Assessment Services Supporting ELs through Technology Systems (ASSETS)

The Wisconsin Department of Public Instruction, in collaboration with the World-class Instructional Design and Assessment (WIDA) Consortium, proposes to develop a next-generation comprehensive and balanced assessment system for English learners (ELs). This project, known as *Assessment Services Supporting ELs through Technology Systems* (ASSETS), will leverage the considerable experience of WIDA as a consortium of states, as a collaborating partner with leaders in the field of academic English and EL assessment, and as a program known for delivering high-quality products and services focused on enhancing the educational opportunities of ELs in U.S. schools. ASSETS will result in an innovative technology-based assessment system that is (a) anchored in WIDA's established English language proficiency (ELP) standards that are aligned with the Common Core State Standards; (b) informed by rigorous ongoing research; and (c) supported by comprehensive professional development and outreach, all of which will be developed within the collaborative framework of a multistate consortium.

The WIDA Consortium, housed within the Wisconsin Center for Education Research at the University of Wisconsin–Madison, was originally established with funding from a U.S. Department of Education Enhanced Assessment Grant and currently includes 26 states and the District of Columbia.[1] Since 2003, WIDA has created and adopted comprehensive ELP standards (2004, 2007, in press) that represent the second language acquisition process and the language of the content areas of language arts, mathematics, science, and social studies. Based on these standards, WIDA has developed:

- A K–12 annual summative ELP test, Assessing Comprehension and Communication in English State-to-State for English Language Learners (ACCESS for ELLs®);
- An initial screener, the WIDA ACCESS Placement Test (W-APT™); and

[1] WIDA Consortium membership as of May 2011: Alabama, Alaska, Delaware, the District of Columbia, Georgia, Hawaii, Illinois, Kentucky, Maine, Maryland, Minnesota, Mississippi, Missouri, Montana, New Hampshire, New Jersey, New Mexico, North Carolina, North Dakota, Oklahoma, Pennsylvania, Rhode Island, South Dakota, Vermont, Virginia, Wisconsin, and Wyoming.

- An on-demand, "off-the-shelf" test of ELP known as WIDA MODEL™ that can be used for placement or for interim assessment.

In addition to its standards and assessments, WIDA pursues a research agenda on behalf of member states. WIDA research explores not only the validity of the assessments, but also areas of interest such as ELP growth rates, correlations between ELP tests and academic tests, and classroom implementation of the ELP standards. Concurrently, WIDA provides extensive professional development opportunities and maintains a comprehensive website (www.wida.us).

The ASSETS project builds on WIDA's successful three-part approach, as illustrated in Table 1.

Table 1

Foundations of WIDA's Successful Approach

Elements of approach	**WIDA Consortium**	**ASSETS**
Consortium	Consortium members voluntarily join together to address EL needs while satisfying state/federal requirements. All members have a voice and benefit from collaboration.	The ASSETS next-generation ELP assessment system will be designed with collaborative input from consortium members to meet their needs.
Integrated systems	WIDA's ELP test—ACCESS for ELLs®—is integrated within an aligned assessment system comprising comprehensive ELP standards, multiple assessments, professional development, and EL-related research.	ASSETS will integrate technology-based assessments and professional development in an innovative and comprehensive system that corresponds with state academic standards, including the Common Core and is compatible with other academic assessment systems.

Elements of approach	WIDA Consortium	ASSETS
"Can do" philosophy	Underlying WIDA's products and services is the belief that developing language proficiency is about tapping children's capacity—not overcoming their limits.	Underlying ASSETS is the belief that valid, reliable assessment equips educators to help students develop language, access grade-level content, and reach college and career readiness.

On March 30, 2011, representatives from 22 states—19 current WIDA member states and 3 additional states—met to plan for the next-generation standards and assessment system. This proposed ASSETS project represents the consensus reached at that meeting on the direction to be taken moving forward. In pursuing these directives, WIDA[2] will take the following steps:

- *Establish consortium structure.* WIDA will give representatives of the respective state educational agencies (SEAs) a voice in the design, implementation, governance, policymaking, and other relevant issues pertaining to the ASSETS assessment system. SEA decision making will include, but not be limited to, establishing a common definition of *English learner*, approving the assessment design, determining accommodations policies and procedures, and establishing data-reporting criteria.
- *Develop ELP assessments.* WIDA will develop, pilot, field-test, and finalize ELP assessments that use technology to allow for (a) more authentic language assessment tasks, including performance-based tasks for all language domains (listening, speaking, reading, and writing); (b) timelier—in some case,

[2] The original WIDA Consortium and the collaborative partnership created for the purpose of realizing the goals of the ASSETS project (WIDA-ASSETS) are different but overlapping entities. For simplicity, we generally use *WIDA* in this proposal to refer to both entities. Members of WIDA-ASSETS include the Wisconsin Department of Public Instruction, the consortium member SEAs, the WIDA project team at the Center for Applied Linguistics, and the WIDA project at the Wisconsin Center for Education Research (*WIDA Central*). In the future, the ASSETS project consortium members may elect to create a new name.

instantaneous—score reporting; (c) reduced burden on test administrators; and (d) compatibility with content-driven assessment systems, including those of the Partnership for the Assessment of Readiness for College and Careers and the SMARTER Balanced Assessment Consortium, as well as with individual state achievement measures. The new assessments to be developed include:

- *A computer-based summative test.* The summative test—to be administered annually in Grades K–12 for accountability and program purposes—will cover (a) the language domains of listening, speaking, reading, and writing and (b) the five WIDA ELP standards, encompassing social and instructional language and the language of language arts, mathematics, science, and social studies. This test will include (a) selected-response options for listening and reading that are machine-scored and (b) constructed-response options for speaking and writing that are digitally recorded or handwritten on paper and centrally scored by human raters.
- *A computer-based on-demand diagnostic (screener) test.* The screener test will be used to determine eligibility for EL services and program placement within those services. The test format will be derived from the summative test, with all scoring done locally.
- *Computer-based classroom benchmark assessments.* A series of benchmark assessments will be organized by language domain and standard at five grade-level clusters: 2, 5, 7–8, 9–10, and 11–12. In the speaking and writing domains, item and task types will be similar to those for the summative and screener tests. The listening and reading domains, however, will incorporate innovative item types, including performance tasks, and innovative response spaces that allow for partial- and full-credit scoring. These benchmarks will provide immediate feedback.
- *Formative assessment.* WIDA will develop a foundation for the formative assessment process to be used by classroom teachers. This process will include the design of language learning progressions corresponding to college and career readiness standards for *incorporation* into instructional assessment for ELs.

- *Create a training program for scorers.* WIDA will create, pilot, and field-test an adaptation of the Multimedia Rater Training Program (MRTP). Developed by our test development partner, the Center

for Applied Linguistics, the MRTP is an interactive software program designed to teach professionals to rate oral language proficiency. The adapted MRTP will provide intensive, on-demand training and practice in scoring speaking and writing. It will be used by educators for scoring the screener and benchmark assessments.

- *Create professional development and outreach materials.* WIDA will develop, pilot, field-test, and finalize materials and methods for (a) professional development on implementation of the assessment system, including appropriate and effective use of assessment results, and (b) outreach to stakeholders, including families, policymakers and researchers.
- *Conduct evaluation.* WIDA will evaluate the assessments and professional development using industry-approved practices and standards in psychometrics, quality control procedures, and qualitative and quantitative research methodologies.
- *Plan for scale-up and sustainability.* WIDA will establish a plan for scaling up the new assessment system and sustaining it beyond the grant period. The plan will cover (a) procuring a post-grant technology-based platform provider and scoring partner through a request for proposals process in accordance with state procurement rules; (b) devising a consortium governance structure that is sustainable for the long term; and (c) working with states to ensure access to the standards, assessments, professional development, and research results.

The ASSETS project represents a critical first step in creating the next generation EL assessment system, which includes standards that correspond to college and career readiness standards, a complete suite of research-based assessments, professional development that is centered on the needs of ELs by focusing on building educators' knowledge and skills, data management systems that allow for meaningful analysis, and research that is timely, actionable and supports ELs. This system maintains and enhances large-scale summative assessment in a technology environment, but more critically, it introduces on-demand, targeted, standards-based benchmark assessments to the classroom that, together with formative assessment processes and resources, can have a powerful and immediate impact on language teaching and learning. WIDA fully expects that with direction from consortium members, the

assessments, professional development and research created and conducted under this grant will continue to improve and expand, offering educators more resources to serve the needs of their students and to guide program development and educational policy. As an example of the anticipated continual improvement to the system, Table 2 illustrates WIDA's vision of how the assessments will evolve from WIDA's current offerings, to what will be developed by the end of the grant, and what is likely to be further developed after the grant period as the system becomes operational.

Table 2

Assessment System: Current, End-of-Grant, and Long Term Post-Grant

	Summative assessment			On-demand screener			Benchmark assessments		
Features	**C**	**G**	**PG**	**C**	**G**	**PG**	**C**	**G[b]**	**PG[c]**
Paper-based	✓	✓	✓[a]	✓	✓	✓[a]	✓		
Computer-based		✓	✓		✓	✓		✓	✓
Semi-adaptive					✓				
Adaptive			✓			✓			
Innovative item types			✓			✓		✓	✓

Note. C = current system. G = system at end of grant period. PG = post-grant system.

[a]Paper-based version available as an accommodation. [b]In selected domains/standards for Grades 2, 5, 7–12. [c]In all domains/standards for all grades.

Absolute Priorities

The ASSETS project will address all five absolute priorities, as discussed throughout this proposal. Here, we summarize briefly:

- *Absolute Priority 1—Collaborations.* ASSETS represents a collaboration among the current WIDA Consortium member states, including Wisconsin; several additional (non-WIDA) states; the Wisconsin Center for Education Research (WCER); the Center for Applied Linguistics (CAL); Data Recognition Corporation (DRC); MetriTech, Inc.; the National Center for Research on Evaluation,

Standards, and Student Testing (CRESST) at UCLA; WestEd; and individual expert consultants. Representative educators from local educational agencies (LEAs) and schools will also participate in aspects of this project.

- *Absolute Priority 2—Use of multiple measures.* The ASSETS project will develop multiple measures of student progress in learning English through several types of assessments and resources mentioned above and detailed below.
- *Absolute Priority 3—Charting student progress.* The assessments will be developed so that the resulting data can be used to chart student progress over time (a) at local classroom and school levels to guide curriculum and instruction, (b) at SEA and LEA levels for accountability purposes, and (c) at the consortium level to inform the field and policymakers.
- *Absolute Priority 4—Comprehensive academic assessment instruments.* ASSETS will result in a system of comprehensive academic language assessment instruments that leverage technology to assess authentic language development more accurately than paper-based tests. The assessment system will include a screener, an annual summative test, periodic benchmark tests and resources for formative assessment.
- *Absolute Priority 5—English language proficiency assessment system.* The ASSETS ELP assessment system will be anchored in WIDA's existing ELP standards (aligned with the Common Core State Standards) and include multiple types of new high-quality assessments designed to (a) monitor student progress, inform instruction, and provide accountability measures; (b) yield actionable data; (c) be compatible with states' assessment systems; and in conjunction with other WIDA resources, (d) provide for the inclusion of all ELs, including students with severe cognitive disabilities.

Competitive Preference Priority

Since 2003, WIDA has grown as a consortium with a governance structure that allows for significant SEA input and an open communications policy that has served it well. The governance structure of the new WIDA-ASSETS Consortium will be similar. This new consortium currently includes twenty-four

states (21 current WIDA Consortium member states and 3 non-WIDA states) from whom WIDA has received signed Memoranda of Understanding.

The new consortium will have two types of members: *advisory members* (SEAs involved with more than one consortium under this grant competition) and *governing members* (SEAs committed only to WIDA-ASSETS). Only governing members will be able to participate in final policy decisions. The goal for final decisions will be consensus, but a simple majority vote will be enough to set policy in most instances. Operational decisions will be made by WIDA Central, the project management partner. WDPI and WCER will manage grant funds.

A subcommittee of the governing states will form a Steering Committee, to be chaired by the representative from the Wisconsin Department of Public Instruction. The role of the Steering Committee will be to provide researchers and test developers with direction and advice to ensure that products and services meet the needs of the states and the requirements of the law. The Steering Committee will also advise WIDA Central on operational decisions. Additional subcommittees may be formed as needed to guide the work of the consortium.

Any state will be able to join the consortium by agreeing to be bound by all statements and assurances in the grant application and executing a memorandum of understanding making the required assurances for adopting and using project products. Advisory members will be able to upgrade their membership status by changing their involvement with other consortia. Member states will be permitted to leave the consortium for any reason during the project period, upon U.S. Department of Education approval.

The timeline for key decisions and project implementation will established by the project plan. The Steering Committee will research and prepare, using working groups as necessary, all policy decisions and required definitions for a full vote by the governing states.

1. THEORY OF ACTION

Modern conceptualizations of test validity center on the *use* made of assessments results (e.g., American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, p. 9). In Section 4, we present our

approaches to validation based on Bachman's (2005) assessment use argument. Here, we simply state (a) the intended uses to be made of each of the four components of our proposed assessment system; (b) the way in which this assessment system can be incorporated into coherent educational systems; and (c) the way in which these educational systems will improve academic achievement for ELs.

The first element in the assessment system is the *screener*. Scores from the screener will primarily be used to identify students as ELs. The screener will also provide a preliminary determination of proficiency as defined by the WIDA ELP standards (Section 2.2) for use in classroom or course placement, initial grouping of students, and progress monitoring.

Students found eligible for services will take the *summative assessment* annually. This assessment will provide a fair, valid, and reliable measure of student performance in the four domains of reading, writing, speaking, and listening, together yielding a comprehensive ELP score. Scores from this assessment are intended for use at state, LEA, school, classroom, and student levels to chart student progress over time, inform decisions about when students should exit English language support programs, and help determine school, LEA, and state effectiveness for accountability purposes. At the local level, scores may serve as one component of principal and teacher evaluations and as an indicator of needed professional development and support.

The *benchmark assessments* will enable schools to chart student progress in finer increments and with more precision than the summative test. Scores from the benchmark assessments are intended for use at LEA, school, classroom, and student levels to monitor language progress on an ongoing basis, to differentiate language instruction or regroup students during the school year, and to inform teachers' collaborative lesson planning and design. At program and district levels, the results will be able to be used to determine ELP benchmarks, contribute to program and district accountability, and inform evaluations of education programs.

In creating these instruments, WIDA will identify key assessable academic language, drawn from the linguistic (e.g., vocabulary, text complexity, forms and conventions) and sociocultural elements in the WIDA ELP standards. Based on these constructs, WIDA, in collaboration with researchers at UCLA, will

then develop language learning progressions, the foundation of the *formative assessment* resources. Ongoing formative assessment will yield meaningful feedback to students and actionable data to teachers so that they can collaborate, adjust their curriculum and instruction, and communicate with families, board members, and other stakeholders in the community.

The ASSETS assessment system will give our consortium states the opportunity to incorporate high-quality assessments and assessment results into their educational systems to improve teaching, learning, and language instruction programs. Through its link to college- and career-ready standards, the assessment system will complement the overall educational system and serve as a pathway for EL access and participation. Each assessment will be designed for use (a) *vertically* with other components of the assessment system to provide a better picture of student language performance and (b) *horizontally* with other components of the overall educational system to contextualize academic achievement data and more accurately profile annual, interim, or ongoing student performance. The fundamental goal of the system, however, is to provide actionable data that leads to improved student outcomes in the belief that when educators know what students can do, they are better equipped to guide those students in building their language base to access grade-level content to reach college and career readiness.

The assessment system will also provide a basis for consortium-led research that will provide data to improve EL achievement nationally. The screener could be used for validity studies of home language surveys for EL identification and determination of eligibility for language support. The summative assessment will provide data for predictive validity studies of performance on achievement measures, determination of the time required for subgroups of ELs to gain full English proficiency, and comparability studies of computer- and paper-based forms. The benchmark data will be useful in predictive validity studies of the summative assessment, reliability studies of performance tasks, research on the use of language data for decision making, and evaluation of language education programs. The formative assessment process will yield data for studies of classroom use of ELP standards, implementation of academic language progressions, and teachers' perceptions of student performance. The combined data will also contribute to evolving definitions of *English learner* and to our

understanding of academic language and its impact on academic achievement. Finally, research using data generated by the proposed assessment system will shed light on the academic achievement of ELs and the field of second language acquisition more generally.

2. ASSESSMENT DESIGN

2.1 Number and Types of Assessments

Building on the solid foundation of its existing annual summative assessment—ACCESS for ELLs®—and the accompanying on-demand screener, the *W-APT™*, WIDA will develop three types of next-generation assessments as part of the ASSETS assessment system:

1. A computer-based annual *summative assessment*;
2. An accompanying computer-based on-demand *screener*; and
3. Computer-based classroom *benchmark assessments*.

In addition, WIDA will undertake much-needed research to begin developing academic English language learning progressions as the foundation for teacher-friendly *formative assessment* resources. Together, the annual summative assessment, screener, benchmarks, and formative assessment resources will constitute a comprehensive set of assessment tools.

To ensure that all consortium members can make a smooth transition to computer-based testing, paper versions of the annual summative assessment and on-demand screener will also be available during the grant period, incorporating appropriate changes to make them as parallel as possible to the next-generation assessments.

By the end of the grant period, WIDA will have parallel computer- and paper-based annual summative and on-demand screener tests in five grade-level clusters (kindergarten, 1–2, 3–5, 6–8, and 9–12), across all five WIDA standards and proficiency levels and in all four language domains (see Section 2.2). Multiple innovative computer-based benchmark tests will be available for Grades 2, 5, 7–8, 9–10, and 11–12 organized by standard, language domain, and proficiency level. To further guide educators, language learning progressions will be drafted to inform the formative assessment resources.

2.2 Assessment of Relevant Standards

The WIDA Consortium preK–12 ELP standards, currently being implemented in 28 states and the District of Columbia, are at the foundation of the proposed project. Grounded in scientifically based research (August & Shanahan, 2006; Francis, Lesaux, Kieffer, & Rivera, 2006; Genesee, Lindholm-Leary, Saunders, & Christian, 2006), linguistic theory (Bailey, 2007; Schleppegrell, 2004; Scarcella, 2003), and best educational practices for ELs (Chamot, 2009; Echevarria, Vogt, & Short, 2008; Freeman, Freeman, & Mecuri, 2002), WIDA's comprehensive ELP standards publications (2004, 2007, in press) illustrate pathways for ELs to become fully proficient in both social and academic English. The WIDA ELP standards, internationally referenced (Gottlieb & Jones, 2008), encompass the language needed for success in school and beyond. They are organized as follows:

- *Five grade-level clusters:* preK–K, 1–2, 3–5, 6–8, and 9–12[3]
- *Five language standards:* social and instructional language; the language of language arts; the language of mathematics; the language of science; and the language of social studies
- *Four language domains:* listening, speaking, reading, and writing
- *Five ELP levels:* 1–Entering, 2–Beginning,[4] 3–Developing, 4–Expanding, and 5–Bridging

The WIDA ELP standards are represented by strands of *model performance indicators* (MPIs) that form matrices for each grade-level cluster. The consortium's existing annual summative assessment, ACCESS for ELLs®, directly measures ELs' ELP in relation to the WIDA standards. All items on the test are designed to allow ELs to demonstrate meeting the standards' MPIs at specified performance levels.

The WIDA ELP standards have been shown to correspond to the academic content standards of all consortium member states, as well as to the Common Core State Standards. An independent alignment

[3] The fall 2011 edition of the WIDA ELP standards will represent the standards by grade level (K, 1, 2, 3, 4, 5, 6, 7, 8, 9–10, 11–12) to mirror the structure of the Common Core State Standards.

[4] The *beginning* level will be changed to *emerging* in the fall 2011 edition of the WIDA ELP standards.

study by the E-TEAM evaluation group (2011) reported "adequate linking across all grade clusters" (p. 1) between the WIDA ELP standards' MPIs and the Common Core State Standards in English language arts (reading, writing, speaking, and listening) and mathematics. Additionally, the new edition of the WIDA ELP standards, to be published fall 2011, uses individual grade-level examples to explicitly connect the WIDA ELP standards and the Common Core State Standards, topically and linguistically, for educators. Based on the results of the E-TEAM study as well as WIDA's ongoing individual state alignment studies offered to new member states, the WIDA ELP standards can be said to reliably represent the language of the content taught in any state that might adopt the consortium's assessments.

In developing the next generation of ELP assessments, the goal is to continue to create items and performance tasks that allow ELs to demonstrate achievement of the WIDA ELP standards' MPIs as students move toward readying themselves for college and careers.

2.3 Required Student Performance Data

The ASSETS next-generation assessment tools will produce all of the required student performance data described in Absolute Priority 5—and more besides. As with WIDA's current ACCESS for ELLs® test, the next-generation *annual summative assessment* will provide fair, valid, and reliable measures of student ELP in the domains of reading, writing, speaking, and listening, which will be combined to form a comprehensive ELP score. As with ACCESS, these scores will be provided as scale scores on a vertical K–12 scale and as interpretive proficiency-level scores that, for each grade level, show the relationship between scores and proficiency levels as defined by the WIDA standards. Educators will be able to use these scores to chart student progress in learning English over time, to inform decisions about whether an individual student should exit from English language instruction educational programs, and to help determine school, LEA, and state effectiveness for accountability purposes. At a more local level, scores may be used, as appropriate, as one of multiple measures to inform principal and teacher evaluations, as an indicator of needed principal and teacher professional development and support, and, together with information from the benchmark assessments and good formative assessment practices, as a tool for identifying strategies to improve teaching, learning, and language instruction education programs.

Student performance on the *on-demand screener* will also be interpreted in terms of the proficiency levels defined by the WIDA ELP standards and will primarily be used for the identification of students as ELs. Performance on the screener will (a) represent a fair, valid, and reliable measure of whether a student's current level of English proficiency is above or below that for identification as an EL, and (b) if below, enable identification of a preliminary level of proficiency as defined by the WIDA ELP standards.

Student performance on the individual *classroom benchmark assessments* will be based on complex demonstrations of comprehension and production of academic English language. Each short benchmark assessment will be designed to assess attainment of MPIs within a language domain (listening, speaking, reading, or writing), in one or two of the five WIDA ELP standards (social and instructional language, the language of language arts, the language of mathematics, the language of science, and the language of social studies), and at a particular proficiency level defined by the WIDA performance levels (1, 2, 3, 4, or 5). The benchmark assessments will provide teachers and local school districts with evidence of student attainment of the standards or, alternatively, feedback on additional needs. They will be useful in checking a student's current proficiency level or marking progress toward the next higher proficiency level. The benchmark assessments will thus give educators multiple measures of student ELP development from multiple sources. They will also constitute comprehensive academic assessment instruments that are performance- and technology-based.

Data on student progress in attaining English proficiency will be disaggregated by EL subgroups such as (a) ELs by years in a language instruction educational program, (b) ELs whose formal education has been interrupted, (c) students who were formerly ELs by years out of the language instruction educational program, (d) ELs by level of English proficiency, such as those who initially scored proficient on the ELP assessment, (e) ELs by disability status, and (f) ELs by native language. The types of data to be produced are described in Section 2.5.

2.4 Availability of Student Data

The introduction of computer-based testing to the large-scale, secure *annual summative assessment* testing program provides several opportunities to improve the scoring turnaround time, from

approximately 8 weeks after the close of a state's testing window with the current ACCESS for ELLs® to as little as 4 weeks. Specifically, computer-based testing will make possible the immediate delivery of test responses to the central office where they will be scored, eliminating delay caused by scanning student responses or digitizing student writing or speaking performances and enabling continuous distributed scoring. Moreover, embedding field-test items in a technology-based assessment is easier than the current field-testing practices, given the 44 different paper forms of ACCESS, and thus pre-equating of refreshed items will also improve turnaround time. Nevertheless, we acknowledge the need for strict quality control of all summative data, including the opportunity for states to review the data before it is publically released, and thus a window of 2–4 weeks is imagined.

For the *on-demand screener*, although the listening and reading portions will be computer-scored, the writing and speaking (performance-based) portions will be locally scored by educators. Scores will be available as soon as local scoring is complete. To improve scorer performance, we are proposing to develop computer-based training programs adapted from the Multimedia Rater Training Program developed by our test development partner, the Center for Applied Linguistics (CAL). We are also proposing to use computer-assisted scoring based on CAL's Computerized Oral Proficiency Instrument. More information on these proposed computer-based tools is provided in Section 2.10.

The *classroom benchmark assessments* for listening and reading will provide immediate scores and feedback useful to students and teachers. As with the screener, performance on writing and speaking tasks will be scored by teachers trained with a computer-based training program and assisted with a computer-based scoring program. These scores will be immediately useful to inform and guide instruction. In addition, performance on the speaking and writing benchmark assessments may be digitally stored and thus enable a portfolio approach to evidencing progress in developing academic language proficiency in those two language domains.

2.5 Types of Student Data

For each student, the data produced by the proposed assessments will meet the requirements of Absolute Priority 5(c). The main data from the annual summative assessment will be *raw scores* by

domain; *scale scores* on a K–12 vertically aligned (within-domain) scale; and a grade level–specific *proficiency-level score* that interprets a student's performance in terms of the proficiency levels defined in the WIDA standards. Vertical scaling makes it possible to measure progress in terms of scale scores as students move across both *grade-level clusters* and *tiers within clusters*. Proficiency-level scores will be presented as whole numbers followed by a decimal (e.g., 2.3) as they are currently for ACCESS for ELLs®. The whole number indicates the student's language proficiency *level* based on the WIDA ELP standards; the decimal indicates the *proportion* within the proficiency-level range that the student's scale score represents, rounded to the nearest 10^{th}.

Following ACCESS for ELLs®, composite scores on the new summative assessment will be derived from weighted scale scores from the four language domains. The *oral language composite* will be equally weighted from the listening and speaking domains. The *literacy composite* will be equally weighted from the reading and writing domains. Following policy guidelines from the current WIDA member states, the *overall composite* will be weighted to reflect a greater emphasis on reading and writing (i.e., listening 15%, speaking 15%, reading 35%, writing 35%) unless the new consortium for the ASSETS project agrees upon different criteria for establishing an overall composite score.

Primary data from the on-demand screener will be an *overall proficiency-level score* that can be used for EL identification and placement in services. Since the screener will be much shorter than the annual summative assessment, the overall proficiency level will be the most psychometrically reliable result. The screener will provide initial proficiency-level scores for each domain that may prove helpful in determining students' English language support needs and in making tier-level assignments on the annual summative assessment.

Results of the short classroom benchmark assessments will be interpreted primarily in terms of the evidence they provide of attainment of MPIs for the grade level, language domain, and standard(s) targeted by the particular benchmark assessment administered.

Table 3 summarizes the types of scores the new assessments will provide.

Table 3

ASSETS Assessments: Types of Scores to Be Produced

	Type of score			
	Raw	Vertical scale	Interpretive ELP	Targeted feedback
Domains (L, S, R, W)	A, S, B	A	A, B	B
Oral composite		A	A, S, B	
Literacy composite		A	A, S, B	
Overall composite		A	A, S, B	

Note. L = listening. S = speaking. R = reading. W = writing. A = annual summative assessment. S = on-demand screener. B = classroom benchmark assessments. Oral composite = 50% L, 50% S. Literacy composite = 50% R, 50% W. Overall composite = 15% L, 15% S, 35% R, 35% W.

The ASSETS Steering Committee and Technical Advisory Committee (see Section 7.5), with consultation from WestEd's Robert Linquanti (Section 7.5), will advise the project on what types of data to collect in addition to test scores, as well as how best to disaggregate groups within the EL population (see Section 2.3) and how these data might be used ultimately to inform policy and improve English language teaching and learning.

2.6 Uses of Student Data

WIDA is committed to making data available—to the degree permissible under the Family Educational Rights and Privacy Act of 1974 (FERPA; 20 U.S.C. § 1232g)—to relevant stakeholders to guide decisions about individual student achievement, program effectiveness, and professional development needs, as well as to inform teaching and learning for ELs generally. To assist in this endeavor, WIDA supports a SQL server–based comprehensive data warehouse of all available assessment and assessment-related information collected from WIDA states. WIDA's data warehouse includes not only ACCESS for ELLs® assessment data, but also data from selected research data collections of the National Center for Education Statistics: the Common Core of Data (CCD), the Schools and Staffing

Survey (SASS), and the National Assessment of Educational Progress (NAEP). These data sets are merged using the CCD identifier for schools, districts, and states. The data warehouse annually combines assessment and national data collection data and creates a longitudinal data system with unique student identifiers. Currently, the data warehouse houses 25 states' data across 6 years with more than 1.5 million student records tracked longitudinally. WIDA's data warehouse is designed to facilitate integration of additional measures, teacher and administrator information, or other relevant educational data.

To support member states, districts, and schools, WIDA has completed piloting a comprehensive online data dashboard of relevant WIDA state assessment data drawn from the data warehouse. WIDA's data dashboard has static assessment information related to ACCESS for ELLs® as well as state and WIDA-wide demographic information (e.g., number and percentage of students by cluster and native language). The dashboard supports two forms of longitudinal data for all domains and clusters from states: *mean growth rates* and *percentile growth charts*. Additionally, it provides state and national NAEP data in the areas of reading, writing, science, and mathematics. The data dashboard is designed to support program improvement, build administrators' and teachers' capacity to access and use EL-related information, and ultimately improve student learning.

The data dashboard is scheduled to go live in June 2011. WIDA is in the process of developing a comprehensive professional development program on how to use the data dashboard and its information to support program improvement and student achievement. Professional development offerings on the dashboard will be available to member states in the fall of 2011. Although outside the scope of this grant, the data dashboard and its accompanying professional development will be expanded to include new types and additional data created through the ASSETS project and beyond.

2.7 Frequency and Timing of Assessment Administration

As seen in Table 4, the large-scale annual *summative assessment* will be administered to students once a year during a state's testing window. This schedule will allow states to collect suitable data in a standardized fashion for federal accountability purposes. The *on-demand screener* will typically be administered only once to a student who is entering a local academic program, for the purposes of

identification as an EL and initial placement into a proficiency level; it is not intended for multiple administrations to the same student. *Classroom benchmark assessments* will be used by educators on an on-demand basis. As part of this grant, professional development materials will be created to help educators make the best use of the benchmark assessments as they guide students toward achieving targeted goals. The proposed *formative assessment resources* will be available on an ongoing basis, giving educators insight into how to assess students continually during the education process.

Table 4

Frequency and Timing of Assessment Administration

	Frequency		
Timing determined	**1/yr**	**Once**	**On demand**
By state	A		
Upon entrance to local program		S	
By district, school, educator			B

Note. A = annual summative assessment. S = on-demand screener. B = benchmark assessments.

2.8 Number and Types of Items

Because the proposed project for new assessments will build on WIDA's experience with ACCESS for ELLs®, we provide some background on that assessment here. Currently, ACCESS is available only in paper format. Drawn from the MPIs, ACCESS incorporates all five standards and ELP levels in sections that correspond to the four domains. The target administration times for each section of the test for Grades 1–12 are: Listening: 20–25 minutes; Reading: 35–40 minutes; Writing: Up to 1 hour; and Speaking: Up to 15 minutes. The kindergarten test is individually administered and takes 40 minutes on average.

The goal of the ACCESS test is to allow students to demonstrate their level of proficiency by demonstrating mastery of the MPIs. However, there are far too many MPIs to present to any single test taker within a reasonable testing session. To reduce the test burden, ACCESS presents test items in three

tiers (A, B, and C) for each grade level cluster. Tier A targets Proficiency Levels 1–3; Tier B, Levels 2–4; and Tier C, Levels 3–5. The tiers overlap to ensure that each is measuring to a common proficiency scale.

The ACCESS test battery is a collection of assessment instruments administered to all ELs across all grades and all proficiencies. Each test form consists of a set of *thematic folders*, or parts, generally containing three items each. This arrangement is intended to give students a context for items, minimizing the cognitive leaps they must make in transitioning from items in one area (e.g., language of math) to items in the next. Because this format has been successful in operationalizing the MPIs of the standards, it will serve as a starting point for work on the ASSETS next-generation assessments. Below, we discuss each of the new assessments in turn.

Annual Summative Assessment

During the grant period, research, development, and operationalization of innovative computer-based item types will focus on the classroom benchmark assessments (discussed below). Once those item specifications have proven stable, the innovative item types will be migrated up to the annual summative assessment. In the interim, with the exception of speaking, the foundational specifications for the item types on the annual summative assessment will remain the same as those for ACCESS for ELLS®. The rationale for this is twofold: (a) the current item types have been very successful and well-liked by the current WIDA Consortium states, having shown through research to be measuring the construct of interest, and (b) during the transition period, it will be critical to keep the computer- and paper-based versions of the annual summative assessment strictly comparable.

This does not mean, however, that the item specifications for the annual assessment are static. Since its inception, the ACCESS test has been research-based. Every year, one third to one half of the items are refreshed (with the performance-based assessment tasks refreshed much more frequently). In a cycle of continual research, item refreshment begins with refinements to item specifications based on what has been learned through research on the test and in the field since the last refreshment. During the transition period, we envision that both versions will follow the same specifications, reflecting the strengths and limitations of each. In other words, the listening and reading test items on the annual assessment will

remain selected response. By the end of the grant, the entire summative assessment will have been refreshed at least once, with parts of it having undergone two refreshment cycles.

We will develop new specifications for the computerized speaking test for Grades 1–12 modeled on the Center for Applied Linguistics' Computerized Oral Proficiency Instrument. While the current test is administered in a one-on-one, face-to-face interview format and scored by the administrator during the assessment, the new specifications will call for a more task-based format. These new specifications for speaking will also be used for the on-demand screener and classroom benchmark speaking assessments.

Unlike the Grade 1–12 assessments, the kindergarten assessment will remain an individually administered assessment. However, technological enhancements will be used to help with the logistics of its administration (Section 2.9).

Tables 5 and 6 outline the number and types of forms and the number of items per form for the new computer-based annual summative test.

Table 5

Annual Summative Assessment: Items and Scoring

	Listening and reading		**Writing and speaking**	
	K	**Grades 1–12**	**K**	**Grades 1–12**
No. test forms	1	3/cluster (Tier A–C)	1	3/cluster (Tier A–C)
Item type	SR	SR	ECR	ECR
Administration	Individual	Group	Individual	Group
Scoring	TA	CS-M	TA	CS-HR

Note. SR = selected response. ECR = extended constructed response. TA = test administrator. CS-M = centrally scored by machine. CS-HR = centrally scored by human raters.

Table 6

Annual Summative Assessment: Number of Items/Tasks per Test Form/Domain (Grades 1–12)

Tier	Listening	Reading	Writing	Speaking	Total
A	18 SR	24 SR	3 ECR	12 ECR	42 SR, 15 ECR
B	21 SR	27 SR	3 ECR	12 ECR	48 SR, 15 ECR
C	21 SR	27 SR	3 ECR	12 ECR	48 SR, 15 ECR

Note. SR = selected-response items. ECR = extended constructed-response tasks.

On-Demand Screener

The on-demand screener will follow the format of the WIDA MODEL™, which was refined through a research project involving iterative cognitive labs and then field-tested with several hundred students in each grade-level cluster. Shorter than the annual summative assessment, the screener will follow a semiadaptive format in both the computer-based and the individually administered mode. Because it is on-demand and can be given at any time, it is designed to be administered one-on-one.

The speaking portion of the Grade 1–12 screener, administered first, consists of eight performance-based tasks at progressively higher proficiency levels. The listening portion is administered next in a two-step process: a student's performance on four items in Step 1 determines the level at which he or she continues the test: low, mid, or high. The writing portion, administered third, consists of a one-minute task designed to assess whether the student has any English writing proficiency. A student who demonstrates writing ability will be presented with one extended constructed-response task. Reading is administered last and in the same manner as the listening test, with a Step 1 followed by one of three levels (low, mid, and high) of Step 2.

Table 7 presents the maximum number of items and performance tasks any student will be administered on the on-demand screener. The test is designed to take 40–90 minutes to administer, with more proficient students requiring more time to demonstrate their full level of proficiency on the extended constructed-response tasks. The longer period of time for high-proficiency students is necessary because

the identification of students as ELs is a high-stakes decision, requiring ample evidence that a student is not in need of English language support.

Table 7

On-Demand Screener: Number of Items/Tasks per Domain (Grades 1–12)

	Speaking	**Listening**	**Writing**	**Reading**
Type of response	S/ECR	SR	S/ECR	SR
No. items/tasks	8 tasks	8 items	2 tasks	8 items

Note. SR = selected response. S/ECR = short and extended constructed response.

The on-demand screener for kindergarten differs from that for Grades 1–12 in several ways. As it is intended for very young children new to schooling, it is organized using a "stair-step" model, beginning with tasks at Proficiency Level 1, with each step increasing one proficiency level. Each step offers the child three opportunities to provide evidence of meeting the MPI for that level. The student stops at the step at which he or she fails to provide adequate evidence (i.e., reaches a ceiling). The listening and speaking portions are combined into one section of the test, administered before the reading or writing sections. Depending on a student's age, schooling, and proficiency level, the kindergarten screener may take anywhere from 5 minutes (very low-proficiency students) to 30 minutes (high-proficiency students).

Classroom Benchmark Assessments

The development of the classroom benchmark assessments will give WIDA the opportunity to research and develop innovative computer-based item types, particularly for reading and listening. Our goal is to go beyond traditional multiple-choice items and develop more complex, machine-scoreable, constructed-response items that allow students to more directly provide evidence of their comprehension of aural or textual input by performing tasks aligned with their proficiency level and the MPIs, with appropriate support. Once these items have been researched, developed, and used in classrooms, our plan is to migrate their specifications up to the computer-based annual summative assessment.

We will develop an array of short, targeted benchmark assessments. Like the MPIs of the WIDA ELP standards, the assessments will be targeted by *grade level*, *language domain* (L = listening, R = reading,

W = writing, S = speaking), *language standard* (SIL = social and instructional language, LoLA = language of language arts, LoMA = language of mathematics, LoSC = language of science, LoSS = language of social studies), and *proficiency level* (1–5). While a vast number of benchmarks could be developed, we seek to develop a more limited number under this grant. These are shown in Figure 1.

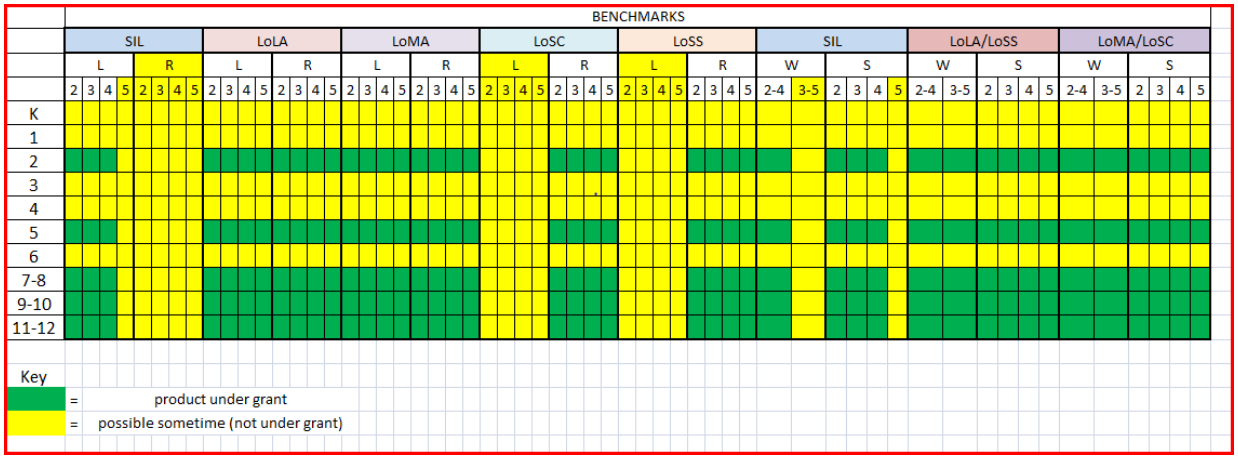

Figure 1. **Benchmark assessments to be developed.**

Our rationale for this approach is as follows. First, we do not propose any benchmarks for Level 1 since reaching that level is not a goal of instruction. Second, since listening—though an important skill—accounts for only 15% of the total score on the annual summative assessment, we will not develop listening benchmarks for LoSC, LoSS, or Proficiency Level 5 of SIL. Third, since SIL is the least academic of the standards, we will not develop reading benchmarks in SIL. Likewise, we will not develop writing benchmarks for Proficiency Levels 3–5 in SIL or speaking benchmarks at Proficiency Level 5 in SIL. Fourth, as in the current specifications for writing, writing tasks will cover three proficiency levels.. Fifth, we propose to fully develop benchmarks for Grades 2, 5, and the clusters 7–8, 9–10, and 11–12 because (a) all of the grade-level clusters (except K) on the annual summative assessment are represented and (b) college and career readiness is especially critical for ELs entering the U.S. educational system at the later grades (middle and high school).

Listening and Reading

The listening and reading inputs will follow research-based specifications that are fully aligned to the WIDA ELP standards and will be used as a foundation for the development of all the new assessments. However, the response spaces for the benchmarks will be designed using multisemiotic representations (including animations, dynamic and static images, and sound) to replace large amounts of text and focus the student on the item target. In this way, we will move away from traditional multiple choice and text-based constructed responses to innovative item types that more directly measure the targeted construct. Responses will be automatically scored, providing automated feedback to students and teachers immediately after each benchmark assessment is completed. We expect each targeted benchmark to include about 10 innovative items. Figure 2 shows an example of one type of innovative response space that can be used to allow students to demonstrate their ELP level in listening in the language of science.

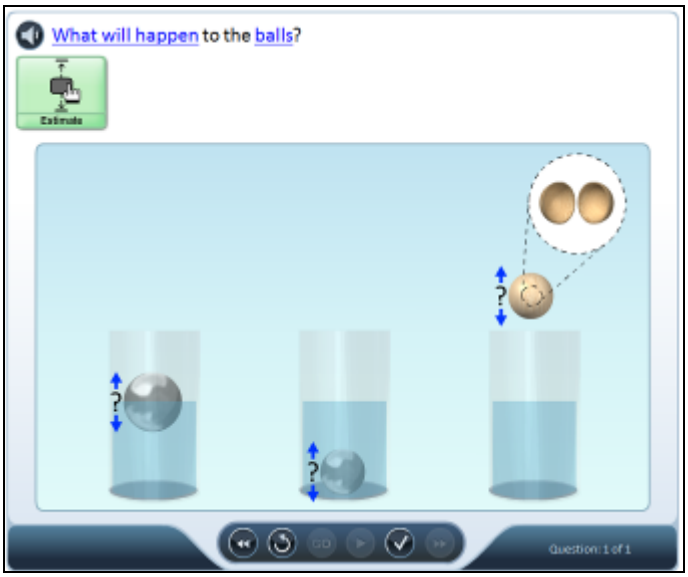

Figure 2. **Example response space: Student manipulates the balls based on aural input.**

Although the example in Figure 2 was originally designed to allow ELs to demonstrate science content knowledge using minimal language, the task (positioning the balls in a certain way by dragging them with a mouse) could also serve to demonstrate comprehension of linguistic input. For example, a listening input could ask students to display comprehension of a simple instruction to place the largest ball at the bottom of the container. More linguistically challenging listening input could ask students to

demonstrate their comprehension of more detailed elements, requiring a more fine-tuned placement of the objects—for example:

> *Emma was demonstrating to the class the outcomes of a scientific experiment. The results showed that the most massive steel ball floated on top of the liquid, whereas the wooden ball fell halfway down the container.*

This example illustrates how one response space could be used to allow students to demonstrate, on different benchmark assessments, comprehension of linguistic input targeted at different levels of ELP.

Table 8 lists selected MPIs from the WIDA ELP standards and innovative response spaces that have already been developed to measure the intended construct more directly than multiple-choice questions. Again, one response space could be used for a range of MPIs and a range of proficiency levels. For example, students could demonstrate the ability to match main ideas in a reading text with their details by completing a picture, not just by physically matching two objects.

Table 8

Sample MPIs and Response Spaces

Sample MPIs	Response spaces
L, Gr. 9–12, LoMA, P2: Create or change graphs, equations, or points on coordinate planes	Creating and adjusting graphs (e.g., shading a portion of a geometric shape, drawing a line graph, extending a bar graph)
R, Gr. 6–8, LoLA, P3: Sequence plots of adventures using visual support	Ordering, completing cycles, completing a picture
R, Gr. 3–5, LoSS, P3: Compare/contrast different time periods or people using graphic organizers and sentences	Creating Venn diagrams, classifying
L, Gr. 1–2, SIL, P5: Match oral descriptions of school areas, personnel, or activities with individual needs or situations	Matching (e.g., dragging words, images, or animations to match stimuli)

Note. L = listening. R = reading. Gr. = Grade. LoMA = language of mathematics. LoLA = language of language arts. LoSS = language of social studies. SIL = social and instructional language. P = proficiency level.

Speaking and Writing

The speaking and writing tasks on the benchmark assessments will follow the same specifications as those on the annual summative assessment. However, on the speaking task, each benchmark assessment will consist of three extended constructed-response tasks at one proficiency level, providing three opportunities for students to meet the relevant MPIs. Each writing benchmark—requiring a much longer response time—will consist of only one writing prompt.

2.9 Mode of Administration

Annual Summative Assessment

The new *annual summative assessment* will be computer-administered. For *listening*, the audio input to be comprehended will be delivered directly via the computer, while for *reading*, written text will be presented on screen. Response options will be presented and selected by students on the computer and be machine-scored. For *writing*, prompts will be presented on screen, with responses either handwritten or keyboarded, depending on which option the student is more familiar with. *Speaking* prompts will be presented on screen, with responses digitally recorded by the computer. Both writing and speaking responses will be centrally scored by trained and monitored raters.

The computer-based mode of delivery for the speaking assessment will increase consistency in administration—that is, all students will experience the same administration language and procedures and hear the same prompts. Moreover, this approach will lessen the local test burden, as compared with the one-on-one administration now used with ACCESS for ELLs®. Centralized scoring by trained and monitored scorers will enable the use of a more refined scoring rubric. The ACCESS rubric is necessarily simple (“meets/does not meet task-level expectations”) because test administrators must quickly score responses while they administer the test. Finally, the digital recording of responses will provide data for research on the development of academic oral ELP.

The computer-facilitated administration of the kindergarten annual summative assessment will simplify one-on-one administration of the assessment when compared to the current kindergarten ACCESS. The role of the administrator will not be eliminated, but it will become less complex. As the kindergarten ACCESS is closely tailored to the proficiency level of each child, the automated scoring and routing features of the new computer-based test will ease the burden of selecting the appropriate next section for the student. The numerous manipulatives and cards used in the paper assessment will be automatically organized and presented on screen. In addition, during the grant period, a second form of the kindergarten assessment will be developed.

(Note that during each state's period of transition from the paper-based ACCESS to the new computer-based annual summative assessment, a paper-based option will be provided. For kindergarten, the current paper-based version of the kindergarten ACCESS will remain an option. For grades 1 to 12, the *listening* input will be media delivered over a playback device and *reading* texts presented in a test booklet, with students selecting responses in the test booklet. *Writing* prompts and written responses will likewise be in the test booklet. For *speaking*, prompts will be delivered via test booklet and audio playback device, with student responses recorded on a second device. Again, both written and speaking responses will be centrally scored with those of the computer-administered version.)

On-Demand Screener

Like the annual summative assessment, the on-demand screener will be computer-administered, (though available during the transition period in both computer- and paper-based versions). The modes of administration and response for each language domain will parallel those for the annual summative assessment. The use of computer technology will eliminate the need for a highly trained test administrator and enable multiple concurrent administrations, while the paper-based version, due to its adaptive nature, will require one-on-one administration. A locally trained language education professional will score the written and spoken responses (see Section 2.10).

The kindergarten screener will also be available in two parallel modalities, with the computer again aiding the administration of the assessment rather than replacing the test administrator, who will need to be well trained and experienced in working with kindergarten-aged children.

Classroom Benchmark Assessments

The classroom benchmark assessments will be available only in a computer-based format because, for the listening and reading assessments, they will contain innovative response spaces that cannot be replicated in a paper-based format. The use of computer technology will enable us to use innovative ways of presenting writing task demands and oral prompts that are more easily comprehended (e.g., through graphics, animations, audio, and video) and encourage the use of more academic English language in response. In addition, the development of computer-based benchmark assessments for use in a non–high stakes environment, across all four language domains, will help local programs make the transition to all computer-based testing in the future.

The classroom writing benchmarks will provide opportunities for ELs to keyboard their responses. However, those who are unfamiliar with computers or just learning to write or keyboard will still have the option of writing responses on paper.

2.10 Scoring Methods

Annual Summative Assessment

The annual summative assessment will be centrally scored. The selected-response listening and reading domains will be automatically scored by machine. Input of student responses will be either direct (for computer-based tests) or from machine-scanned test booklets (for paper-based tests).

Similarly, student responses in the writing domain will either be digitized during administration (if keyboarded) or scanned (if handwritten). Oral student responses to the speaking prompts will be digitized. These performance-based responses will be centrally scored by trained raters. For the field testing during the grant period, MetriTech, Inc., will score all constructed responses because of their familiarity with the WIDA rubrics and their high-quality training of raters. The current writing rubric for ACCESS for ELLs® has worked well and will be minimally updated to reflect the amplification of the WIDA ELP

standards. A new speaking rubric will be developed, however, to allow greater differentiation between responses to tasks than is currently possible. CAL has already conducted research on a more refined, 4-point rubric and will undertake its further development and implementation under this grant.

As discussed in Section 2.4, WIDA anticipates a turnaround time of approximately 4 weeks from the time students take the annual summative test until they receive their scores.

On-Demand Screener

Responses to the selected-response listening and reading items of the Grades 1–12 screener will be dichotomously scored automatically by the computer or by hand if the paper version was used. Extended constructed responses for writing and speaking will be captured on computer, paper, or recorder and scored by locally trained scorers.

Two innovations will be incorporated in the administrator scoring of the screener writing and speaking items and all locally scored performance-based assessments (screener and benchmarks). First, we will adapt CAL's Multimedia Rater Training Program (MRTP) to provide intensive training and practice in scoring. The adapted MRTP will have the following characteristics: (a) interactive, technologically supported training, (b) content-based quizzes that provide diagnostic feedback as a rater learns material, (c) scoring quizzes that provide full justifications for scores, (d) computer-adaptive scoring practice that focuses rater-trainees' attention on aspects of scoring they find difficult, using an underlying pool of pre-rated samples, helping raters internalize the criteria, and (e) a digitized library of pre-rated responses that can be used as benchmarks against which to compare trainee performance. While CAL's MRTP has been developed for use in scoring task-based speech performance using speaking proficiency guidelines of the American Council on the Teaching of Foreign Languages (ACTFL), it will be easy to adapt the program for training on WIDA's generic scoring rubric that operationalizes the proficiency levels of the WIDA ELP standards for scoring speaking and writing.

Turning to the second innovation, local scoring will be supported by a computer-based scoring model that provides raters with the resources they need to produce accurate and reliable ratings for the speaking and writing tasks in the on-demand screener. This scoring interface will be modeled on the one used in

CAL's Computerized Oral Proficiency Instrument (COPI). Features to be included from the COPI scoring module include (a) access to information about the task the student completed (e.g., instructions, graphics, audio, video animations), (b) benchmark speech samples, and (c) explanations and justifications of the ratings for each benchmark speech sample. Using the computer-supported program to record scores for each domain helps eliminate errors, simplify score recording for the test administrator, and achieve cleaner score reporting. It also allows raters to keep diagnostic notes on performance for student records, which can be appended to students' score reports if desired.

The kindergarten screener will be scored by administrators while the test is administered, but technological support for their training will be analogous to that described above.

Classroom Benchmark Assessments

Listening and Reading

A key feature of the innovative response spaces to be used in the listening and reading benchmarks is that they will be automatically scored and provide instantaneous feedback to teachers and students. In automatic scoring of innovative response spaces, the computer program keeps track of the location of objects on the screen and their relationship to each other. Tolerances are set for partial and full credit, and programming governs where objects can be placed, how many can be placed in particular locations, and whether they snap to locations. All of this information is synthesized using research-based automatic scoring algorithms developed during the piloting of the new items and confirmed through field testing.

In the benchmark assessments, partial-credit scoring may be used to provide diagnostic feedback on listening and reading comprehension. Based on (a) finely detailed specifications for the linguistic components (vocabulary level, morphology, syntax, discourse structure) that apply to the proficiency levels defined by the WIDA standards and (b) the levels and types of nonlinguistic support provided, auditory and textual input can be tagged at the word, phrase, and discourse level. Then, for example, in response to a task requiring comprehension of that input, if a student correctly manipulates only some on-screen elements, it may be deduced that the student only partially comprehends the input. The linguistic aspects of the portion not comprehended may be delineated for the teacher. If these form a pattern across

items in a benchmark assessment, specific feedback may be given on aspects of language needing more practice and development. Since each benchmark assessment will be aligned with a specific proficiency level and standard, at the very least the student and teacher will receive feedback on whether the student has provided evidence of that level of proficiency on that standard or whether more practice to develop proficiency is needed.

Speaking and Writing

Teachers will use the generic rater-training programs described above (the adapted MRTP and COPI) for the speaking and writing portions of the benchmark assessments. Computer-based scoring modules will be developed for each speaking and writing task. These will provide all the supports needed for an educator to transfer training in the generic scoring to scoring the specific task. These supports will include task-specific notes, anchor samples at each score point with rationales, and alternative benchmarks for each score point. As mentioned previously, the scoring interface will allow a rater to make notes about a student's performance. Test scorers will be encouraged to note strengths and weaknesses of a student's spoken or written response and to identify interventions, activities, or practice might help the student continue to develop spoken or writing proficiency.

2.11 Reports Based on Assessments

Score reports resulting from the ASSETS project will build on WIDA's 7-year experience with delivering meaningful, uniform score reports customized to the needs of the various stakeholders of the consortium. The specific score reports that will be generated for ASSETS assessments will be determined by the SEAs of consortium member states; however, we anticipate that score reports will target audiences similar to those for ACCESS for ELLs®. Currently, WIDA provides the following reports for the ACCESS assessment: (a) a *parent/guardian report* that presents test results visually and numerically to help parents and guardians to understand their children's ELP levels in the individual language domains, oral language, literacy, as well as comprehension, and that also provides a composite proficiency level and scale score; (b) a *teacher report* that provides more detailed information to educators, including scale scores for all domains and combinations of domains and raw scores for each of the WIDA ELP standards;

(c) a *student roster report* that gives teachers and administrators an overview of the proficiency levels and scale scores for all domain and composite scores for ELs in a school; (d) a *school frequency report* that shows teachers and administrators the distribution of ELs according to their language proficiency levels for each domain and combination of domains in a school, and (e) a *district frequency report* that provides the same information for an entire district. In collaboration with SEAs and LEAs, WIDA is already providing translations of the parent/guardian report in more than 30 languages and will continue this practice with the ASSETS assessments.

3. ASSESSMENT DEVELOPMENT PLAN

3.1 Approach to Test Development

Development of Test Items

Because language is a complex system of knowledge, skills, and abilities, language testers at the Center for Applied Linguistics (CAL) have for years used the approach recently categorized by Mislevy and his colleagues (e.g., Mislevy, Steinberg, & Almond, 2003) under the rubric of *evidence-centered design*. In the development of ACCESS for ELLs®, the principles of evidence-centered design were adhered to as CAL's language testers sought to operationalize the WIDA ELP standards in a large-scale assessment. The *domain analysis* is provided by the WIDA ELP standards, as described in Section 2.2. *Domain modeling* is used to think about the evidence needed from students to demonstrate that they are meeting the MPIs of the WIDA ELP standards at the different proficiency levels. In designing test items, each item specification, each level of review, and each statistical analysis seeks to achieve a positive answer to the question:

> *If a child answers this question correctly or performs at this level on this constructed-response task, has he or she provided evidence of meeting the MPI at the given performance level, for the given standard, in the given domain, and at the given grade-level cluster?*

The overall design of the test (*the assembly model*) is intended to collect sufficient measurement information while keeping the assessment short and practical. During the annual refreshment cycle, the *student model* is continually supplemented by additional research conducted by CAL and others that

provides further insight into our understanding of the construct. The *evidence model* likewise is refined annually through research, experience, and psychometric analyses that ensure that the assessment's items and tasks fit the demands of the Rasch measurement model (Wright & Stone, 1979). In particular, the *task model* for each type of selected-response item or performance-based task is critically examined through cognitive labs, piloting, and field testing before it becomes operational.

Language testers at CAL find evidence-centered design particularly powerful because of its connection with Toulmin's (2003) structure of arguments. In each phase of development, testers ask whether, for example, tasks are eliciting performance related to the construct (and what warrants and backings exist for that construct) and what alternative hypotheses (other than language proficiency) might account for student performance. This approach ensures that a research perspective covers all aspects of test development.

In addition, CAL follows the principles of *universal design*. All the assessments in the ASSETS system will adhere to style guides, graphics guidelines, layout templates, and other resources and specifications CAL has developed over the years in applying universal design principles to language proficiency tests for ELs in Grades K–12. With every innovation in CAL's approach to testing, new resources and specifications are developed. These resources and specifications detail, for example, the use of easy-to-read text; clear, high-contrast graphics and visuals in color or black and white; and clean, simple layouts. Graphic organizers, maps, and other visuals used in the ASSETS assessments will follow specific guidelines to ensure consistent presentation across all assessments that is appropriate to the grade level for which the assessment is intended.

New applications of the principles of universal design are piloted in cognitive labs, and resources and specifications are updated annually. The success of CAL's internal guidelines and specifications is confirmed through annual content and sensitivity reviews conducted on each new item or task, as well as through feedback from the field and statistical information on the performance of the test items and tasks.

Development Phases and Personnel Involved

CAL follows a rigorous approach to developing language tests involving four phases: (a) initial development, (b) piloting, (c) field testing, and (d) operationalization. Below, we briefly describe each of the phases of CAL's approach and then indicate the types of personnel involved in each.

Phase 1—Initial Development

The goal of the initial phase is to create agreed-upon plans for the design and development of the assessment, including initial test and item specifications, written clearly enough for all stakeholders to understand. At the end of this phase, carefully selected prototypes of actual items are ready for piloting.

Phase 2—Piloting

Piloting is an iterative process in which all aspects of the prototypical items, administration instructions, and scoring procedures are carefully researched, including the computer-user interface. Piloting is exploratory in nature, and each pilot has its own research questions. Evidence to test hypotheses about the test items and procedures is collected through a variety of qualitative and quantitative methods. For example, cognitive labs may be conducted with students to ensure that they understand the computer-user interface, the task demands, and the best way to respond; and focus groups may be held with scorers to ensure that the rubric is clear. Validity evidence is sought to support hypotheses about the task and evidence model, and when alternative hypotheses cannot be adequately disconfirmed, prototypes, rubrics, and scoring procedures are revised based on the research findings.. With successive pilots, more items may be developed and join the piloting pool as item specifications become tighter and more refined.

Phase 3—Field Testing

The goal of field testing is to confirm hypotheses about all aspects of the test items, ensure that all aspects of the administration and scoring work as intended, and collect data to link performance on new items with performance on existing items. For most assessments, more items will be developed than will ultimately be needed. For the ASSETS project, we plan to conduct the field test together with the annual operational test, appending a small set of field-test items and tasks to the operational test.

Phase 4—Operationalization

The goal of the operationalization phase is to finalize all materials to ensure they are ready for large-scale use.

Mentioned above and discussed in greater detail in Sections 3.5 and 4.1, some of the development work and validity research, rely on linking scores on the operational ACCESS for ELLs test with student performance on the new tests. This will require the involvement of current WIDA ACCESS consortium members. However, non-WIDA-ACCESS states (non-ACCESS states) can be involved in all phases of the development of the new assessments. During phase 1, development, non-ACCESS states will give their feedback and input into the written framework and test and item specification documents as well as review prototypes. During phase 2, piloting, LEAs and students in non-ACCESS states can serve as sites for the cognitive labs and piloting. During phase 3, field testing, LEAs and students in non-ACCESS states can participate in the data collection to examine the psychometric properties of the new assessments. Finally, during phase 4, operationalization, students in non-ACCESS states can participate in the full administration of the new assessments.

Personnel

Table 9 sets forth major steps in each of the ASSETS development phases and identifies the personnel to be involved in each.

Table 9

Development Phases, Tasks, and Principal Personnel

Phases and tasks	Principal personnel
Initial development	
Create initial overall test design and development plan, initial test, and item specifications, with input and consensus from all stakeholders	Researchers and specialists in ESL education, applied linguists, language testing specialists, psychometricians, experts on career readiness standards, educators, computer testing experts (with consensual approval from SEA representatives and Technical Advisory Committee)
Develop and review item pool for piloting and assemble pilot-test form(s)	Language testing specialists
Develop ancillary materials for pilot test (including administrator/scorer materials)	Language testing specialists, professional development specialists
Pilot testing	
Conduct and score pilot test	Language testing specialists, qualitative researchers
Conduct pilot-test item analyses, reliability studies, and validation studies	Qualitative and quantitative researchers, psychometricians
Revise overall test design and development plan, including test and item specifications	Language testing specialists (with reviews by SEA representatives and Technical Advisory Committee)
Field testing	
Develop and review item pool for field test	Language testing specialists

Phases and tasks	Principal personnel
Assemble field-test forms	Language testing specialists, technology experts
Develop ancillary materials for field test (including administrator/ scorer materials)	Language testing specialists, professional development specialists, technology experts
Conduct field test	Language testing specialists, technology experts
Score field test	Language testing specialists, technology experts, educators
Conduct field-test item analyses, standard setting (as needed), and reliability studies	Psychometricians
Conduct field-test validation studies	Psychometricians, qualitative and quantitative researchers
Finalize design of all components of operational testing program	Language testing specialists, professional development specialists, technology specialists
Operationalization	
Assemble final operational test forms	Language testing specialists, technology specialists
Develop final score reports and score reporting system	Language testing specialists, psychometricians, educators, SEAs
Develop ancillary materials for operational test	Language testing specialists, technology specialists
Finalize plan for training and monitoring administrators/scorers and assemble final training materials	Language testing specialists, professional development specialists, technology specialists

Phases and tasks	Principal personnel
Conduct additional reliability studies	Psychometricians
Conduct additional validation studies	Psychometricians, qualitative and quantitative researchers
Finalize plan for continual monitoring and evaluation of operational testing program	Language testing specialists

3.2 Approach to Accommodations

WIDA has a proven history of working with SEA partners and national experts to develop and recommend appropriate accommodation policies for the ACCESS for ELLs® assessment. We will continue to do so in our work on ASSETS. The current edition of WIDA's *Guidelines for Accommodating English Language Learners with Disabilities* (WIDA, n.d.) provides guidance for the following aspects of testing ELs with disabilities: (a) test directions, (b) presentation format, (c) setting format, (d) timing/scheduling, and (e) response format. WIDA will work with SEA partners and national experts to ensure that appropriate accommodations can be provided on the new ASSETS technology-based assessments for ELs with disabilities.

In addition, WIDA is implementing a new performance assessment, Alternate ACCESS for ELLs™, designed specifically for ELs who have the most significant cognitive disabilities and thus are unable to take regular ELP tests, such as ACCESS, even with accommodations. WIDA plans to draw on the expertise gained from developing the Alternate ACCESS assessment in developing the ASSETS assessments.

Finally, WIDA will extend research it has conducted on options for providing an ELP testing solution for blind ELs that does not confound ELP and Braille proficiency constructs—a problem that almost inevitably arises when an existing language test, such as ACCESS, is simply translated into Braille.

Although not part of the ASSETS project, WIDA will continue researching and refining Alternate ACCESS and an instrument for blind ELs during the grant period and will endeavor to ensure their compatibility with the new ASSETS assessments, incorporate the assessment results into overall reporting systems, and, if possible and appropriate, create a computer-based form of the alternate tests.

3.3 Approach to Developing Scoring Materials

Annual Summative Assessment

As mentioned in Section 2.10, selected-response listening and reading items on the annual summative test will be scored by computer, whereas the speaking and writing constructed responses will be scored by human raters. For the grant period, MetriTech, Inc., will score the constructed response items using its online scoring system (MTscore). MetriTech has a proven record of hiring well-qualified raters using criteria such as completion of at least a bachelor's degree at an accredited college or university, work experience (particularly teaching or education-related experience), and pre-employment test scores. Many scorers have backgrounds in education and are active or retired teachers. Potential scorers participate in a rigorous online training program and must demonstrate mastery of the scoring rubrics, methods, and task types before scoring any operational tasks.

To ensure that the scoring rubric is being applied consistently across scoring sessions, specially prepared calibration sets are routed to each scorer daily. To the scorer, these look like regular student responses. However, master scorers have already reviewed each response in these sets and created a key of expected scores. Once the scorer completes the set, the system checks his or her scores against those in the master key. This approach allows for the immediate detection and correction of "scorer drift," the tendency for a scorer to begin deviating from the rubric over time. Agreement between the active and master scorer must exceed the standards established for the project (80% exact agreement) or the scorer is locked out of the system until he or she has successfully completed a retraining with the master scorer.

As a final check to ensure interrater reliability throughout the scoring process, 20% of all constructed responses are rescored by a master scorer. The master scorer then has the opportunity to provide feedback

to individual scorers to keep their scoring at the highest level of accuracy possible. Interrater information is kept for future analysis, inclusion in technical reports, and daily feedback to the individual scorers.

On-Demand Screener and Classroom Benchmark Assessments

Listening and Reading

For the on-demand screener, the scoring of the listening and reading selected-response items will be automatically scored by the computer or hand-scored by the test administrator using a scoring key.

For the classroom benchmark assessments, the innovative item types to be used in the listening and reading benchmarks will be only computer-administered and -scored and pose no special operational issues. The development of the scoring criterion, especially for partial-credit scoring, will follow a rigorous process during the test development stages, as described in Section 3.1. In particular, based on the linguistic features of the input and the response task, test developers will make hypotheses about the performance of students who demonstrate full comprehension, partial comprehension, and incomplete comprehension. These hypotheses will be initially confirmed, revised, or rejected through iterative rounds of cognitive labs (pilots) with the items. During the labs, examinee performance on the items will be carefully watched, and examinees queried about their comprehension of the input (e.g., by being asked to paraphrase what they understood in their own language or in English) and their reasons for choosing to complete the assessment task the way they did. Since each benchmark assessment will target a certain proficiency level, student participants in the cognitive labs for that benchmark will be both below and at the targeted level (based on current designations of proficiency according to the WIDA standards through ACCESS for ELLs® or W-APT scores, corroborated by teacher judgment).

The revised hypotheses about scoring, including those relating to response patterns that indicate partial understanding (partial credit), will be further confirmed through item field testing. Through empirical data modeling, partial-credit scoring will be examined to see whether or not it fits the Rasch measurement model and adds measurement information to total scores.. Items for which partial-credit modeling does not hold will be scored dichotomously in the operational benchmark assessment. For items for which partial-credit modeling does hold, a final check on the interpretation of the partial-credit

scoring—in light of the linguistic demands of the input to be comprehended—will be reviewed and approved by language education specialists and applied linguists before becoming part of the operational interpretation of the meaning of the performance.

Speaking and Writing

As described in Section 2.10, the scoring of the performance-based speaking and writing tasks will be done locally by trained teachers. Technology will support this scoring in two ways: through a generic multimedia rater-training program and through a task-specific rating module.

The development of the scored materials (e.g., anchor performances, benchmark performances, and scored performances for rating practice) will follow a vigorous procedure during the test development process. Obtaining performances to be used for these purposes will involve the following steps: (a) collection of a wide variety of performances during field testing (at least 500 per task); (b) internal calibration by CAL staff; (c) initial scoring of a subset of responses by at least five CAL staff members; (d) training of external scorers; (e) double or triple scoring of the entire set by external scorers; and (f) analysis of interrater agreement and selection of final performances. Development of the rater-training materials will involve these additional steps: (a) initial development of the materials for internal CAL use and review; (b) revision of the self-training materials and preparation for piloting by external scorers; (c) use of the materials as preparation for the external scorers, who will provide organized feedback on the materials and performances on calibration sets designed to test the materials' efficacy; (d) further revisions and review by CAL staff; and (e) a final pilot with untrained scorers, followed by their performances on rating calibration sets.

At the local level, some programs may want to increase the reliability of the ratings on the speaking and writing benchmarks. WIDA will provide guidance on how to incorporate a local scoring program. Such a program might involve requiring that items be scored by two independent raters, instituting procedures for arbitrating discrepant ratings, and ensuring that raters do not rate their own students. Such a program could be put in place for the on-demand screener, the speaking and writing benchmarks, or all rater-scored assessments.

3.4 Approach to Developing Reporting System

The governing states of the WIDA-ASSETS Consortium, especially the Steering Committee, will play a key role in designing a reporting system that meets the needs of multiple stakeholders and can be integrated with other state assessment systems. Colleagues from WestEd will consult on the types of data to be included and compatibility with other systems, while Data Recognition Corporation (DRC), the platform provider for the assessments during the grant period, will assist with technical aspects of operationalizing the reporting system.

3.5 Approach to Quality Control, Piloting, and Field Testing

WIDA follows a rigorous approach to quality control in all its test development and PDSR (i.e., printing, distribution, scoring, and reporting) activities, and it will continue this practice with the ASSETS project. Every year, a team representing consortium SEAs conducts a quality control site visit to CAL, in which documentation of all processes and procedures, responsibilities of the test development team, and qualifications of CAL staff members are reviewed and reported on to the full consortium board. The quality control procedures followed at CAL include clearly delineated multiple levels of internal review and signoff, external content reviews, external bias and sensitivity reviews, external post–field test reviews for final selection of items on operational forms, and internal and external editorial reviews and signoffs. All of CAL's psychometric work is also overseen by a technical advisory committee. A similar process occurs at our PDSR vendor, MetriTech, Inc., to verify the accuracy of score reporting and will be implemented at DRC for the work performed under this grant.

As described in Section 3.1, we follow a very rigorous approach to test development. Access to students for participation in piloting and field testing is coordinated through WIDA Central, working with SEAs. While the primary focus during piloting is an ever-expanding group of prototypical items, specific attention is paid to testing accommodations for ELs with disabilities and collecting data on their suitability. In addition, because all items are targeted to the proficiency levels across the entire continuum, very low- and very high-proficiency students are always selected for participation in pilots in order to collect data on the items that are intended to allow them to show what they can do.

Our approach to field-testing all items—for whatever domain or assessment (annual summative, on-demand screener, benchmarks)—will be to include them as part of the annual testing program during Year 3 of the project. Students selected to participate in the field testing will take one test (or a portion of one test) in one domain shortly after taking the annual test. Students in states that have made the transition to computer-based testing will take the computer-based items (including benchmark assessments), whereas those who take the paper-based version will be given the paper-based versions of annual assessment and on-demand screener items. Working with consortium states' SEAs—and through them, LEAs—we will ensure inclusion of a wide variety of ELs—from those recently arrived, to those who have been reclassified, to those with disabilities.

4. RESEARCH AND EVALUATION

4.1 Plan for Psychometric Analyses

Annual Summative Test

Because the ASSETS annual summative test will be built on the foundation of the current ACCESS for ELLs®, we will illustrate our planned approach to psychometric analysis of the new test with examples from our well-established existing procedures for analyzing items and test forms. In addition, to support the validity of interpretation of the quantitative analyses, we include descriptions of qualitative analyses that will be conducted.

Measurement Models

As with ACCESS for ELLs®, the Rasch measurement model (Wright & Stone, 1979) will form the basis of the psychometric analysis undertaken in developing the ASSETS summative test. The Rasch model (e.g., Wilson, 2005) guided all measurement decisions throughout the development of the ACCESS assessment (Kenyon, 2006). Careful analysis based on Rasch fit statistics guided decisions about the inclusion, revision, and deletion of items during the development and field-testing of the ACCESS test forms. For listening, reading, and speaking, a dichotomous Rasch model was used; for writing, a Rasch rating scale model was used. A similar approach will guide psychometric analyses during the development of the ASSETS summative test.

Equating and Scaling

The equating and scaling procedure used with ACCESS for ELLs® was designed through adjacent grade-level cluster testing to derive a single, vertically equated scale from kindergarten to Grade 12 so that progress could be measured across all grade levels (Kenyon, 2006; Kenyon, MacGregor, Li, & Cook, in press). In addition, horizontal equating was conducted across the three tiers of ACCESS within each grade-level cluster so that progress could be measured across tiers. In brief, this scaling was accomplished during the field test based on an elaborate common item design—across both tiers and grade-level clusters—that spanned two series of complete test forms. Concurrent calibration was used to determine item difficulty measures. These item difficulty measures were used to create the ACCESS scale scores used for reporting results on the test. Such careful procedures will be used to ensure the stability of the next-generation ASSETS scale.

Item Refreshment and Annual Equating

WIDA has a well-established plan for annual refreshment of ACCESS for ELLs® to prevent item exposure and ensure continuous improvement of the assessment. Annually, between one third and one half of the items on the operational test forms are refreshed. Annual equating is conducted to place results on new series onto the ACCESS score scale through a common-item equating procedure. Items that are not revised are anchored to the difficulty values from previous series. Through a similar careful process, the score scale will be maintained as we transition to the next generation of assessments.

Reliability and Errors of Measurement

For ACCESS for ELLs®, a variety of approaches—including Cronbach's alpha and stratified alpha—are used to provide estimates of the test reliability by domain and composite score. In addition, item response theory (IRT) information function and IRT-based conditional standard errors of measurement are provided by domain. Complete information on the reliability of each year's ACCESS, including interrater reliability for the writing test, is provided in the ACCESS annual technical report. Such techniques will be used to analyze the reliability of each assessment in the transition to the next generation of assessments.

Tier system

ACCESS for ELLs® was designed to measure a wide range of proficiency levels as described in the WIDA ELP standards. To make the test appropriate for students across proficiency levels, the test items are presented in three overlapping tiers for each grade-level cluster (Section 2.8). The development of thematic folders presenting items at three adjacent proficiency levels—arranged across the five main proficiency levels of the WIDA ELP standards in listening, reading, and writing—makes it possible to administer the test level that is most appropriate for the student's proficiency, which in turn increases the reliability of the measurement and decreases the amount of measurement error. This approach will be used during the period of transition, while parallel paper- and computer-based versions of the new annual assessment are used across states in the WIDA-ASSETS Consortium.

Validity

In examining the validity of ACCESS for ELLs®, we use an argument-based approach (e.g., Kane, 2006; as applied to language testing, see Bachman, 2005; Mislevy, Steinberg, & Almond, 2002). This approach combines a focus on the assessment (*assessment argument*) with a focus on its use (*use argument*). Central to this approach is a clear statement of proposed interpretations of test results. Similar to ACCESS, the overarching purpose of the assessments developed under the ASSETS project will be to assess the developing ELP of English learners in Grades K–12 in the United States following WIDA's ELP standards. Additional purposes include (a) identifying the ELP level of students with respect to the WIDA ELP standards; (b) identifying students who have attained ELP; (c) assessing annual ELP gains using a standards-based assessment instrument; (d) providing districts with information that will help them evaluate the effectiveness of their ESL/bilingual programs and determine staffing requirements; (e) providing data for meeting federal and state accountability requirements with respect to student assessment; and (f) providing information that enhances instruction and learning in programs for ELs.

Below, we describe studies addressing the validity of ACCESS, stating the claim to be investigated, evidence collected or to be collected to support the claim, and the methodology used or to be used to test the validity of the claim. Some validation studies of ACCESS have been completed, and more are

planned. We present these here as concrete examples of the types of validity studies we will undertake for the ASSETS project, as the current assessment transitions to the next generation.

Construct validity. Construct validity—what test scores mean and what kinds of inferences they support—is the central concept underlying the ACCESS for ELLs® test validation process. Evidence for construct validity integrates evidence from both content- and concurrent-related validity. For interpretation of students' performance on ACCESS to be meaningful, test scores must correlate highly with independent measures of language proficiency. Patterns of correlation with other ELP assessments should demonstrate convergent and discriminant validity—that is, tests designed to measure similar skills should correlate more highly than tests designed to measure distinctly different skills. Here, we discuss our approach to the ACCESS test validation process to illustrate the approach we will take in establishing the validity of the new ASSETS computer-based tests and to emphasize that the foundation for the new ASSETS annual assessment—ACCESS for ELLs®—has already accumulated much validity evidence.

The first steps in establishing the construct validity of the ACCESS assessment are careful specification of content and review of the ACCESS items assessing the MPIs. Empirical evidence—especially item-level data such as item fit and point-biserial correlation—is used to identify the presence of construct-irrelevant elements. Another indication of construct irrelevance might be differential item functioning (DIF). To minimize construct-irrelevant variance that can occur when raters score constructed-response items more leniently or more severely than established standards, a number of procedures and interrater reliability checks are instituted, which will be used throughout the ASSETS project. Finally, since research on the interaction between student characteristics (e.g., native language, sociocultural background) and item features suggests (a) that items may be *psychometrically* but not necessarily *psychologically* equivalent (Ferrara & Chen, 2011; Pearson & Garavaglia, 2003; Sato, 2011) and (b) that different item features may present construct-irrelevant challenges to students that could affect their processing of information and thus their demonstration of knowledge related to the targeted construct, cognitive interviews will be conducted with a purposeful sample of EL and non-EL students. These interviews will use both concurrent and retrospective structured protocols and require students to

provide verbal reports on their processing of a selected set of items in order to examine whether any construct-irrelevant factors are affecting their access to and engagement with the items and whether the items are measuring the targeted constructs as intended (Almond et al., 2009; Ericsson & Simon, 1980; Ercikan, 2006; Leighton, 2004; Paulsen & Levine, 1999; Sato, Rabinowitz, Gallagher, & Huang, 2010).

Minimizing item bias. Minimizing item bias is essential in ensuring that ACCESS measures students' ELP without introducing construct-irrelevant elements in the performances on which the measurement is based. Tests that require students to have specific cultural knowledge and skills not taught in school can result in bias because of differences in student background and out-of-school learning (Camilli & Shepard, 1994). Three measures that are taken to minimize bias in the ACCESS assessments will be adopted in the ASSETS project. First, careful attention will be paid to content validity during the item-writing, cognitive-lab, field-testing, and item-review processes. Second, every item will be approved by a bias and sensitivity panel before it is administered to any student. Third, operational data will be examined to identify items with high levels of DIF. Such items will then be examined to determine if item performance differences between identifiable subgroups of the population are due to extraneous or construct-irrelevant information, making the items unfairly difficult for a certain subgroup. The inclusion of such items will be minimized in the test development process.

In the past, DIF of ACCESS items has been assessed for males vs. females and Hispanics vs. non-Hispanics at each cluster in which the items are administered for each ACCESS series, with findings reported in the ACCESS annual technical reports. Items with high levels of DIF are removed from the operational test at the earliest possible moment. The collection of additional ancillary background data on ELs in the ASSETS project will make it possible to create additional groupings of students.

Concurrent validity. The results from a study of the relationship between ACCESS and four other ELP tests (IDEA Proficiency Test; Language Assessment System; Language Proficiency Test Series; Maculaitis Assessment of Competencies Test of English Language Proficiency II; Gottlieb & Kenyon, 2006; Kenyon, 2006) showed a moderate to strong correlation between performance on ACCESS and performance on these tests. This finding provides strong support for the claim that performance on

ACCESS represents an assessment of ELP. However, the absence of very high correlations provides some support for the claim that the standards-based ACCESS is assessing the construct of ELP somewhat differently from the other, older tests.

Consequential validity. Some testing experts use consequential validity to refer to the social consequences of using a particular test for a particular purpose. The use of a test is said to have consequential validity to the extent that society benefits from use of the test. Other testing experts believe that the social consequences of using a test—however important they may be—are not properly part of the concept of validity. Despite these disagreements, it is believed that consequential validity should be an integral part of test validation.

The most important consequence of ELP tests is the use of the test data to make judgments about the proficiency of ELs in K–12 programs. Since such decisions are usually made on the state and local levels, the consequential validity of ACCESS is best addressed through a series of carefully planned research and evaluation studies with input and involvement from state and local stakeholders. In moving from ACCESS to the next generation of ELP testing, the most relevant consequential validity issues are (a) whether the assessments are being implemented as designed and (b) whether the theory of action (Section 1) is being realized, including whether the intended effects on individuals and institutions are being achieved. To this end, the WIDA-ASSETS Consortium has been working on evaluating the impact of adopting the WIDA ELP standards and the ASSETS assessments on state and local levels. Details of these research and evaluation efforts are addressed in Section 4.2. Necessary in supporting these efforts and informing the interpretation of their findings is the systematic analysis of documents (e.g., administration manuals, training and professional development materials, scoring protocols, score reports, interpretation guides, proficiency-level descriptors) to ensure that the purpose, uses, ELP domain and language modality definitions, and population definitions are consistently and accurately represented. Research suggests that inconsistent or insufficient documentation and communication of these critical factors can affect assessment implementation and thus the validity of interpretation of assessment results (AERA, APA, & NCME, 1999; Assessment and Accountability Comprehensive Center, 2009; Crooks,

Kane, & Cohen, 1996; Gorin, 2007; Kane, 2007; Lane, Parke, & Stone, 1998; Linn, 1997). This document analysis will help ensure that (a) the potential for misuse or misunderstanding of the assessment resulting in negative unintended consequences is mitigated and (b) the potential for fidelity of implementation and implementation conditions that facilitate the intended consequences of the assessment is improved.

Comparability Between Paper- and Computer-Based Annual Summative Assessment

Because we anticipate having both a paper-based and a computer-based version of the summative assessment, it is critical that the tests measure proficiency in the same construct and with the same degree of precision and that their scores be interchangeable. Five validity hypotheses should be tested: (a) test content and content specifications are the same, (b) scores have the same factor structure, (c) scores have the same measurement precision, (d) score distributions differ only in difficulty and hence are equitable, and (e) scores are highly related to one another.

These five validity hypotheses will be examined using experimental data gathered via a within-subjects design for each domain assessed by the summative test. During the field-test phase, students will be administered paper- and computer-based versions of at least one domain of the same summative test. The paper-based version will be administered during the operational testing window, and the computer-based version, within 2 weeks. A minimum of 500 students from each grade-level cluster will be enrolled as participants. The item parameters from the paper-based test will be used for common-item equating, making it possible to equate studies to determine mode effects. The results of the study will also inform future computer-based test development to minimize potential mode effect during the time of transition to computer-based only.

On-Demand Screener

All developmental activities and analyses associated with the new summative assessment will be applied to the new computer-based on-demand screener as well. For example, the items on the screener test will be reviewed by a bias and sensitivity panel, and no items will be administered to students unless approved by the panel.

To put the results of the screener test on the same scale as the summative test, the screener will be field-tested by domain using a common-person equating design. In addition, items that have been retired from previous summative assessment administrations may be included in the screener, also enabling common item linking. During the field-testing phase, for each domain in each grade-level cluster and each version (paper- or computer-based), students who take the summative test will subsequently be administered one domain of the appropriate screener version. Results from the field test will be analyzed using the software program WINSTEPS. A concurrent calibration of the two tests across students will allow us to (a) estimate the item difficulty of the screener on the same scale as the summative assessment and (b) investigate any possible differences between a domain on the summative assessment and the same domain on the screener.

Placing results of the new screener on the same scale as the summative assessment will allow us to interpret results in terms of the proficiency levels of the WIDA ELP standards. Because the screener will be shorter than the summative assessment, additional steps will be taken to ensure accurate placement decisions. First, the screener will include extra items designed to test ELP at Levels 4 and 5. Second, we will recommend that the ASSETS policy committee adjust upwards the cut score for placement out of language support services by one or two standard deviations to reduce the incidence of false positives (i.e., students who need English language support services despite performance on the screener indicating that they do not).

One important goal for the screener is that it accurately predict how well a student will perform on the summative test. In service of this goal, we will collect and analyze data during the operational year on the performance of students on the screener and their later performance on the annual summative assessment.

Classroom Benchmark Assessments

The main purpose of the classroom benchmark assessments is to provide information to stakeholders on whether a student is achieving targeted growth in developing ELP in a given language domain for a given standard. A secondary purpose is to bring concrete examples of the proficiency levels of the WIDA ELP standards into classroom instruction. Therefore, the benchmark tests should provide maximum

information around the relevant cut scores. To ensure that these assessments serve this purpose, each will be targeted very specifically.

During the field-test phase, these short tests will be administered to students who have recently completed the annual summative assessment. Results from this common-person design will be analyzed to (a) confirm that the items or tasks are at the intended levels, (b) collect initial validity evidence of their use as benchmark measures of targeted performance vis-à-vis performance on the annual assessment, and (c) provide an interpretation of the score in terms of the WIDA ELP scale score. Items or performance tasks whose empirical results confirm that they are operating at the intended proficiency level will be chosen for the benchmark tests.

4.2 Plan for Examining Washback

To determine whether the ASSETS assessments are being implemented as intended and whether their intended effects are being achieved, a consequential validity study will be conducted. This study will focus on all ASSETS assessment components: the annual summative assessment, classroom benchmarks, on-demand screener, and formative assessment resources. Two methods will be used to collect consequential validity evidence. First, an online survey will be administered. The survey instrument will look at a number of areas associated with implementation of the ASSETS assessments, including but not limited to:

1. Teachers' perceptions of what is being assessed;
2. Teachers' and administrators' preparation for assessment administration;
3. Teachers' and administrators' interpretation and use of results;
4. Types of professional development activities engaged in to support the assessment;
5. Types of curricular material (if any) adopted as a result of participation in the assessment;
6. Parents' perceptions of the assessment and its purposes; and
7. Students' perceptions of the assessment and its purposes.

Second, several focus group sessions will be conducted with SEAs and LEAs on the topics listed above. Data from the surveys and focus group sessions will be analyzed, interpreted, and published as part of the

annual technical report. Results from this study will be reviewed by the assessment development team and consortium members. If unforeseen consequences are found, assessment design, materials, administration, or implementation will be modified accordingly.

5. PROFESSIONAL CAPACITY AND OUTREACH

5.1 Plan for Supporting Teachers and Administrators

WIDA will develop a comprehensive professional development system to help educators implement the ASSETS project's annual summative assessment, on-demand screener, and classroom benchmarks. The assessment system will include administration manuals, training materials, sample items for practice scoring, and logistical information. Particular efforts will go toward preparing materials for teachers on the benchmark assessments—when and why to use them and how to interpret their results to guide instruction. These materials will be available in a portable electronic format and online to facilitate access, whether through self-training or district or school trainings. Educators will also have opportunities to attend face-to-face trainings with certified consultants. These trainings will be coordinated with state personnel to meet the specific needs of the different states throughout the consortium. To ensure the effectiveness of training sessions and materials, WIDA will collect evaluations from participants in face-to-face and online trainings. The variety of training formats and venues will not only help build capacity, but also accommodate different learning styles.

Educators will have access to an interpretative guide that will provide general information about the assessments and guidance in the interpretation of their scores. Additional help in interpreting scores and using results in instructional planning and decision making will be available in the form of training materials, including case scenarios, and blended trainings. Districts and states will receive guidance on the analysis of longitudinal data. Like the administration manuals and other pre-assessment materials, these post-assessment materials will be available in a variety of formats to accommodate different learning styles and maximize dissemination.

Using these tools and the ELP standards, WIDA will prepare materials to help educators integrate academic language formative assessment into their educational practice. Educators will be trained in how

to use the standards to set language targets and mine data from the ASSETS assessments to inform and improve their educational practice. All of these materials will also enhance educators' communication with families and the community about the education of ELs, as discussed next.

5.2 Strategy and Plan for Informing the Public and Key Stakeholders

The ASSETS project will implement a multifaceted strategy for informing consortium members, other key stakeholders, and the wider public. A dedicated website will offer information, news, and materials for the general public. A password-protected area of the website will provide training materials, videos, and confidential information about the assessments for consortium members. An annual meeting will give consortium members the opportunity to share information, provide input, vote on key policy decisions, and network with colleagues from other member states. Further, the consortium Steering Committee will meet two to four times each year to provide direction for the project and will share the results of these meetings with the member states. Conference calls, a bulletin, and webinars will also help keep members up-to-date. Through all of these avenues, SEAs will have timely information to share with stakeholders within their states. In addition, SEA representatives and project staff will attend regional and national meetings to share information about the project and to learn from others to inform ASSETS development. Finally, the WIDA Help Desk will be available during business hours by phone (toll-free) and email to answer questions or make referrals to other staff when greater expertise is required.

6. TECHNOLOGY APPROACH

6.1 Uses of Technology and Rationale

WIDA is committed to using technology to the maximum extent appropriate to develop, administer, and score the assessments and report assessment results. In close collaboration with WestEd, WIDA will develop all assessment items to an open-licensed interoperability standard that is industry-recognized and approved by the U.S. Department of Education. The interoperable design will support (a) test-test content portability; (b) transfer of assessments from one technology platform to another; (c) consistent assessment delivery across consortium states; (d) consistent application of accessibility features, including universal design of items; and (e) coordination and compatibility of the system with relevant practices of the Race

to the Top (RttT) assessment consortia. To maximize the interoperability of the assessments, a corpus of codes or tags will be used to specify (a) key elements of each test item; (b) alternative ways of presenting test content to maximize accessibility; (c) characteristics of the range of student test-takers; and (d) student and system behaviors expected to result when particular codes or tags are applied to an item.

The Accessible Portable Item Profile (APIP) standards are one example of the type of codes or tags that will be used to maximize interoperability. APIP incorporates key elements of established Question and Test Interoperability specifications, Access for All specifications, and the National Instructional Materials Accessibility Standard to create a single standard for accessible item file format, accompanied by documentation of intended behaviors when the standardized APIP tagging structure is applied to test items. The RttT assessment consortia have been discussing standards such as APIP, and the WIDA-ASSETS Consortium will adopt and use standards consistent with those selected by the RttT consortia in order to maximize the interoperability of the ASSETS assessments and their items. Documentation will allow full interoperability across different delivery platforms in the WIDA-ASSETS Consortium states.

WIDA recognizes the urgent importance of moving large-scale student assessment from paper-based to computer-based test delivery, as well as the need to leverage technology to deliver more robust reporting tools and instructional resources to educators as part of a comprehensive and interactive computer-based assessment system. Following a rigorous selection process, WIDA has chosen the INSIGHT Online Learning System—a product of the Data Recognition Corporation (DRC)—as the delivery platform for all field-testing activities under this grant. DRC INSIGHT is a secure system that pairs maximum control and flexibility with features critical to the ASSETS project. These features include (a) the ability to deliver summative, formative, screener, and benchmark assessments in computer-adaptive or fixed-form formats; (b) support for interactive, innovative test items that fully leverage the available technologies of a computer-based test environment, such as animation (dynamic graphics and simulations), graphing, drag-and-drop, short answer, completion, and hot spot; and (c) the ability to capture and store spoken responses for the speaking component of the ASSETS tests. DRC

INSIGHT will deliver approximately 56,000 items for the ASSETS project assessments to at least 7,000 students during field testing.

6.2 Strategies for Addressing Technology-Related Barriers

WIDA is sensitive to the varying technological capabilities of districts and schools and is committed to working with districts to minimize the need for extensive technology upgrades. DRC INSIGHT was designed to work with the technology commonly available in schools; the desktop-based online testing interface is fully compatible with Microsoft Windows (2000+) and Apple Mac OS X (10.4+) operating systems and has the capability to automatically update itself with the newest published version of code. When needed, DRC will conduct a technology survey to assess current district capabilities and areas of need related to school online testing hardware and make recommendations or provide assistance to resolve technological barriers.

Slow or intermittent Internet connectivity can have an impact on the e-testing experience. To mitigate this impact, DRC has developed a local caching service that allows test data to be housed locally on school-owned hardware, minimizing the need for a speedy Internet connection. DRC also has systems in place to ensure that capacity requirements do not impede the implementation of online testing. The DRC INSIGHT infrastructure was developed to be fully scalable, providing the flexibility needed to accommodate each state's capacity needs. DRC carefully monitors current system usage and capacity requirements to plan for future needs and adds more application or web servers as required to ensure smooth, fully supported student data handling and system downloads.

7. PROJECT MANAGEMENT

7.1 Workplan and Timeline

WIDA has established a workplan and timeline for the ASSETS project (Table 10) that takes into account the multiplicity of activities required for successful completion. These include the need for National Development Advisory Group and Steering Committee meetings, assessment development activities, interoperability planning, validity research, professional development, and policymaking.

Table 10

Project Timeline (see note at end of chart for key)

Activities	Lead	Year 1				Year 2				Year 3				Year 4			
Clarify roles, relationships, goals, & tasks	PI,W																
Convene AG, ST, & TAC meetings	PI,W																
Establish subcommittees[a]	PI,W																
Create initial test design & development plan	C																
Establish corpus of codes	WE																
Create item development style guide	WE																
Conduct item writer training	WE																
Create initial test & item specifications	C																
Develop & review item pool for piloting	C																
Assemble pilot-test form(s)	C																
Develop ancillary materials for pilot test	C																
Conduct validation research	WE,C,U																
Create PD materials	W																
Conduct pilot test	C																
Score pilot test	C																
Develop consortium materials re: accommodations	WE																
Conduct pilot test on PD materials	W																
Conduct pilot-test item analyses	C																
Conduct pilot-test reliability & validation studies	C																
Revise overall test design & development plan	C																

Activities	Lead	Year 1				Year 2				Year 3				Year 4			
Develop & review item pool for field test	C																
Assemble field-test forms	C																
Develop ancillary materials for field test	C																
Conduct field test	C																
Score field test	C,MT																
Conduct field-test item analyses	C																
Conduct field-test reliability & validation studies	C																
Finalize design of operational testing program	C																
Establish plan for scale-up & operationalization	W																
Conduct final review of PD materials	W																
Assemble final operational test forms	C																
Develop final score reports & reporting system	C																
Develop ancillary materials for operational test	C																
Finalize plans for administrator training/	C																
Assemble final materials for administrator training	C																
Review interoperability	WE																
Conduct additional reliability & validation studies	C																
Finalize test monitoring/plan	C																
Convene project closeout meetings	W																
Submit final reports to USED	PI																

Note. Lead = responsible entity. AG = advisory group. ST = Steering Committee. TAC = Technical Advisory Committee. PI = Wisconsin Department of Public Instruction. W = WIDA. C = Center for Applied Linguistics. WE = WestEd. U = UCLA. MT = MetriTech. PD = professional development.
[a]Subcommittees: accommodations, score reporting, EL definition.

7.2 Identifying, Managing, and Mitigating Risks

In this proposal, we have outlined a solid development strategy carefully thought out to mitigate risk. We have a clear plan for the gradual change from paper-based to computer-based testing. Our plan for the annual summative assessment is conservative, incorporating a period when paper- and computer-based tests will be used simultaneously and the states—with the support of DRC during the field testing and whoever the platform provider might be after the grant period—will gradually be helped to make the transition to computerized assessment. Likewise, the on-demand screener will be available in both paper- and computer-based formats.

The riskiest parts of the proposal lie in the innovative item design for listening and reading in the benchmark assessments. Our vision is to migrate the item specifications from the benchmarks to the annual summative assessment once the research on these innovative item types is complete and the items themselves have been shown to function with stability. In addition, because there cannot be strictly parallel paper- and computer-based versions of these tasks, we plan to wait to migrate item specifications up to the annual summative assessment until the majority of students are using the computer-based version. Since the annual summative test is on an annual refreshment plan, it will be possible to transition to the new item types gradually.

We have strong experience with good project management practices. ASSETS will adopt these proven practices—including weekly communication between partners, quarterly face-to-face meetings, and meetings with the Executive Committee, full Board of Directors, and Technical Advisory Committee—to closely monitor progress and ensure that resources are being used efficiently.

7.3 Adequate Budget

Based on WIDA's experience managing a large consortium and developing comprehensive and complex assessments, we believe that our budget is reasonable and adequate to support the development of assessments that meet the requirements of the priority as outlined in this proposal.

7.4 Estimated Costs and Plans for State Implementation

Current WIDA Consortium states pay $23 per student tested with ACCESS for ELLs®. Based on development costs and projected costs for computer-based distribution, scoring, and reporting, we estimate that this next generation of assessments—particularly the annual summative and screener tests—will be comparable or, more likely, lower in price. WIDA states have indicated that they anticipate no change in the level and source of funding for required assessments for accountability purposes.

7.5 Quality and Commitment of Personnel

The work of each of the partnering entities in the ASSETS project—WDPI, WCER, CAL, WestEd, UCLA, DRC, and MetriTech—will be directed and managed by well qualified professionals who have been working and leading in their respective disciplines for many years and who have developed the capacity of their staff to ensure that the work on this project will be accomplished with a high degree of quality and timeliness as conceptualized by the SEA consortium members and partnering entities. Furthermore, all the members of the National Development Advisory Group and the Technical Advisory Committee are nationally recognized leaders in their fields and whose experience and scholarship will provide wise counsel from multiple perspectives. The key personnel for the project, the directors and managers of various components of the project in most cases, are listed below by organization.

Wisconsin Department of Public Instruction: Project Director **Lynette K. Russell**, PhD, is the Director of Educational Accountability WDPI. She has a background in educational administration and supervises the statewide assessment system for all Wisconsin public schools. She will guide and oversee all reporting requirements, implementation and coordination of grant activities, and all work of the contractor to ensure that all goals, objectives, and deliverables are met. WDPI Assistant Project Director, **Philip Olsen**, MA, is the Assistant Director of Educational Accountability in WDPI. He oversees the

statewide assessment system for the alternate assessments for students with disabilities and accommodations for English language learners and will guide the work of two additional WDPI staff members: **(1) Standards Based Assessment Consultant** (To be assigned), who will have extensive experience in consulting, coordinating, and communicating with interdepartmental resources, administrators, district staff, educators, and policy makers at the local, state, and national level; and (2) WDPI **Consortium Project Coordinator** (To be hired), who will have extensive experience in communication, test administration and project coordination. In this full-time position, he or she will oversee the day-to-day coordination of the steering committee activities to ensure the products and services meet the needs of the states and federal requirements. In addition, this person will work closely with the WIDA Project Manager on the assessment development related organizational and logistical aspects of the project.

Wisconsin Center for Education Research (WIDA "Central"): As the managing partner of the ASSETS project, WIDA staff at WCER will coordinate, and in some cases direct, the efforts of all entities, including ASSETS consortium activities and subcontracts. In addition WCER will spearhead professional development, and engage in research related to the project. These efforts will be directed by Principal Investigator **Timothy Boals**, PhD, executive director of the WIDA Consortium. He is responsible for leadership, strategic planning, operations, and board relations. He has a background in curriculum, language education, Spanish language and literature, bilingual/ESL, and educational policy. Co-Principal Investigator **Elizabeth Cranley**, PhD, is associate director of the WIDA Consortium and is responsible for products and services. She has a background in comparative education and ESL. She will direct test development from WCER. WIDA Lead Developer **Margo Gottlieb**, PhD, is the primary developer of the WIDA ELP standards and has been a key force in the development of all major WIDA initiatives. Her background is in English proficiency testing, bilingual and ESL education, EL policy, and standards development. For ASSETS, Dr. Gottlieb will serve as a senior advisor and guide operationalization of the ELP standards at no cost to the grant.

Additional key WCER staff includes **Carsten Wilmes**, PhD, who, with a background in language testing, manages all WIDA operational assessments and assessments-in-development and is the primary liaison with the CAL office; **Mariana Castro**, MA, who directs and manages development and implementation of all WIDA professional development and teacher resource development, including those associated with ASSETS; Research scientist **H. Gary Cook**, PhD, who directs research for WIDA, including non-psychometric research for ASSETS; **WIDA project manager** (to be hired) who will have extensive experience in management of large, complex projects. The manager will direct the goals and deliverables of the grant at WCER, including compliance with requirements for research with human subjects, contract administration, report preparation, and dissemination. This person will also serve as the primary liaison with the consortium member states and oversee the organizational and logistical aspects of the project; and the **WIDA assessment project coordinator** (to be hired) who will have experience in professional development, test administration, and project coordination. This individual will oversee the day-to-day operations of the ASSETS test development process and help coordinate research and development of curriculum and technology-based training materials.

Center for Applied Linguistics: **Dorry M. Kenyon**, PhD, is a CAL vice president and director of the CAL Language Testing Division. He directs or serves as senior advisor on a variety of projects related to EL assessment. He also serves as CAL's chief psychometrician and the leader of its Psychometrics/Research Team. Active in research on language testing, Dr. Kenyon is particularly interested in the application of new technology to language assessment problems and will serve as director of the ASSETS project at CAL and as a senior project advisor for the project at no cost to the grant. **Margaret E. Malone**, PhD, is senior testing associate at CAL and co-director of the National Capital Language Resource Center. Dr. Malone will serve a project advisor for the development of the ASSETS assessments, particularly the speaking tests.

Other key CAL staff who will contribute to the ASSETS project include **Dr. David MacGregor**, who manages the Psychometrics/research Team for ACCESS for ELLs® as well as the development of WIDA's forthcoming Spanish academic language test and; **Jennifer Christenson**, who oversees test

development for the ACCESS for ELLs® and Alternate ACCESS assessments, including a staff of approximately 21 people (10 FTE); **Dr. Shu Jing Yen**, who serves as a psychometrician for ACCESS for ELLs®; **Catherine Cameron**, who manages operations activities for research on the technology-based ONPAR assessment in science and math for ELs and other students with reading challenges; **David Gabel**, who manages item development for ONPAR; and **Anna Z. Todorova**, who serves as project manager for the ACCESS program and also leads the quality control program at CAL. Each of these individuals brings expertise that will inform and direct the ASSETS assessment development project, including the management of approximately 15 additional staff members (approximately 10 FTEs).

WestEd: **WestEd** staff will serve as technical advisors to WIDA, providing expert guidance on issues related to the project's goals and objectives. **Robert Linquanti** will consult regarding EL population definition, policy, and accountability. **Dr. Aida Walqui** will bring expertise is in teacher professional development, teacher quality, and models for professional capacity building and outreach. **Dr. Edynn Sato** will facilitate coordination with RttT consortia (e.g., on technology approach; interoperability; item types; population definitions; accessibility and accommodation strategies), correspondence with Common Core State Standards, and validity frameworks for special population assessment. **Jeffrey Eng**, with his strong technology background and role as project manager and liaison with SMARTER Balance Assessment Consortium, will work on tasks related to interoperability.

University of California, Los Angeles: **Dr. Alison Bailey** and **Margaret Heritage** of UCLA will work with WIDA personnel to develop learning progressions for the English language associated with school success and career readiness. They will work with other UCLA staff to study the impact of the progressions on teacher assessment and instructional practices with ELs in a sample of K–12 content and ESL teachers. They will conduct analyses of these practices and report results to WIDA to refine ongoing assessment development.

Data Recognition Corporation and MetriTech, Inc: **Ms. Ara Lotzer**, Senior Project Manager at DRC, will provide overall coordination for the delivery of the online field testing test platform and

customer support. She has nine years of project management experience and more than five years of experience with online assessments.

MetriTech, Inc.: As MetriTech, Inc.'s Vice President of Operations, **Ms. Susan Feldman**, has more than 20 years of corporate and governmental experience organizing procedures and personnel to achieve time-sensitive objectives, including work for the U.S. Department of Commerce. She has managed the printing, distribution, scoring, and reporting of the WIDA Consortium's ACCESS for ELLs test since its inception in 2005 and has overseen the processing and scoring of millions of items, including over 30 million ELP constructed-response items since 2009.

Advisory Groups to guide and review ASSETS project work

- A **Steering Committee** composed of consortium SEA representatives will provide direction for the scope of the project, the design of the components, and policy directives.
- A **National Development Advisory Group** will advise on test content development. In addition to WCER and CAL key personnel, members of the advisory group will include **Alison Bailey** and **Margaret Heritage**, UCLA; **Edward Roeber**, Michigan State University; **Mary J. Schleppegrell,** University of Michigan; and **Robert Linquanti**, **Edynn Sato**, and **Aida Walqui**, West Ed; and an expert in computer-based teaching and learning and/or testing.
- A **Technical Advisory Committee** (TAC) will advise primarily on issues of psychometrics and other technical aspects of assessment. TAC members will include **Carol Chappelle**, University of Iowa; **Jamal Abedi**, UC Davis; **Aki Kamata**, University of Oregon; **Michael Hock**, Vermont Department of Education; and **Carol Myford**, University of Illinois at Chicago; **Lyle Bachman**, UCLA.