic examination of actual school discourse and the materials with which both EL students and mainstream monolingual English-speaking students are engaged is a much needed area of research (e.g., Bailey & Heritage, 2014; Hiebert, 2013; Schleppegrell, 2012). Such research will also help to define what being English proficient means for EL students and to set reasonable expectations for the development of ELP. Further, it will advance knowledge about operationalizing academic or school language that would be assessed in ELP assessments.

Alignment Among Standards, Curricula, Instruction, and Assessments

The challenges and issues discussed regarding the development of new ELP assessments and appropriate use of ELP assessments for EL designation point to the importance of empirical investigation of alignment among new standards, curricula, instruction, and assessment. Research into the alignment of ELP standards, curricula, instruction, and assessment is relatively scarce compared to similar research in the context of academic content areas. With the current transition to new ELP standards and assessments, research on the alignment across these four areas is particularly pressing to ensure that ELP assessment results are used validly for EL students. Alignment evidence between the assessments and standards should be gathered via research studies conducted not only by the test developers (or others with direct involvement in the design and development process) but also by independent researchers. In the course of such studies, examination of the correspondence between the language skills represented in ELP standards and assessments and those in content standards and assessments will be necessary.

ELP standards should be understood in the same way by test developers and practitioners, as intended by standards developers. Empirical investigation of a common understanding across standards, curricular materials, instructional practice, and assessments needs to take place. Thus, research is needed to examine practitioners' actual interpretations of the assessment results and instructional practices based on the standards and assessment results. An investigation of the types of professional support to implement the standards and understand the assessment results will be an important piece of alignment research. This line of research will not only help to ensure fairness and equity in the use of the assessment

results but will also offer useful guidance to support the teachers' understanding of the relationships among the standards, curricula, instruction, and assessments.

Another area of alignment issues lies in the assessment reporting practice. As described earlier, the current accountability requirement has been interpreted by many states and test developers to report four discrete domain scores (listening, speaking, reading, and writing). This reporting brings up challenges for states that are moving toward implementing new ELP standards reflecting the integrated nature of language skills as they are actually used in the classroom. That is, the examination of alignment will become more complicated given the discrepancy between the reporting requirements and the language skills as described in the standards. Research to offer guidance regarding alignment practices will be useful to inform policies on the reporting requirement.

Subgroups of English Learners and Contextual Factors

Thus far, we have described the research areas that were explicitly related to the topics presented and discussed during the Working Meeting. One additional area of research underlying the discussions that we the authors would like to raise is examination of variation within an EL group. Despite a wide recognition of the EL group's heterogeneity, students' background characteristics and the relationship between these characteristics and performance on ELP and content assessments have not been well researched. For example, despite EL students speaking more than 400 languages and coming from various ethnic and racial backgrounds, all the students are grouped similarly and their academic results are reported in the aggregate as EL or non EL. EL students differ with respect to home language use and proficiency in the home language as well as English (e.g., balanced bilinguals or dominant bilinguals), formal schooling experiences, length of stay in the United States (e.g., US born or newcomers), socioeconomic status, types of instructional program (e.g., English immersion, dual language immersion, transitional bilingual education), and more.

Investigating the existing data may provide further distinguishable classes within the EL group (Guzman-Orth & Nylund-Gibson, 2013). Research has shown that EL students tend to be densely clustered in economically disadvantaged areas or schools (Capps et al., 2005). Yet, it is not clear to what extent this contextual factor impacts student performance. Students' background information can be used by score users to interpret score information to arrive at a more meaningful understanding. This understanding can inform subsequent decision making to support the teaching and learning of students. Thus, research to delve into EL students' background characteristics and other contextual factors and their relation to student performance would be of considerable value.

Along this line of research to examine the background characteristics within the EL group, it is important to examine the students' opportunity to learn the language being assessed in ELP and content assessments. For example, some students' low scores on ELP assessments can be related to their length of stay in the United States, their formal schooling experience, or the instructional services they receive. Hence, it is critical to examine such important background variables and their association with EL students' progress and performance over time to inform adequate policy guidelines.

Further Validation of English Language Proficiency Assessment Uses

Validation of assessments involves a continual process of accumulating various strands of evidence and making integrated arguments supporting the validity of the intended uses of assessments. The appearance of new assessments and the constant changes in the students taking the assessments warrant the importance of this continual validation investigation. The aforementioned research areas already encompass various sources of validity evidence. Considering that summative ELP assessments are primarily intended to be used for accountability purposes and for EL designation purposes, validity research to support these uses is critical. As mentioned above, empirical research on the constructs of the assessments; the alignment of the assessments with the standards, curricula, and instruction; and the consequences of assessment uses for EL students' designation needs to be continuously conducted.

Additionally, further validity research on the methods of standard setting and creating composite scores needs to be conducted; this is one of the pivotal issues related to the impact of assessment uses for schools and students that was raised in the Working Meeting. Comparability across consortia assessments and various ELP assessments is also an important validity research area. This area of research will provide important insights into EL classification and the definition of EL status.

Conclusion

The purpose of this paper was to summarize the important conversation that took place at the K–12 ELP Assessment Working Meeting on summative ELP assessments and to identify critical policy and research issues. The discussion from the Working Meeting informed the critical areas in need of additional attention from policy and research communities as next-generation ELP assessments are being designed and implemented. As noted, the paper focused on summative ELP assessments and does not address other important ELP assessment purposes (e.g., screeners, formative assessment, interim assessments, and classroom-based assessments).

One emerging theme from the meeting is that policy, practice, and research in this area are deeply linked and continue to inform one another to advance the assessment uses. This relationship is underscored by the diversity and sheer numbers of EL students across the United States, making the task of addressing ongoing issues such as the achievement gap and the definition of EL status both challenging and complex. While this paper has highlighted the challenges associated with measuring ELP in a diverse group of EL students, we would like to remind readers that these challenges can be better contextualized as opportunities. This shift in perspective is in recognition of the hard work practitioners, policymakers, and researchers have contributed — and are still contributing — to the goal of better serving the EL population. EL students and their families bring into our schools unique and important multilingual and cultural benefits and resources. Finding ways to capitalize on these benefits is an important endeavor needed to embrace and support EL students through shared goals and active communication across federal agencies, state agencies, local agencies, researchers, and practitioners.

Acknowledgments

The authors would like to thank the following for their review and valuable comments on earlier drafts of this paper: Nancy Doorey, Pat Forgione, Kenji Hakuta, Robert Linquanti, Alexis Lopez, Jennifer O'Day, Scott Paris, Don Powers, Charlene Rivera, and Joyce Wang. Special thanks goes to Pat Forgione and Nancy Doorey for organizing the English Language Proficiency Assessment Research Working Meeting and supporting the production of this paper. In addition, the authors would like to acknowledge Ian Blood, Kim Fryer, and Ayleen Gontz for their helpful editorial assistance. Any errors remain the responsibility of the authors. A previous version of this report was published in December 2014 on ets.org.

Notes

1 We use *EL designation* in this paper to refer to classification and reclassification of ELs. By the classification of ELs, we mean determining the students' ELP levels, and by reclassification of ELs, exiting the students from EL status.

2 For more information about Understanding Language, see http://ell.stanford.edu.

References

Abedi, J. (2006). Psychometric issues in the ELL assessment and special education eligibility. *Teachers College Record, 108*(11), 2282–2303.

Bailey, A. L. (Ed.). (2007). *The language demands of school: Putting academic English to the test*. New Haven, CT: Yale University Press.

Bailey, A. L., & Heritage, M. (2008). *Formative assessment for literacy: Building reading and academic language skills across the curriculum*. Thousand Oaks, CA: Corwin.

Bailey, A. L., & Heritage, M. (2014). The role of language learning progressions in improved instruction and assessment of English language learners. *TESOL Quarterly, 48*(3), 480–506.

Bailey, A. L., & Huang, B. H. (2011). Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing*, *28*(3), 343–365.

California Department of Education. (2012). California ELD standards. Retrieved from http://www.cde.ca.gov/sp/el/er/eldstandards.asp

Capps, R., Fix, M., Murray, J., Ost, J., Passel, J. S., & Herwantoro, S. (2005). *The new demography of America's schools: Immigration and the No Child Left Behind Act*. Washington, DC: The Urban Institute.

Cheuk, T. (2013). *Relationships and convergences among the mathematics, science, and ELA practices* [Refined version of diagram created by the Understanding Language Initiative for ELP Standards]. Stanford, CA: Stanford University.

Cook, H. G. (2005). *Aligning English language proficiency tests to English language learning standards*. Washington, DC: Council of Chief State School Officers.

Cook, H. G. (2014, June). *Research-based findings and recommendations to improve inferences and reporting*. Presentation given at the English Language Proficiency Assessment Research Working Meeting: Summative Assessments by the K-12 Center at ETS, Houston, TX.

Cook, H. G., Linquanti, R., Chinen, M., & Jung, H. (2012). *National evaluation of Title III implementation supplemental report: Exploring approaches to setting English language proficiency performance criteria and monitoring English learner progress*. Washington, DC: U.S. Department of Education.

Council of Chief State School Officers. (2012). *Framework for English language proficiency development standards corresponding to the Common Core State Standards and the Next Generation Science Standards*. Washington, DC: Author.

DiCerbo, P. A., Anstrom, K. A., Baker, L. L., & Rivera, C. (2014). A review of the literature on teaching academic English to English language learners. *Review of Educational Research*, *84*(3), 446 – 482.

Faulkner-Bond, M., Shin, M., Wang, X., Zenisky, A., & Moyer, E. (2013, April). *Score reports for English proficiency assessments: Current practices and future directions*. Paper presented at the annual conference of the National Council on Measurement in Education, San Francisco, CA.

Francis, D. (2014, June). *Conceptualizing the relationship between English-language proficiency and academic performance*. Presentation given at the English Language Proficiency Assessment Research Working Meeting: Summative Assessments by the K-12 Center at ETS, Houston, TX.

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

Gottlieb, M., Katz, A., & Ernst-Slavit, G. (2009). *Paper to practice: Using the TESOL ELP Standards in preK-12*. Alexandria, VA: Teachers of English to Speakers of Other Language.

Gottlieb, M., & Wilmes, C. (2014, June). *WIDA English language development standards and assessment system*. Presentation given at the English Language Proficiency Assessment Research Working Meeting: Summative Assessments by the K-12 Center at ETS, Houston, TX.

Guzman-Orth, D., Laitusis, C., Thurlow, M., & Christensen, L. (2016). *Conceptualizing accessibility for English language proficiency assessments* (Research Report No. RR-16-07). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12093

Guzman-Orth, D., & Nylund-Gibson, K. (2013, April). *When is an English language learner not an English language learner? Exploring individual differences in developmental language and literacy acquisition for at-risk learners: A latent transition approach*. Paper presented at the annual meeting of American Educational Research Association. San Francisco, CA.

Hakuta, K. (2011). Educating language minority students and affirming their equal rights: Research and practical perspectives. *Educational Researcher, 40*(4), 163 – 174.

Hakuta, K. (2014, June). *Summary and conclusions of the working meeting*. Presentation given at the English Language Proficiency Assessment Research Working Meeting: Summative Assessments by the K-12 Center at ETS, Houston, TX.

Hauck, M. C., Wolf, M. K., & Mislevy, R. (2016). *Creating a next-generation system of K-12 English learner (EL) language proficiency assessments* (Research Report No. RR-16-06). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12092

Hiebert, E. H. (2013). Supporting students' movement on the staircase of text complexity. *The Reading Teacher, 66*(6), 459 – 468.

Ho, M. (2014, June). *A new assessment system for English language learners*. Presentation given at the English Language Proficiency Assessment Research Working Meeting: Summative Assessments by the K-12 Center at ETS, Houston, TX.

Hopkins, M., Thompson, K. D., Linquanti, R., Hakuta, K., & August, D. (2013). Fully accounting for English learner performance: A key issue in ESEA reauthorization. *Educational Researcher, 42*(2), 101 – 108.

Linquanti, R. (2014, June). *Defining an "English proficient" performance standard: Current issues and opportunities*. Presentation given at the English Language Proficiency Assessment Research Working Meeting: Summative Assessments by the K-12 Center at ETS, Houston, TX.

Linquanti, R., & Cook, G. (2013). *Toward a "common definition of English learner": Guidance for states and state assessment consortia in defining and addressing policy and technical issues and options*. Washington, DC: Council of Chief State School Officers.

Linquanti, R., & Hakuta, K. (2012). *How next generation standards and assessments can foster success for California's English learners* (PACE Policy Brief No. 12-1). Stanford, CA: Stanford University.

Lopez, A. A., Pooler, E., & Linquanti, R. (2016). *Key issues and opportunities in the initial identification and classification of English learners* (Research Report No. RR-16-09). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12090

Lopez, A., & Wolf, M. K. (2014, June). *Conceptualizing and operationalizing the construct of standards-based ELP assessments*. Presentation given at the English Language Proficiency Assessment Research Working Meeting: Summative Assessments by the K-12 Center at ETS, Houston, TX.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 *et seq*. (2002).

Parker, C., Louie, J., & O'Dwyer, L. (2009). *New measures of English language proficiency and their relationship to performance on large-scale content assessments* (Issues & Answers Report No. 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. Retrieved from http://ies.ed.gov/ncee/edlabs

Rivera, C. (2014, June). *The promise of a new generation of ELP assessments*. Presentation given at the English Language Proficiency Assessment Research Working Meeting: Summative Assessments by the K-12 Center at ETS, Houston, TX.

Scarcella, R. (2003). *Academic English: A conceptual framework* (Technical Report No. 2003-1). Santa Barbara, CA: University of California, Linguistic Minority Research Institute.

Schleppegrell, M. (2004). *The language of schooling: a functional linguistics perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Schleppegrell, M. (2012). Academic language in teaching and learning. *The Elementary School Journal, 112*(3), 409 – 418.

Snow, C. E., & Uccelli, P. (2009). The challenge of academic language. In D. R. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy* (pp. 112 – 133). New York, NY: Cambridge University Press.

Thompson, K. (2012, April). *Are we there yet? An analysis of how long it takes English learners to be reclassified*. Paper presented at the annual meeting of American Educational Research Association, Vancouver, British Columbia, Canada.

Thompson, K. (2014, June). *Illustrating empirical relationships between English-language proficiency and academic performance*. Presentation given at the English Language Proficiency Assessment Research Working Meeting: Summative Assessments by the K-12 Center at ETS, Houston, TX.

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: Author.

Valdes, G., & Wong Fillmore, L. (2011, June). *Common core standards and English language learners: A conversation*. Paper presented at the National Conference on Student Assessment, Orlando, FL.

van Lier, L., & Walqui, A. (2012). *Language and the common core state standards. Understanding language initiative*. Retrieved from http://ell.stanford.edu/papers/language

Wolf, M. K., Everson, P., Lopez, A., Hauck, M., Pooler, E., & Wang, J. (2014). *Building a framework for a next-generation English language proficiency (ELP) assessment system* (Research Report No. RR-14-34). Princeton, NJ: Educational Testing Service.

Wolf, M. K., & Farnsworth, T. (2014). English language proficiency assessments as an exit criterion for English learners. In A. Kunnan (Ed.), *The companion to language assessment* (pp. 303 – 317). New York, NY: Wiley-Blackwell.

Wolf, M. K., Kao, J., Griffin, N., Herman, J. L., Bachman, P., Chang, S. M., & Farnsworth, T. (2008). *Issues in assessing English language learners: English language proficiency measures and accommodation uses — Practice review* (CRESST Technical Report No. 732). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Wolf, M. K., Wang, Y., Huang, B. H., & Blood, I. (2014). Investigating the language demands in the Common Core State Standards for English language learners: A comparison study of standards. *Middle Grades Research Journal, 9*(1), 35 – 51.

Working Group on ELL Policy. (2010). *Improving educational outcomes for English language learners: Recommendations for ESEA reauthorization*. Questions and answers. Palo Alto: Author. Retrieved from http://ellpolicy.org/wp-content/uploads/QA.pdf

Appendix A

List of Attendees
English Language Proficiency Assessment
Research Working Meeting
June 11, 2014

Jennifer Alexander
Houston Independent School District

Alison Bailey
UCLA

Sandra Alvear
Rice University

Noelia Cortez Benson
Northside Independent School District

Rosa Aronson
TESOL International Association

Tim Boals
ASSETS

Maureen Ayers
Cypress-Fairbanks Independent School District

Patricia S. Cantu
Alief Independent School District

Alma Cardenas-Rubio
Brownsville Independent School District

Maria S. Carlo
University of Texas Health Science Center

Emilio Castro
San Antonio Independent School District

Magda Chia
Smarter Balanced Assessment Consortium

Fen Chou
Council of Chief State School Officers

Yvonne Colmenero
Corpus Christi Independent School District

H. Gary Cook
Wisconsin Center for Education Research

J. Chris Coxon
Communities Foundation of Texas

Nancy Doorey
K – 12 Center at ETS

Peggy Estrada
University of California-Santa Cruz

Pat Forgione
K – 12 Center at ETS

David J. Francis
University of Houston

Luz Garcia-Martin
San Antonio Independent School District

Elizabeth Garza
Northeast Independent School District

Daniel Gohl
Houston Independent School District

Arlen Benjamin Gomez
New York State Education Department

Margo Gottlieb
ASSETS

Gracie Guerrero
Houston Independent School District

Danielle A. Guzman-Orth
ETS

Betsabe Haisler
Aldine Independent School District

Kenji Hakuta
Stanford University

Mark Hansen
UCLA

Maurice Cogan Hauck
ETS

Margaret Heritage
CRESST/UCLA

Margaret Ho
Washington Department of Education

Maria Hohenstein
Austin Independent School District

Mary Jadloski
Cypress-Fairbanks Independent School District

Julia Lara
National Education Association

Robert Linquanti
WestEd

Alexis Lopez
ETS

Giselle Lundy-Ponce
American Federation of Teachers

Carlos Martínez
U.S. Department of Education

Martha I. Martinez
Oregon Department of Education

Abdinur Mohamud
Ohio Department of Education

Scott Norton
Council of Chief State School Officers

Jennifer O'Day
American Institutes for Research

Cecilia Oakeley
Dallas Independent School District

Phil Olsen
Wisconsin Department of Public Instruction

John Oswald
ETS

Scott G. Paris
ETS

P. David Pearson
University of California, Berkeley

Berta Alicia Peña
Brownsville Independent School District

Delia Pompa
National Council of La Raza

Emilie Pooler
ETS
Justin Porter
Texas Education Agency

Sara Ptomey
Aldine Independent School District

Beatriz C. Ramirez
San Antonio Independent School District

Tamara Reavis
PARCC

Charlene Rivera
George Washington University

Lily Roberts
California Department of Education

John Segota
TESOL International Association

Karen Thompson
Oregon State University

Dora Torres-Morón

Dallas Independent School District
Maria A. Trejo

Cypress-Fairbanks Independent School District

Charlotte "Nadja" Trez
North Carolina Dept. of Public Instruction

Guadalupe Valdés
Stanford University

Kathleen Vanderwall
Oregon Department of Education

Xavier Vasquez
Fort Worth Independent School District

Vince Verges
Florida Department of Education

Dan Wiener
Massachusetts Department of Elementary and Secondary Education

Carsten Wilmes
WIDA Consortium

Mikyung K. Wolf
ETS

Santiago V. Wood
National Association for Bilingual Education

Nancy Zayas
Arlington Independent School District

Gloria Zyskowski
Texas Education Agency

Appendix B

Working Meeting Agenda
The English Language Proficiency Assessment
Research Working Meeting

AGENDA

7:30 – 8:30 a.m. ***Continental Breakfast***

8:30 – 8:35 a.m. ***Welcome and Introductions***
Pat Forgione, Executive Director, K – 12 Center at ETS

8:35 – 9:55 a.m. **Session 1:** ***Summative ELP Assessments for a New Era***
Moderator: Charlene Rivera, George Washington University

- Setting the Context — Charlene Rivera
- English Language Proficiency Re-examined in the Era of College- and Career-Ready Standards
- The ELP Assessment Consortia
- Carsten Wilmes, WIDA — ASSETS and Margo Gottlieb, ASSETS
- Margaret Ho, Washington DOE — ELPA21
- Conceptualizing the Construct of ELP
- Alexis Lopez, ETS
- Open Floor Discussion

9:55 – 10:10 a.m. ***Break***

10:10 a.m. – 12:15 p.m. **Session 1,** ***Continued***

- Research-Based Findings and Recommendations to Improve Inferences and Reporting — H. Gary Cook, University of Wisconsin
- Panel Discussion: Technical and Policy Implications
- Tim Boals, University of Wisconsin — ASSETS
- Kenji Hakuta, Stanford University — ELPA21
- Lily Roberts, California DOE
- Arlen Benjamin Gomez, New York DOE
- Tamara Reavis, PARCC
- Magda Chia, SBAC
- Table Work
- Open Floor Discussion

12:15 – 1 p.m. ***Lunch***

1 – 3:15 p.m. **Session 2:** ***Defining an "English Proficient" Performance Standard***

- Moderator: Robert Linquanti, WestEd
- Current Issues and Opportunities — Robert Linquanti, WestEd
- Conceptualizing the Relationship Between English Language Proficiency and Academic Performance — David Francis, University of Houston
- Illustrating Empirical Relationships Between English Language Proficiency and Academic Performance — Karen Thompson, Oregon State University
- Panel Discussion: Research and Policy Implications
- Dan Wiener, Massachusetts DOE, an ASSETS/PARCC state
- Phil Olsen, Wisconsin DOE, an ASSETS/SBAC state
- Martha Martinez, Oregon DOE, an ASSETS/Smarter Balanced state
- Abdinur Mohamud, Ohio DOE, an ELPA21/PARCC state
- Gloria Zyskowski, Texas DOE, an independent state
- Table Work

3:15 – 3:30 p.m.	***Break***
3:30 – 4 p.m.	**Session 2, *continued***
	• Report Out and Open Floor Discussion
4 – 4:30 p.m.	**Session 3: *Synthesis and Closing*** Moderator: Kenji Hakuta, Stanford University
	• Gathering of Major Take-Aways for Report to Policymakers — Facilitated by Kenji Hakuta, Stanford University
	• Closing Comments and Next Steps — Nancy Doorey, Director of Programs, K – 12 Center at ETS
4:30 p.m.	***Adjournment***

Suggested citation:

Wolf, M. K., Guzman-Orth, D., & Hauck, M. C. (2016). *Next-generation summative English language proficiency assessments for English learners: Priorities for policy and research* (Research Report No. RR-16-08). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12091

Action Editor: Donald Powers

Reviewers: Alexis Lopez and Joyce Wang