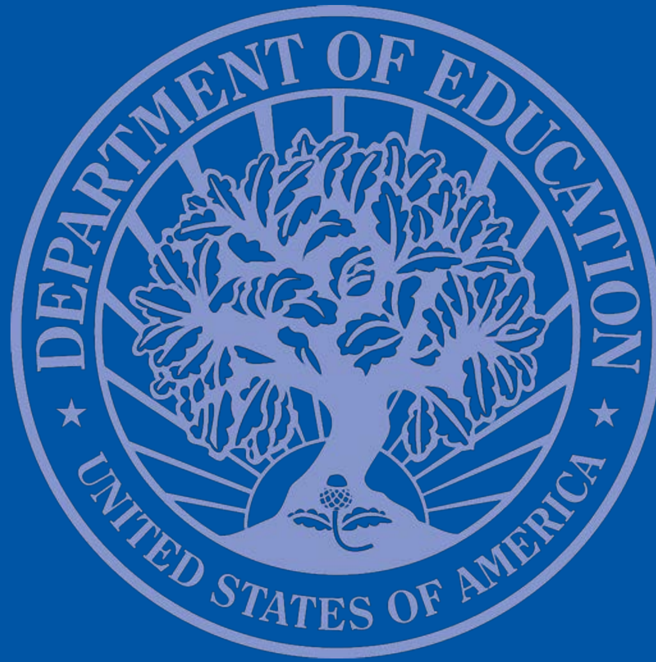

Literature Review Related to Assessment and Accountability Provisions Relevant to English Learners

U.S. Department of Education
Office of English Language Acquisition

This report was produced under U.S. Department of Education Contract No. ED-04-CO-0025/0020 with American Institutes for Research. Melissa Escalante served as the contracting officer's representative. The views expressed herein do not necessarily represent the positions or policies of the Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service, or enterprise mentioned in this publication is intended or should be inferred. For the reader's convenience, this publication contains information about and from outside organizations, including hyperlinks and URLs. Inclusion of such information does not constitute an endorsement by the Department.

U.S. Department of Education
John King
Secretary

Office of English Language Acquisition
Libia Gil
Assistant Deputy Secretary

September 2016

This report is in the public domain. Authorization to reproduce it in whole or in part is granted. Although permission to reprint this publication is not necessary, the citation should be U.S. Department of Education, Office of English Language Acquisition, *Literature Review Related to Assessment and Accountability Provisions Relevant to English Learners* Washington, DC, 2016.

This report is available on the Department's website at
http://www2.ed.gov/about/offices/list/oela/index.html

Availability of Alternate Formats

Requests for documents in alternate formats such as Braille or large print should be submitted to the Alternate Format Center by calling 202-260-0852 or by contacting the 504 coordinator via e-mail at om_eeos@ed.gov.

Notice to Limited English Proficient Persons

If you have difficulty understanding English, you may request language assistance services for Department information that is available to the public. These language assistance services are available free of charge. If you need more information about interpretation or translation services, please call 1-800-USA-LEARN (1-800-872-5327) (TTY: 1-800-437-0833), or e-mail us at ed.language.assistance@ed.gov. Or write to: U.S. Department of Education, Information Resource Center, LBJ Education Building, 400 Maryland Ave. SW, Washington, DC 20202.

Content Contact:

Melissa Escalante
202-401-1407
melissa.escalante@ed.gov

Literature Review Related to Assessment and Accountability Provisions Relevant to English Learners

September 2016

Prepared for:

U.S. Department of Education
Office of English Language Acquisition (OELA)

Prepared by:

Diane August and Rachel Slama

American Institutes for Research
1000 Thomas Jefferson Street, NW, Suite 200
Washington, DC 20007-3835

Robert Linquanti, Gary Cook, Ed Roeber, Li Cai

Technical Advisors to the Project

Contents

Page

Introduction ... 1
Process Used to Conduct the Literature Review ... 1
Parameters Guiding the Literature Review ... 1
Characterization of the Research Literature ... 2

Section 1. Development and/or Adoption of State ELP Standards ... 3
Development and Adoption of Appropriate and Challenging ELP Standards and Inclusion of Stakeholders in the Process ... 3
Methods to Evaluate Implementation and Uptake of ELP Standards ... 4
Coherence, Rigor, and Alignment or Linkage to Content-Area Standards ... 5

Section 2. Design and Development of the ELP Assessment System ... 8
Purposes, Claims, and Uses ... 9
Fairness ... 9
Standard Setting ... 10

Section 3. Technical Quality ... 14
Validity ... 14
Reliability ... 17
Scoring and Scaling ... 18
Comparability ... 19
Test and Data Security ... 19

Section 4. Using the ELP Assessment System for Accountability ... 20
Defining Empirically Based AMAO Criteria and Targets ... 21
Establishing and Validating the Title III Accountability System ... 23
Reporting ... 24

Areas for Further Research ... 25

References ... 26

Additional Resources ... 31

Appendix A. Panelists, Staff, and Partners ... 33
Panelists ... 33
Project Staff and Partners ... 34

Appendix B. Characteristics of Reviewed Studies ... 35

Introduction

This document reviews the literature related to assessment and accountability provisions relevant to English learners (ELs).

This review begins with a description of the process used to conduct the literature review, parameters for the review, and the characterization of the literature. The body of this review consists of four sections: (1) Development and/or Adoption of State English Language Proficiency (ELP) Standards, (2) Design and Development of the ELP Assessment System, (3) Technical Quality, and (4) Uses of an ELP Assessment System for Accountability. The research questions are described at the beginning of each section.

At the end of the document, we suggest *Areas for Further Research,*[1] some of which are related to outstanding problems of practice.

Process Used to Conduct the Literature Review

The project team convened an expert panel and a technical working group who contributed to the literature review through interviews and meetings. Appendix A includes the names of panelists and project partners along with their affiliations and areas of expertise.

The process used to generate the literature review consisted of the following six steps: (1) outline the parameters to guide the review, (2) identify the pool of possible studies to be included in the review, (3) reduce the pool of studies by determining their relevance to the research questions, (4) code study characteristics for relevant information, (5) review each resource to answer the research questions, and (6) compile summaries for each question. This process yielded an initial list of 112 articles, chapters, reports, and other resources that were candidates for inclusion in the literature review. The final review includes 48 studies.

Parameters Guiding the Literature Review

The review incorporates theoretical and empirical sources, including methodological and statistical analyses (empirical analyses of state and district data); technical reports; chapters in edited volumes; center reports; practice reports, such as the Council of Chief State School Officers (CCSSO) *Framework for English Language Proficiency Development Standards Corresponding to the Common Core State Standards and the Next Generation Science Standards*; dissertations; policy analyses; peer-reviewed journal articles and papers; and state reports. We excluded editorial and opinion pieces.

The project team employed the following inclusion criteria: The studies had to be (1) relevant to the research questions (independent of the method employed), (2) published between 1990 and 2014, and (3) focused on prekindergarten–Grade 12 education. Typically, the studies reviewed were specific to ELP assessments. The review draws on several studies related to content

[1] The studies reviewed were recommended by a panel of experts for their relevance to the research questions and do not represent a comprehensive search of all studies in the field.

assessments when there was no corresponding study to review that focused on ELP assessments. Occasionally, additional resources were used to provide context for the question summaries. The additional resources are listed at the end of the reference list.

Characterization of the Research Literature

Appendix B provides a summary of the studies reviewed, characterizations of the studies (publication type, study type, unit of analysis), and the section(s) that they inform. The literature review differs from literature reviews of experimental studies. To the extent possible, the resources reviewed were those that addressed questions related to Title III assessment and accountability provisions. Additional literature was incorporated when it provided general background information or when there were lessons to be learned from analogous studies conducted with content assessments that could inform ELP assessments. Relatively few of the studies were experimental studies. Instead, the literature was predominantly technical in nature and focused on preK–12 education. In terms of publication type, the most frequently appearing studies were peer-reviewed journal articles (10 studies), technical reports (eight studies), and policy papers (six studies). “Test” was the most common unit of analysis across the resources reviewed (14 studies). Other common units of analysis across the studies reviewed were other (nine studies), ELP standards (five studies), accommodation (four studies), content standards (two studies), and performance-level descriptors (three studies).

Section 1. Development and/or Adoption of State ELP Standards

The studies in Section 1 focus on the development of appropriate and challenging ELP standards and the inclusion of stakeholders in the development and adoption of ELP standards; methods used to evaluate the implementation or uptake of ELP standards; and the coherence, rigor, and correspondence or linkage[2] of ELP standards to academic content standards.

Title III of the No Child Left Behind Act of 2001 (NCLB) required states to establish state standards for English language proficiency that were derived from the four domains of speaking, listening, reading, and writing, and that aligned with achievement of challenging state academic content and student academic achievement standards as described in Title I, section 1111(b)(1) (sec. 3113(b)(2)).

Title I of the Every Student Succeeds Act (ESSA) also requires states to establish state ELP standards derived from the four domains of speaking, listening, reading, and writing, and aligned with the challenging state academic standards (sec. 1111(b)(1)(F)(i); s.1111(b)(1)(F)(iii)). ESSA further specifies that ELP standards must also address the different proficiency levels of ELs (sec. 1111(b)(1)(F)(ii)).

Development and Adoption of Appropriate and Challenging ELP Standards and Inclusion of Stakeholders in the Process

Before delving into studies that focus on the development and adoption of ELP standards, we present findings related to the development of academic achievement standards more generally. The same procedures used in establishing academic achievement standards could be applied to creating ELP standards. To set academic achievement standards, Hambleton (2001) suggested careful consideration of the composition and size of panels brought together to set expectations, panelists' understanding of the assessment used to measure student knowledge and skills associated with these expectations and the uses of this assessment, qualifications of the panelists, opportunities for panelists to take portions of the assessment, adequacy of panelist training, and panelist participation in evaluation of the process. With respect to time and resources, the evaluation criteria included opportunities for field-testing and revision of the method used to set the standards. With respect to the appropriateness of assessment methodology, evaluation criteria include the use of clear performance-level descriptions, the use of feedback (whether an iterative process was used), process efficiency, the grounding of judgments in performance data, the

[2] Some scholars make a distinction between the terms "correspondence" and "alignment." The Council of Chief State School Officers (CCSSO, 2012) noted that alignment typically refers to a comparison between equivalent "artifacts," such as standards, assessments, or curricula (e.g., ELP standards to an ELP assessment). Correspondence refers to a comparison between nonequivalent artifacts. For example, the English Language Proficiency Development Standards Framework corresponds to the Common Core State Standards (CCSS) and Next Generation Science Standards (NGSS) because the language practices do not encompass all standards in the CCSS and NGSS (CCSSO, 2012). Some scholars (e.g., Bailey, Butler, & Sato, 2007) used the term "linkage" to refer to the linking of standards across different content areas on a common dimension. Consistent with the ESSA, from this point forward this report uses the term alignment to refer to both alignment and correspondence. It also uses the term linkage as defined by Bailey et al. (2007).

gathering of evidence concerning the validity of standards, thorough documentation of all aspects of the process, and clear communication of results.

According to the English Language Proficiency Development Framework (CCSSO, 2012), the development of ELP standards should be guided by both theory and research (CCSSO, 2012). ELP standards should reflect "a sequence of language development that is grounded in theoretical foundations, responsive to the various backgrounds of students, and attuned to the growth trajectories of different ELLs" (CCSSO, 2012, pp. 4–6).

States generally involve relevant stakeholders in the following phases and activities related to standards development and adoption: (1) development of ELP standards, (2) evaluation of the alignment between ELP and content standards, and (3) evaluation of the alignment of English language arts and ELP standards. For example, in the development phase of the ELP standards in California, the process for developing and adopting standards involved numerous focus groups, public hearings, public comment, and the convening of an expert panel to review drafts of the standards and provide ongoing input and guidance (California Department of Education, 2012).

Methods to Evaluate Implementation and Uptake of ELP Standards

The methods used to evaluate the implementation and uptake of ELP standards may depend on the level at which states are interested in examining standards implementation. Standards implementation may occur at state, district, and school levels. Information related to implementation has been collected through large-scale surveys (e.g., CCSSO & Wisconsin Center for Education Research [WCER], 2010; Tanenbaum et al., 2012), questionnaires, one-on-one interviews, and document reviews (e.g., CCSSO & WCER, 2010).

At the state level, Tanenbaum et al. (2012) interviewed state Title III officials on the use of their ELP standards. State Title III officials were asked whether ELP standards were used to inform professional development or support, select or develop the state ELP assessment, approve instructional programs or materials, or monitor classroom instruction. At the time of the study interviews (2009–10), 23 states and the District of Columbia reported using their state ELP standards to inform professional development. Six states reported using ELP standards to develop the ELP assessment, five states reported using the ELP standards to approve instructional programs and materials, and five states reported using ELP standards to monitor classroom instruction. At the district level, Tanenbaum et al. (2012) gathered data through 12 case studies on how districts used ELP standards to guide classroom instruction. Administrators in five case study districts reported using ELP standards to make decisions about district-required curricula. Two districts reported developing and requiring high school curricula based on ELP standards. One district official noted that because the district curricula were grounded in the ELP standards, teachers were more aware of the standards and could adapt their instruction according to their students' needs. Seven case study districts reported that ELP standards drove the planning of professional development activities for teachers of ELs and mainstream teachers in districts with high concentrations of ELs.

At the school level, the Survey of Enacted Curriculum (CCSSO & WCER, 2010) surveys the teachers responsible for the English language development of ELs on topics such as instructional

activities for ELs, instructional influences, classroom instructional readiness, teacher opinion and beliefs, professional development, and formal course preparation.

An example of a question that relates to the implementation and uptake of ELP standards is asking teachers to indicate the extent to which a list of factors, including the state ELP standards and the district's curriculum framework, standards, or guidelines, supports or constrains their practice in teaching ELs. The survey results are reported online, and provide teachers and leaders a method to evaluate the degree of alignment between instruction and state standards and assessments.

Coherence, Rigor, and Alignment or Linkage to Content-Area Standards

A strong linkage between ELP standards and state content standards in the core subject areas will help ensure that ELs are exposed to types of language that will help them be successful in academic contexts (Bailey et al., 2007). The WIDA standards, for example, reference social and instructional language, the language of English language arts, the language of mathematics, the language of science, and the language of social studies (Kenyon, MacGregor, Li, & Cook, 2011).

Bailey and colleagues (2007) noted that ELP standards should provide both detailed descriptions of the "degree of complexity of the lexical and grammatical forms expected of students at each [English language development] level" (p. 75) and the language demands required for demonstrating content-area mastery (Bailey et al., 2007).

Under NCLB, states were required to show "linkage"[3] between state content standards and ELP standards (see Bailey et al., 2007, for a review). However, the U.S. Department of Education did not provide guidance to states about the procedures or methodology by which this linkage should be established. Guidance would have been welcome because unlike alignment between assessments and standards—which have a relatively well-developed set of methodologies to determine how well test items measure the skills described in content standards documents—there were no set procedures for establishing evidence of linkage between language proficiency standards and standards for different content areas (Bailey et al., 2007).

Bailey and colleagues suggested establishing a linkage by identifying language demands common in ELP and content standards. Language demands can be associated with discrete linguistic skills, such as syntactic structures used to communicate particular types of information (e.g., compare and contrast structures such as *greater than* and *less than*). Language demands can also be at the functional level of text, such as explanations and descriptions in textbooks and classrooms. However, identifying language demands common across content and ELP standards is particularly challenging, because content standards typically do not reference the language structures and functions that students need to access the content or to demonstrate proficiency in a given content area.

[3] Bailey et al. (2007) referred to examining linkage as the evaluation of the degree to which content standards (e.g., English language arts or science) overlap with ELP standards with respect to the language demands (both implicit and explicit) placed on students.

Because of this challenge, Bailey et al. (2007) described a second approach to identifying the language required for a student to demonstrate proficiency in a particular content area. This approach entailed comparing the language demands in the ELP standards with published content-area instructional materials, such as standards-based textbooks, lesson plans, and other published curriculum, rather than with the content standards. The authors suggested that the results from this linking procedure could help the state determine how well the state ELP standards capture the language demands required of students to meet state content standards.

In addition to the approaches described by Bailey et al. (2007), CCSSO developed English language arts, mathematics, and science tables that identify features of classroom language, students' language use, and language tasks in each of the disciplines to support the development of ELP standards that reflect the language expectations and underlying language practices embedded within the CCSS and NGSS standards (CCSSO, 2012). The tables also support the evaluation of extant standards (an example is provided in Exhibit 1 of this report).

Exhibit 1. Sample K–3 English Language Proficiency Descriptors Supporting the Common Core State Standards

Level	Modality		Selected Language Practices Identified in the CCSS	
			Construct Explanations (ELA, Math, Science)	Argue from Evidence (ELA, Math, Science)
Level 1	Oral	Receptive	• Can begin to guess intelligently at topic. Continues to listen past frustration to make sense of incoming speech.	• Can comprehend that speakers disagree by relying on his/her experience in L1 interaction.
Level 1	Oral	Productive	• Can respond to choice questions in which an explanation is presented.	• Can begin to express agreement or disagreement with gestures, basic utterances, memorized chunks, L1, and intonation.
Level 1	Written	Receptive	• Can guess intelligently at the topic of written explanations when these are accompanied by illustrations.	• No examples of this practice at this age-band.
Level 1	Written	Productive	• Can reproduce drawings or diagrams of known items or ideas used in class that explain how something works.	• No examples of this practice at this age-band.
Level 2	Oral	Receptive	• Can comprehend most teacher explanations if supported by gestures, illustrations, and other scaffolds.	• Can comprehend main points of others' arguments if provided with background information and other scaffolds.
Level 2	Oral	Productive	• Can draw from and build upon others' explanations using gestures, pictures, and memorized language chunks.	• Can ask questions for clarification about others' arguments. Can draw from and build upon segments of others' arguments.
Level 2	Written	Receptive	• Can comprehend written explanations when he/she has knowledge about the topic and can draw from images.	• Can identify argument and evidence given in a text if provided with support and examples.
Level 2	Written	Productive	• Can draw from and build upon basic illustrated written explanations if provided with examples.	• Can draw from and build upon written arguments and statements presenting evidence if provided with examples.
Level 3	Oral	Receptive	• Can comprehend almost all key points of teacher explanations that are not supported by gestures or other scaffolds.	• Can comprehend almost all points of disagreement in a discussion. Can distinguish arguments not supported by evidence.
Level 3	Oral	Productive	• Can draw from and build upon explanations produced by other students, using appropriate disciplinary terminology.	• Can draw from and build upon others' arguments and statements that provide evidence using gestures, pictures, memorized language chunks and other communicative strategies.
Level 3	Written	Receptive	• Can comprehend written explanations of topics covered in class. Will rely to some degree on illustrations and other graphic materials.	• Can comprehend arguments and identify evidence in age-appropriate written texts on topics covered in class. Will rely to some degree on illustrations and other graphic materials.
Level 3	Written	Productive	• Can produce written explanations of processes with the support of examples, can begin to rely less on illustrations.	• Can write out the arguments and supporting evidence he/she can produce orally. Can continue to draw from and build upon examples.

Source: Reproduced with permission from CCSSO (2012, p. 46).

Section 2. Design and Development of the ELP Assessment System

The studies in Section 2 focus on the design and development of the ELP assessment system; the appropriateness of using ELP assessment results for various purposes (e.g., initial EL identification, instructional program placement, monitoring language development, and classifying ELs as English proficient); the ways in which states ensure during the design and development phase that their ELP assessments are fair and accessible; and the methods and processes currently used for establishing ELP levels, including use of the content assessment to set the "English proficient" performance standard on an ELP assessment.

Titles I and III of NCLB require school districts to have procedures in place to identify students who had a primary or home language other than English. Students who had a primary or home language other than English had to be assessed using valid and reliable measures to determine their levels of English proficiency in four domains (speaking, listening, reading, and writing). School districts had to provide notice within 30 days from the beginning of the school year to all parents of ELs regarding their student's identification and placement in a language instruction educational program (sec. 1112(g)(1)(A) and sec. 3302(a)).

NCLB provided a definition for who was considered an EL[4] (sec. 9101(25)), although states were granted flexibility in operationalizing who was included in the EL subgroup. States and districts varied in the criteria used to exit students from EL status although in 2009–10. Tanenbaum et al. (2012) reported that 49 states and the District of Columbia required or recommended the use of the state ELP assessment to exit students from EL status.

Title I of ESSA requires all ELs—including those with disabilities—to participate annually in ELP assessments that are aligned with ELP standards and measure the four domains of language proficiency: reading, writing, speaking, and listening (sec. 1111(2)(G)(i)).

Title I of ESSA established in law (sec.1111(b)(1)(B)(E)) alternate performance standards for alternate content-area assessments for students with the most significant cognitive disabilities, but there is no provision related to alternate performance standards for ELP assessments for ELs with the most significant cognitive disabilities.

Title III of ESSA requires states to establish, in consultation with local education agencies (LEAs)—representing the geographic diversity of the state—standardized statewide procedures for identifying students as ELs and for classifying ELs as English proficient. This provision is intended to reduce the variation within states of which children are considered ELs.

Title III of ESSA also requires LEAs to assess all students with a primary or home language other than English within 30 days of school enrollment (sec. 3111 (b)(2)(A)).

[4] No Child Left Behind used the term "limited English proficient" to refer to English learners.

Purposes, Claims, and Uses

An ELP assessment system typically includes several major components that correspond to the multiple purposes that such a system must serve:

- The identification, screening, and classification process associated with ELs;
- Summative assessments that measure ELs' attainment of English language proficiency and track progress for accountability purposes;
- Interim assessments that provide useful feedback to improve instructional planning and other decisions regarding a student's schooling;
- Alternate assessments for ELs with the most significant cognitive disabilities; and
- Any other associated measures that are formally required as part of the ELP assessment system.

The assessments that constitute an ELP system are used for a variety of purposes: to classify students into EL/non-EL categories and to reclassify them when they are proficient enough in English to fully participate in general education classrooms and content-area assessments administered in English; to determine placement in language instruction educational programs; and to monitor students' progress while in those programs for various purposes, such as grouping students for instruction, detecting gaps in the curriculum, planning professional development, and guiding local improvement (Tanenbaum et al., 2012).

The assessments that make up an ELP system serve multiple purposes. However, according to the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council of Measurement in Education [NCME], 2014), no assessment will serve all desired purposes equally. Therefore, test users should decide which uses of the ELP assessment are of highest priority and should develop their assessment in accordance with these uses. In addition, for the designated purposes, test users must determine whether the assessments that compose the ELP system are accessible and provide valid inferences for the students being assessed.

Fairness

The studies in the *Fairness* subsection of Section 2 focus on the design and development of fair and accessible ELP assessments. The studies in the fairness subsection of Section 3 focus on fairness during the administration of the assessment, such as by providing carefully considered accommodations; and during interpretation of test performance, such as by considering student background characteristics. This subsection and the corresponding subsection in Section 3 focus primarily on ELs with disabilities. Among other things, the sections report on best practices for determining when ELs with disabilities should receive the alternate ELP assessment.

Assessment systems should be designed to be accessible and to provide valid inferences for the widest range of students possible (Albus & Thurlow, 2008). To ensure fair and accessible ELP assessments for ELs with disabilities during the development phase, states and consortia should understand the population of students being tested, ensure the test development team includes

individuals with expertise in relevant areas of test and item development, use universal design principles in the construction of test and item development, include students with disabilities in item try-outs and field testing, and ensure items and tests go through committee review (Thurlow, Liu, Ward, & Christensen, 2013).

Standard Setting

The studies in the *Standard Setting* subsection focus on the establishment of ELP levels, including the number of levels that should be established and the use of additional information (including content-area assessments in setting the “English proficient” performance standard). These research questions are linked, because once the number of ELP levels has been decided, performance level descriptors (PLDs) are created and used to describe each performance level on an ELP test. Then, experts determine the test scores associated with each performance level, and notably, the “English proficient” performance standard—the level at which states and consortia believe students have developed sufficient English and do not require further language instructional support (Gary Cook, personal communication, July 11, 2016).

The studies reviewed did not provide information related to evidence-based methods and processes for establishing ELP levels. We draw on the literature on standard setting for content assessments because of the parallels between the processes used to set proficiency levels on content assessments and those used to set proficiency levels on ELP assessments.

Evidenced-based standard setting uses procedures that include selecting and training panelists, collecting panelists’ judgments of where the cut scores should be set, and selecting numerical cut scores that reflect panelists’ judgment about what the PLDs mean. Furthermore, data from external measures are used to validate claims about expected student performance at each proficiency level. This process has been applied to ELP assessments. For example, Powers, Williams, Keng, and Starr (2014) described how evidence-based standard setting was used to recommend performance standards on the Texas ELP assessment. Student data were analyzed to determine whether students scoring at the proficient level on the ELP assessment would also score proficient on the state content assessment one year later.

Justifying the Number of ELP Levels

Cook (Gary Cook, personal communication, July 11, 2016) noted that statistical models can identify proficiency levels on ELP assessments, but the results vary depending on which of the four domains (listening, speaking, reading, or writing) is examined. Research also has suggested that even within domains, particular linguistic features vary in the number of linguistic levels that could be associated with them.

In practice, however, standard developers select the number of ELP levels based on how they believe teachers of ELs organize their instruction to support ELs with varying levels of language proficiency (Gary Cook, personal communication, July 11, 2016) rather than on the number of actual levels that exist. This observation is consistent with California’s rationale for the use of three ELP levels because it mirrors the common practice in the state to group ELs into three groups for instructional purposes. Although WIDA has five ELP levels, Gottlieb (2013) suggested that three levels might be better because teachers commonly create three groups of

language learners when differentiating language instruction (beginning, intermediate, and advanced). She suggested possibly collapsing WIDA's five ELP levels into three levels with some overlap (1–3, 2–4, and 3–5) so that they are more consistent with teaching practices. Although this example suggests the number of performance levels on ELP assessments may be a practical decision, the literature reviewed on content-area assessments[5] has suggested that the number of performance levels set on content-area assessments should also demonstrate a logical or empirical connection to the assessment purpose (Camilli, Cizek, & Lugg, 2001; Kane, 2001).

Process for Establishing Performance Levels

The studies focusing on establishing performance levels on content-area assessments (e.g., Hambleton, Pitoniak, & Copella, 2012) have suggested a systematic, transparent, and well-documented process for establishing performance levels, with input from multiple relevant stakeholders, largely through panels.

During the standard-setting process on content assessments, several types of data may be used to help panelists validate claims about characteristics of students at each performance level (Hambleton, 2001). Panelists may examine actual performance data such as item difficulty values when discussing how they rated student performance on particular test items or tasks. Panelists might be presented with overviews of ratings from all of the panelists. In addition, standard-setting panels may be provided with "consequences data" such as the percentage of test takers who would be classified into each performance category (e.g., advanced, proficient, basic, below basic) if the panel relied on the performance standards generated during the standard-setting session. Providing standard-setting panels with additional data may influence panelists' behavior (and the resulting performance standards), and therefore it is important to carefully consider whether and how this feedback is provided (Hambleton et al., 2012).

Using Content-Assessment Data to Set the "English Proficient" Performance Standard

The required alignment between ELP standards and content standards may provide some justification for the use of ELs' content test performance to set the "English proficient" performance standard. The federal definition of English learner[6] as a student who has difficulties in speaking, reading, writing, or understanding the English language that may be sufficient to deny the individual the ability to meet challenging state academic standards (CSSO, 2016) suggests that having adequately addressed ELs needs is linked to their performance on state content assessments (Cook et al., 2012). Cook et al. (2012) noted, "These aspects of the federal law imply an expected relationship between students' ELP and levels of academic proficiency when content is assessed in English" (p. 7). Accordingly, state policymakers can use the state content assessment to help determine what levels of linguistic and academic performance

[5] Content-based performance standards and English proficiency levels are related: they are socially mediated artifacts categorizing distinct representation of students' ability. Standard setting and establishing ELP levels are both socially mediated processes, informed by research (Gary Cook, personal communication, July 11, 2016). We draw on literature related to establishing performance levels on content assessments that may be relevant to establishing ELP levels on ELP assessments.
[6] Cook, Linquanti, Chinen, and Jung (2012) referred to the definition of limited English proficient students under NCLB. The definition of English learner under ESSA is nearly the same (see CCSSO, 2016, for a review).

determine their definition of EL (Cook et al., 2012). The authors (Cook et al., 2012) illustrated how states can examine the ELs' performance on ELP and content assessments to determine where the English proficient performance standard should be set.

However, it should be noted that scholars have concerns related to construct validity, measurement, and accountability when content tests are used to make judgments about ELs' level of English language proficiency (e.g., Working Group on ELL Policy, 2011).

Kenyon et al. (2011) cited studies conducted by Cook (2009) that used EL performance on content-area assessments to determine the point on a language proficiency assessment at which ELs should be exited from EL services. More specifically, Cook (2009) examined the relationship between student performances on state reading and mathematics content assessments and on WIDA ACCESS for ELLs. It was found that between proficiency levels 4.5 and 5.0 (depending on the state, grade, and subject), state content tests may be more accurately measuring students' content knowledge, "with less interference" (as cited in Kenyon et al., 2011, p. 398) from students' level of ELP.

Alignment of the ELP Assessment With ELP Standards

This subsection focuses on the methods that can be used to demonstrate that an ELP assessment adequately measures the knowledge and skills described in the ELP standards. The degree to which an assessment (through its test items) covers what the test is intended to measure is referred to as alignment (Cook, 2006). In the past, studies that evaluated the degree of alignment between an assessment (typically content assessments) and content standards focused on the matching of test items to standards. In the context of aligning ELP assessments to ELP standards, the goal is to examine (1) the degree to which the ELP standards are "covered" on the test, (2) the alignment between the linguistic complexity of test items and the complexity of the ELP standards, (3) breadth of coverage of the ELP standards (composed of the range of what is covered and the balance of what is covered), and (4) linkage to state academic content standards (Cook, 2006).

Cook (2006) outlined a three-step procedure for aligning ELP tests to ELP standards, which includes (1) setting up the alignment, (2) conducting the alignment study, and (3) analyzing the degree of alignment between ELP standards and the ELP assessment. A summary of each step is described next.

This first step, *setting up the alignment*, consists of identifying relevant stakeholders, developing alignment protocols, deciding who participates in the alignment review, and deciding how to collect alignment information. The second step, *conducting the alignment study*, consists of convening an independent alignment committee for preassigned ELP[7] standards and grade spans; training the alignment committee on the alignment process; having the alignment committee assign linguistic difficulty levels to ELP standards and proficiency levels, if these are separate; having the alignment committee independently assign, to each test item, a linguistic difficulty level and a match to an ELP standard and a proficiency level; and having the alignment

[7] Note that Cook (2006) referred to ELP standards as English language development (ELD) standards, but we use the more common term ELP standards in summarizing the information from that report.

committee identify links to the state's academic content standards (e.g., English language arts and/or mathematics). The third step, *analyzing the degree of alignment*, consists of identifying the assessment's coverage of ELP standards, identifying the assessment's linguistic appropriateness relative to ELP standards, identifying the assessment's breadth of ELP standards coverage, and identifying the assessment's linkage to the state's content standards.

Section 3. Technical Quality

The studies in Section 3 focus on the validity and reliability of ELP assessments, scoring and scaling, comparability (specifically the comparability of scores on paper-delivered and computer-delivered accommodations on the ELP assessment), and test and data security.

As noted in Section 2, Title I of NCLB required states to administer annual assessments of ELP that measured students' oral language, reading, and writing skills in English (sec. 1111(b)(7)). Title VI of NCLB made grants available to states in part to ensure the continued validity and reliability of state assessments. For example, section 6111(2)(B) specified that funds could be used to develop or improve assessments of ELP necessary to comply with section 1111(b)(7).

Title I of ESSA specifies that state education agencies must have appropriate procedures and safeguards in place to ensure the validity of the assessment process (sec. 1111(g)(2)(I)). As described in Section 2, states are required by federal law (ESSA sec. 1111(2)(G)(i)) to assess their ELs annually based on standards that are derived from the four recognized domains of language proficiency: reading, writing, listening, and speaking (ESSA sec. 1111(b)(1)(F). In addition, states must monitor ELs' progress in these domains and in comprehension. Based on these requirements, states typically report separate domain scores (listening, speaking, reading, and writing) and composite scores (composed of oral literacy, comprehension, and overall).

The Department of Justice and Office for Civil Rights' (2015) letter also provides guidance related to valid and reliable assessment of English proficiency for all ELs, stating: "[T]he English language proficiency assessment must meaningfully measure student proficiency in each of the language domains, and, overall, be a valid and reliable measure of student progress and proficiency in English" (p. 33).

The Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014) describe the characteristics of high-quality assessments in general, as well as the processes that states can employ to ensure that tests are valid and technically defensible for their intended uses. In applying these standards to ELs, two considerations emerge: (1) There is a need to collect evidence about the technical quality of all measures that make up the ELP assessment system, not only about the language proficiency assessment used to designate students as ELs; and (2) there is a need to pay special attention to the gathering of data in support of validity, reliability, fairness, and comparability of assessment results for students with the most significant cognitive disabilities who are ELs and require alternate ELP assessments.

Validity

The studies in the *Validity* subsection focus on the importance of alignment between the ELP assessment and the knowledge and skills ELs are expected to master, the use of the ELP assessment and other empirical data collected on ELs for various purposes (e.g., instructional placement, EL classification, progress, attaining English proficiency, and exit), and the validity of the ELP assessment for EL subpopulations.

Alignment Across the ELP Assessment, Curriculum, and Instruction

Pellegrino, Chudowsky, and Glaser (2001) noted the importance of alignment across the assessment, curriculum, and instruction if testing is to achieve its intended goals. A potential negative consequence of testing is the narrowing of instruction to address only the particular aspects of the curriculum that are assessed. “This curricular reductionism often shortchanges children because they fail to encounter the full richness of the subjects they are studying” (Popham, Keller, Moulding, Pellegrino, & Sandifer, 2005, pp. 126–127).

The level of alignment of the content of the ELP assessment with ELP standards is one way to assess this alignment (AERA, APA, & NCME, 2014). See the *Alignment of the ELP Assessment With ELP Standards* in Section 2 for further elaboration.

Use of the ELP Assessment and Other Empirical Data on ELs

For an assessment to be valid for a particular purpose (e.g., initial program placement), there must be evidence that supports using the assessment for that purpose. For example, data should show that an assessment intended to help make placement decisions about ELs does in fact lead to appropriate placement decisions.

Although the studies in this section described the purposes of ELP assessments (e.g., guiding instructional decisions and EL instructional placement, as well as documenting progress and ELP attainment and exit), they generally did not provide evidence of the appropriateness of using ELP assessment for particular purposes.

Additionally, research on the use of other empirical data on ELs (e.g., content assessment scores, teachers’ judgment, parents’ input) to determine a student’s ELP level or exit from EL instructional services are limited. However, there is some empirical support for the use of ELs’ content-area assessment performance data (e.g., outcomes on English language arts assessments) to set the English proficient performance standard. Cook, Boals, Wilmes, and Santos (2008) suggested that the “English proficient” performance level could be set by identifying the level of performance on the ELP assessment at which language proficiency no longer inhibits ELs’ performance on state content assessments. Kenyon et al. (2011) found that lower levels of English language proficiency interfered with performance on content assessments. At higher levels of English language proficiency—particularly levels 4.5 to 5 on WIDA ACCESS for ELLs®—these assessments better reflected content knowledge because English language proficiency no longer interfered.

Several of the reviewed studies (Abedi, 2008; Douglas & Mislevy, 2010) are relevant to procedures used to exit ELs from EL instructional services. Douglas and Mislevy (2010) highlighted the general importance of using multiple measures to make high-stakes testing decisions. The authors’ (Douglas & Mislevy, 2010) guidance can be applied to the use of multiple criteria to reclassify ELs as fluent English proficient. Other researchers have noted that external criteria such as demonstrated learning in authentic instructional environments may be useful for evaluating ELs’ language development (Gottlieb, 2013).

Validity of the ELP Assessment for EL Subpopulations

The studies in the *Fairness* subsection in Section 2 of this report focus on the development of fair and accessible ELP assessments. The studies in this subsection focus entirely on accommodations, because accommodations can improve the validity of inferences from the ELP assessment for different subpopulations of ELs when background or disability is believed to impede measurement of the intended construct (Albus & Thurlow, 2008). Accommodations are changes in test presentation (i.e., how test items are presented to the test-taker) or response formats (i.e., changes in how a test-taker may respond while being assessed) that do not alter the focal construct. On content-area assessments administered in English, ELs need accommodations to address their limited English proficiency, and ELs with disabilities need accommodations to address their disability. On ELP assessments, however, ELs with disabilities primarily need accommodations related to their disability, with support related to their limited English proficiency if test instructions are not comprehensible. Therefore, this section focuses on accommodations for ELs with disabilities.

Christensen et al. (2014) provided examples of accommodations on ELP assessments for ELs with disabilities. Those related to test presentation include using large print for ELs who are vision impaired, and interpreting directions in sign language for ELs who are hearing impaired. Examples of accommodations on the ELP assessment related to response format include allowing ELs with specific learning disabilities to write in test booklets or to use a proctor or scribe.

States vary in the types of accommodations that are permitted or prohibited on the ELP assessment overall as well as the accommodations that are permitted or prohibited for each domain of reading, writing, speaking, and listening (Albus & Thurlow, 2008; Christensen, Albus, Liu, Thurlow, & Kincaid, 2013; Liu et al., 2015). In 2013, 37 states required accommodations (as needed) for ELs with disabilities on ELP assessments. Some accommodations—such as Braille, large print, amplification, and magnification equipment—are widely accepted across states. Other accommodations are increasingly being permitted, such as native language translation of instructions, whereas some are more controversial (e.g., screen readers). There is variation across states in the use of sign language for reading directions and questions or expressing responses.

With regard to EL participation in alternate assessments, the group of ELs eligible to take the alternate ELP assessments is typically students who are considered to have the most significant cognitive disabilities (National Center on Educational Outcomes [NCEO], 2014; WIDA, n.d.). NCEO (2014) reported that 32 percent of states surveyed indicated that their ELs with significant cognitive disabilities participated in alternate ELP assessments, compared with approximately 70 percent of states that indicated that their ELs with significant cognitive disabilities participated in some or all of the general ELP assessment.

Consistent with federal guidelines (U.S. Department of Education, 2014, 2015; U.S. Department of Justice & U.S. Department of Education, 2014), many of the resources reviewed (e.g., California Department of Education, 2015; NCEO, 2014; WIDA, n.d.) emphasized the central role of the individualized education program (IEP) team in deciding the extent to which ELs participate in alternate ELP assessments. When making assessment participation decisions about

ELs with disabilities, it is general best practice for IEP teams to include representation of all professionals tasked with educating ELs with disabilities (e.g., English language development specialists, special education teachers, content teachers), parents, and the student, when appropriate (Thurlow et al., 2013). Written guidelines and documentation should be provided to determine the use of alternate assessments, and states should provide evidence of training for educators who administer alternate assessments (Christensen et al., 2004).

WIDA ACCESS guidelines also indicate that IEP teams should play a key role in determining the ELs who participate in alternate ELP assessments. States follow their individual state education agency’s guidelines to determine participation in alternate ELP assessments. ELP assessment consortia also provide additional participation guidelines. For example, ACCESS guidelines provide recommendations for WIDA states. These guidelines suggest that ELs should meet all of the following criteria in order to participate in the alternate ELP assessment: (1) the student is classified as an EL, (2) the student has a significant cognitive disability and receives special education services under the Individuals with Disabilities Education Act of 2004, (3) the student requires “extensive direct individualized instruction” and support to make measureable gains in the grade- and age-appropriate curriculum, and (4) the student is participating or will participate in the alternate content-area assessment based on alternate achievement standards.

English Language Proficiency Assessment 21 (ELPA21), a consortium of states designing and developing an ELP assessment system, does not offer an alternate ELP assessment for students with the most significant cognitive disabilities. This group recommends that individual states work to develop their own alternate assessments.

Best practices related to accommodations for ELs on content-area assessment might be extended to ELP assessments to improve their validity. These best practices include conducting statewide studies to examine the appropriateness of accommodations (e.g., evaluating the impact of a particular accommodation on assessment scores). Other practices include providing materials to clarify who should receive accommodations (e.g., a flowchart of accommodations decision making), the accommodations that are permitted (e.g., a table demonstrating accommodations that are permitted/not permitted), and for whom the accommodations are permitted (e.g., guidelines that state which accommodations are allowed for IEP students and those with 504 plans). Moreover, decisions concerning accommodations should be made for individual students (rather than groups of students) by an IEP team with appropriate membership and training (Thurlow et al., 2013).

State education agencies can also increase the validity of ELP assessments by clearly communicating federal and state participation requirements related to ELs with disabilities to districts and schools (Liu et al., 2013).

Reliability

Reliability refers to the consistency of measurement and includes internal consistency, generalizability of scores and classification accuracy. The studies in this *Reliability* subsection focused entirely on classification accuracy—measures of the extent to which decisions classifying an EL into a particular language proficiency level on the basis of student performance on the ELP assessment agree with decisions made on the basis of results of that same student

being hypothetically tested on all parallel forms of the assessment (i.e., accuracy) or on a parallel form of the assessment (i.e., consistency; National Research Council Committee on the Assessment of 21st Century Skills, 2011). Typically, summary indices are generated that capture the overall accuracy and consistency of a particular ELP assessment, such as the percentage of students who are consistently placed above and below the cut scores (based on statistical models) denoting the different language proficiency levels.

In the case of the ELP assessment, because scoring at the “English proficient” performance standard has consequences for ELs (e.g., placement in particular instructional programs, access to particular courses; exit from EL services), examining classification consistency estimates is important. However, none of the studies specifically addressed the level of classification consistency estimates required for ELP assessments to be considered reliable; nor did they provide an adequate level of classification consistency by grade, grade band, or discrepancies in consistency across EL subgroups, especially as the estimates relate to score values that characterize ELs at the English proficient performance standard.

Scoring and Scaling

The studies in the *Scoring and Scaling* subsection focus on establishing weighted composites—or the statistical importance assigned to each of the four domain scores of the ELP assessment (listening, speaking, reading, and writing) to generate one score that captures students’ overall English language proficiency. This subsection examines states’ justification for use of a particular method of establishing weighted composites and methods for validating composites scores.

The reviewed literature describes common state and consortium practices for constructing weighted composites and the implications of variations in these practices for accountability and progress monitoring (e.g., Linquanti & Cook, 2013). Some states weight the four domain scores equally (e.g., California), whereas other states and consortia weight reading and writing more heavily than speaking and listening (e.g., ACCESS for ELLs). Cook (2014) addressed the creation of optimal weighting for composite scores, specifically for alternate ELP assessments, but the method could easily be applied to regular ELP assessments (Gary Cook, personal communication, September 21, 2016).

Linquanti and Cook (2013) recommended that states articulate a clear rationale for the construction of their ELP assessment composites and provide evidence of the efficacy of their weightings as an important part of the validity argument for the ELP assessment that may inform the interpretation of test results (see also AERA, APA, & NCME, 2014). Although the studies reviewed suggested that states articulate a clear rationale for the construction of their ELP assessment composite and that states provide evidence of the efficacy of their weighting, most studies reported on the procedures used rather than on the states’ rationale and justification for how they established weighted composites.

Researchers who work in this area have recommended that ELP composite scores, including alternate composite scores, should predict student outcomes, including performance on content assessments (Cook, 2013; Kenyon et al., 2011). Accordingly, scholars have tested methods of

validating ELP composite scores using student performance on content-area assessments (Cook, 2013; Francis, Tolar, & Stuebing, 2012; Kenyon et al., 2011).

In a study to identify a procedure for creating composite scores on ELP assessments without using all four domain test scores, Cook (2013) found that ELP composite scores predicted performance on content assessments, concluding that performance on content assessments could be used to compute composite scores.

Francis et al. (2012) found that the reading and writing domains are the most predictive of ELs' content-area performance of the four ELP domains, which may provide empirical support to justify weighting reading and writing more heavily than speaking and listening.

Comparability

Although the reviewed literature does not specifically address the question of comparability for scores on paper-delivered and computer-delivered accommodations of an ELP assessment, three of the reviewed studies (Abedi, 2014; Abedi & Ewers, 2013; Liu, Ward, Thurlow, & Christensen, 2015) suggested benefits as well as cautions related to administering accommodated computer-delivered content-area assessments compared with accommodated paper-and-pencil assessments.

Computer-delivered assessments solve many of the challenges related to providing accommodated paper-and-pencil tests, such as those related to standardization, validity, differential impact of accommodations on different students, and feasibility of implementation (Abedi & Ewers, 2004). For example, on computer-delivered assessments, accommodations can be easily turned on or off for individual students. On the ELPA21 assessment, each testing platform has a Personal Needs Profile (PNP), which designates features a student needs and that will be made available to that student during the assessment. All embedded designated features must be activated via the PNP prior to testing (ELPA21, 2015).

Liu et al. (2015) cautioned that for ELs with disabilities, there might be a tendency for IEP teams to select every possible accommodation on content-area assessments, with the impression that more accommodations cannot negatively impact a student's test score, but that some of the accommodations that do not directly address an EL's disability or disabilities could actually exacerbate the disability. Abedi (2014) countered that algorithms for selecting accommodations could be programmed into a computer based on individual student data and thus ameliorate this issue.

Test and Data Security

The study reviewed in this section (Olson & Fremer, 2013) is a guidebook for states[8] on many issues relating to test security for large-scale assessment programs. Topics include preventing and detecting test security irregularities (e.g., cheating, test piracy) and investigating suspected or confirmed cases of potentially improper or unethical testing behavior. The report is focused on

[8] The Family Educational Rights and Privacy Act protects the privacy of personally identifiable information in a student's education records, including the disclosure or redisclosure of Personally Identifiable Information. The reader can download the guidebook at http://www.ccsso.org/Documents/TILSATestSecurityGuidebook.pdf.

"large-scale assessment programs" in general, but the recommendations might apply to ELP assessments.

We highlight what the authors consider a "limited number of critical recommendations" (Olson & Fremer, 2013, p. 6) in each of the three report sections: prevention, detection, and follow-up investigation. To prevent testing irregularities, the authors recommend the following: (1) announce that steps are being taken to ensure adherence to testing rules and include information about consequences of failing to follow state testing rules, (2) assign personnel to be responsible for test security and to monitor effectiveness of security efforts, (3) limit the testing window when possible, and (4) provide training on current security policies and procedures to new staff and to all staff periodically.

To detect testing irregularities, stakeholders should: (1) make sure all plans for employing data forensics are reviewed by legal and communications teams, (2) employ data forensics analysis—the use of analytic methods to identify or detect potential instances of cheating—on a regular basis for all high-stakes assessment programs, and (3) develop interpretive guidelines for use of data forensics and include the interpretation and use of data forensics in staff trainings.

Follow-up investigations of potential improper or unethical testing behavior should: (1) focus data forensics findings on a select number of statistically significant findings—the "worst of the worst" (Olson & Fremer, p. 8), (2) respect the confidentiality of all individuals involved in any data forensics investigations, and (3) maintain records of any forensics investigation suitable for sharing in a court of law.

Section 4. Using the ELP Assessment System for Accountability

The studies in Section 4 focus on establishing Title III accountability systems and reporting out student outcomes.[9] Most of the studies focus on uses of an ELP system for accountability as defined by NCLB, but the methods described in these studies can help inform accountability systems established under ESSA.

As noted in previous sections, Title III of NCLB required an annually administered ELP assessment aligned to the ELP standards that measured the four domains of listening, speaking, reading, and writing (sec. 3113(b)(2)). Title III accountability provisions required states to define criteria for progress (Annual Achievement Outcome 1) and attainment in learning English (Annual Achievement Outcome 2) and to establish targets for annually increasing the number or percentage of ELs meeting these criteria (sec. 3122).

Under ESSA, Title III requires that states implement standardized, statewide procedures for identifying ELs ("entrance procedures") and for determining when special language services are no longer needed ("reclassification procedures"). Title III also requires states to disaggregate English learners with a disability from English learners without disabilities. Title I requires annual ELP assessments.

[9] This section is tied to NCLB accountability provisions, but *Section 5* findings remain relevant to states as they set up their accountability systems measuring ELs' growth in language proficiency and attainment of the English proficiency performance standard.

Under ESSA, Title I requires that each state set up its own accountability system. The accountability system must incorporate at least four indicators, including English language proficiency, and one nonacademic indicator, such as student engagement, educator engagement, access to and completion of advanced coursework, or school climate. Two options are provided for including recently arrived ELs (sec. 1111(3)(A)):

1. Exclude recently arrived ELs from one administration of the reading or English language arts assessment and assess and incorporate these students' test results after they have been enrolled in a U.S. school for one year.
2. Assess recently arrived ELs, but exclude the results from the state accountability system in their first year in a U.S. school, include a measure of student growth in the second year, and include proficiency on the academic assessments beginning in the third year.

Although states will be responsible for establishing their own accountability systems, the systems must be submitted to the U.S. Department of Education. Plans will be peer reviewed; the reviewers' names will be made public, and states could have a hearing if their plan is turned down.

Under ESSA, former ELs may be included in the EL subgroup for up to four years after reclassification as fluent English proficient (sec. 1111(3)(B)), which represents two years longer than the period former ELs could be included in the EL group under NCLB.

Title III of NCLB required Title III districts to submit comprehensive evaluation reports to their state education agencies biennially. The reports described the types of instructional programs and activities supported with Title III funds and how well current ELs and ELs exited from EL status within the previous two years met state Annual Measurable Achievement Objectives (AMAOs).

Title III of ESSA specifies several LEA biennial reporting requirements, including the types of instructional programs and activities supported with Title III funds (sec. 3121(a)(1)); ELs making progress in English language proficiency in the aggregate and disaggregated by English learners with a disability (sec. 3121(a)(2)); ELs attaining English language proficiency (sec. 3121(a)(3)); ELs exiting EL status based on their attainment of English language proficiency (sec. 3121(a)(4)); ELs meeting academic standards for each of the four years after exit (sec. 3121(a)(5)); and ELs not attaining English language proficiency within five years of initial EL classification and first enrollment in a state's LEA (sec. 3121(a)(6)).

Defining Empirically Based AMAO Criteria and Targets

As noted in the previous section, Title III of NCLB required states to determine annual increases in the number or percentage of students who make progress in learning English (AMAO 1) and the number or percentage of students who attain English language proficiency (AMAO 2) as determined by the state ELP assessment. States varied in their methods used to establish AMAO targets—the percentages of ELs making progress in learning English and attaining English language proficiency—and target structures—annual increments to AMAO targets that LEA subgrantees and the state had to meet.

The studies in this section describe key decisions that had to be made in defining AMAO 1 and AMAO 2 criteria and targets, and provides a detailed example (California) (Linquanti & George, 2007). Although these studies are specific to NCLB, they may be useful in considering methods used to set growth and attainment expectations under ESSA.

Key decisions in measuring AMAO 1 included: (1) determining the metric used to measure growth in ELP; (2) determining an annual target for growth; (3) setting starting (2003–04) and ending points (2013–14) for the target—that is, the percentage of students within each LEA expected to meet their annual growth target at the beginning of the period of analysis (2003–04) compared with the percentage expected to meet these respective targets by the end of the period of analysis (2013–14); and (4) setting an annual rate of growth between the start and end points. Key decisions required in measuring AMAO 2 included: (1) determining the level of English proficiency at which ELs are considered proficient level; (2) determining the EL cohort to be included in the analysis; (3) setting the starting (2003–04) and ending points (2013–14) for the AMAO 2 targets; and (4) setting an annual rate of growth between the start and end points (as described in Linquanti & George, 2007).

With regard to establishing AMAO targets and target structures in California, LEAs (with more than 25 ELs with two consecutive years of ELP data) were ranked based on the proportion of students within each LEA who were meeting the progress or attainment criteria for each respective AMAO. The starting points for both AMAO 1 and AMAO 2 were determined to be the observed proportions of students meeting each criterion (progress, attainment) for districts at the 20th percentile of the distribution of LEAs on each criterion (i.e., 51 percent of ELs meeting their annual growth target within each LEA). The ending points were determined to be the proportion of students meeting each criterion at the 75th percentile—a target viewed as being "attainable yet rigorous" (Linquanti & George, 2007, p. 5). In other words, by 2014, all LEAs were expected to reach the point that the top 25 percent of LEAs (i.e., 75th percentile of the distribution) had attained in 2001–02. This translated into 64 percent of ELs making progress (AMAO 1) and 46 percent of the AMAO 2 cohort attaining the English proficient level (AMAO 2). Incremental growth targets (between the starting and ending points) were established to allow for slow initial growth, followed by steady growth in later years.

As detailed in the preceding example, expectations for growth and attainment should be "reasonable but challenging" (Cook et al., 2008, p. 6). Grounding progress and attainment performance criteria in second language acquisition research and real student data are best practices that can help achieve this goal. Examining patterns and rates of ELs making progress toward and attaining English proficiency on the ELP assessment, including variation by proficiency level and grade, can provide a basis for determining how progress might be defined as well as what reasonable or achievable targets might be established (Linquanti & George, 2007).

Research has suggested that it takes approximately three to five years for ELs to develop oral English language proficiency and four to seven years to develop English proficiency in all four domains of proficiency (reading, writing, listening, speaking). The actual time can vary widely based on the grade level at which ELs enter a U.S. school, their initial level of proficiency, and the quality of instruction (for a review see Linquanti & George, 2007; Working Group on ELL Policy, 2011). Research also has suggested that it is reasonable to expect more rapid growth in

language development for ELs with lower levels of English proficiency; growth slows down as ELs become more proficient (Cook et al., 2008; Gottlieb, 2013; Kenyon et al., 2011). For these reasons, Cook et al. (2008) recommended setting growth and attainment targets conditioned by grade and initial proficiency level.

Linquanti and Cook (2016) provided guidance on EL reclassification policies and practices. Reclassification has significance within state and local accountability systems more generally because it signals that the determination has been made that a student no longer requires legally mandated support services to meaningfully participate in an English instructional environment. Failure to be reclassified after a certain amount of time or at a certain point in a student's schooling can have consequences for these learners. ELs who remain in the subgroup for most of their schooling trajectory (i.e., long-term ELs) or during their secondary schooling may be unable to meaningfully access courses required for college and career readiness. Linquanti and Cook (2016) reviewed current reclassification criteria across the 50 states and Washington, D.C. They reported that 29 states and the District of Columbia relied on one criterion, the ELP assessment, to reclassify ELs, whereas 10 states relied on the ELP assessment and one additional criterion (e.g., content assessments, teacher input, grades, writing samples). The authors (Linquanti & Cook, 2016) provided guidelines to districts states and consortia for working toward establishment of common EL reclassification criteria and procedures. Tanenbaum et al. (2012) found that in 2009–10, 16 states took into account time in EL program or the amount of time that students had received EL services in setting progress and attainment expectations. How states incorporated "time in program" into their AMAOs differed substantially across states. For example, in 2009–10, Texas required 12 percent of the ELs enrolled in an EL instructional program for one to four years to attain proficiency and 20 percent of ELs who had been enrolled in an EL program for more than five years to attain proficiency. Five states required a specific percentage of ELs to attain proficiency after enrollment in an EL instructional program for a set number of years, and two states established a weighting system where ELs would count more or less towards reaching the AMAO target based on the time they had been enrolled in an EL program. States that did not incorporate time in EL program into their AMAOs cited the difficulty of making an accurate determination of the time that an EL had received services, particularly for ELs who transferred from other states or districts where they may have received EL services.

Establishing and Validating the Title III Accountability System

Establishing Title III Accountability Systems

Cook, Linquanti, Chinen, and Jung (2012) suggested that accountability systems be "ambitious, realistic, and meaningful" (p. xiv). As described in the previous section, they also argue that systems be set up using real data and that growth and attainment targets be conditioned by grade and initial proficiency level.

In terms of acceptance, Linquanti and George (2007) found that presenting empirical data to stakeholders (including policymakers and educators) and providing opportunities for them to provide input could help increase their acceptance of state accountability systems. For example, the acceptance of the Title III accountability system in California may be due to the fact that policymakers were presented with options for defining different progress expectations under

AMAO 1 and for attaining English language proficiency under AMAO 2. The options were based on empirical data. Educators provided input during the development of the system.

Validating ELP Accountability Systems

Few studies in this review provided models that might help validate the ELP accountability systems. While not directly validating entire state accountability systems, recent empirical work (e.g., Robinson-Cimpian & Thompson, 2015; Umansky & Reardon, 2014) might help validate policies related to setting targets for growth and attainment. Using 12 years of longitudinal data on ELs in one large district, Umansky and Reardon (2014) examined the time it took ELs in four different instructional environments (English immersion, transitional bilingual, maintenance bilingual, and dual immersion) to be reclassified as fluent English proficient. ELs enrolled in two-language programs took longer to exit EL programs but had better outcomes overall by the end of the study (higher overall reclassification, English proficiency, and content proficiency). Their study provides an example of how the time frames for ELs' outcomes may vary based on the instructional goal of the program (e.g., biliteracy compared with English-only outcomes). Robinson-Cimpian and Thompson (2015) examined the impact of a change in exit criteria in California in 2006–07 on ELs' content-area performance and graduation outcomes in the Los Angeles Unified School District. They found that the move to more stringent reclassification criteria (i.e., students had to meet higher criteria in order to be deemed fluent English proficiency) improved student outcomes as evidenced by higher English language arts achievement and improved graduation rates. Examining the time it takes ELs to meet exit criteria (Umansky & Reardon, 2014) and the impact of exit criteria on student outcomes (Robinson-Cimpian & Thompson, 2015) is particularly relevant given the new ESSA provision requiring that states have standardized exit procedures (sec. 3113(b)(2)). Furthermore, changes to district exit criteria in response to this new provision might create opportunities to empirically evaluate the impact of different exit criteria on student outcomes.

Reporting

The *Reporting* subsection focuses on reporting formats for ELP assessment results recommended for different audiences and purposes. It first cites some general best practices for reporting state assessment data (Thurlow et al., 2013). Although the practices focused on reporting content assessment data for ELs with disabilities, the guidelines could be extended to ELP assessment results for ELs with disabilities or other subgroups of ELs or generalized to the broader EL population. This section also includes a specific state example of reporting ELP assessment results.

Some best practices in reporting assessment data on ELs with disabilities highlighted in Thurlow et al. (2013) include (1) using disaggregated data to account for diversity in demographic characteristics and proficiency level (when the number of students in each reporting group is sufficient); (2) providing interpretive guides to educators that include information on how to use assessment results for multiple uses, including program evaluation, group performance analysis, and summative analysis; and (3) creating multiple versions of score reports for parents and students (which may also include native-language reports and in-person meetings) to ensure that families are well informed. Thurlow et al. (2013) also suggested providing cross-state reporting when applicable for states that share common assessments.

The California Department of Education established an online reporting system[10] that allows any individual to view state English language proficiency assessment (in California, the California English Language Development Test [CELDT]) at multiple levels (e.g., state, county, district, school) by selecting from a drop-down menu. Furthermore, the viewer can review assessment results by student subgroups (e.g., gender, students with disabilities, primary language, and program participation). A particularly useful feature of the reporting system is the "stoplight" system, which color codes cells depending on whether groups of students increased their ELP performance level from the previous year (green), experienced no change in performance level (yellow), or decreased their performance level (red). The reporting function generates tables that display the percentage of students by performance level on the current CELDT compared with those students' most recent previous CELDT performance level.

Areas for Further Research

To support the development and implementation of valid and reliable ELP assessment systems, future research might examine the following variables:

- The academic language needed for content-area success. According to the California ELP standards document, academic English encompasses discourse practices, text structures, grammatical structures, and vocabulary. Some characteristics of academic English span disciplines, while others are limited to particular disciplines (see California Department of Education, 2012, Appendix B). CCSSO (2012) provided guidance to "craft the next generation of ELP standards" by providing examples of discipline-specific language practices that all ELs must acquire to master the CCSS and NGSS and a framework for unpacking the language demands of the CCSS and NGSS (Section 1).
- The appropriateness of using one ELP assessment for multiple purposes (e.g., instructional program placement, initial EL identification, progress, attaining English proficiency, and exit; Sections 2 and 4).
- The classification consistency estimates for particular ELP assessments, where classification consistency estimates are defined as measures of how consistently and accurately the ELP assessment would classify students into the same performance categories if they were (in theory) administered the assessment on multiple occasions (Section 4).
- Methods or procedures for ensuring the validity and fairness of the ELP assessment for EL subpopulations such as students with interrupted formal education, migrant students, and students from different language groups (Section 2).
- The effectiveness of specific accommodations on ELP assessments for ELs with specific disabilities (Section 4).
- Models to help validate a state's accountability system for ELs (Section 4).
- How to create assessment and accountability systems that take into account ELs who are acquiring proficiency in more than one language (Section 4).

[10] The reader can view this reporting system at this website: http://dq.cde.ca.gov/dataquest/.

References

Abedi, J. (2008). Measuring students' level of English proficiency: Educational significance and assessment requirements. *Educational Assessment,* 13, 193–214.

Abedi, J. (2014). The use of computer technology in designing appropriate test accommodations for English language learners. *Applied Measurement in Education, 27*(4), 261–272.

Abedi, J., & Ewers, N. (2013). *Accommodations for English language learners and students with disabilities: A research-based decision algorithm*. Olympia, WA: Smarter Balanced Assessment Consortium. Retrieved from https://www.smarterbalanced.org/wp-content/uploads/2016/08/Accommodations-for-under-represented-students.pdf

Albus, D., & Thurlow, M. L. (2008). Accommodating students with disabilities on state English language proficiency assessments. *Assessment for Effective Intervention, 33*, 156–166.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council of Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Bailey, A., Butler, F., & Sato, E. (2007). Standards-to-standards linkage under Title III: Exploring common language demand in ELD and science standards. *Applied Measurement in Education, 20*(1), 53–78.

California Department of Education. (2012). *English language development standards.* Sacramento, CA: Author. Retrieved from http://www.cde.ca.gov/sp/el/er/eldstandards.asp

California Department of Education. (2015). *California English Language Development Test. 2015–16 CELDT information guide*. Sacramento, CA: Author. Retrieved from http://www.cde.ca.gov/ta/tg/el/

Camilli, G., Cizek, G. J., & Lugg, C. A. (2001). Psychometric theory and the validation of performance standards: History and future perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 445–475). Mahwah, NJ: Lawrence Erlbaum.

Christensen, L. L., Albus, D. A., Kincaid, A., Christian, E., Liu, K. K., & Thurlow, M. L. (2014). Including students who are blind or visually impaired in English language proficiency assessments: A review of state policies. Minneapolis, MN: University of Minnesota, Improving the Validity of Assessment Results for English Language Learners with Disabilities (IVARED).

Christensen, L. L., Albus, D. A., Liu, K. K., Thurlow, M. L., & Kincaid, A. (2013). *Accommodations for students with disabilities on state English language proficiency assessments: A review of 2011 state policies*. Minneapolis, MN: University of Minnesota, Improving the Validity of Assessment Results for English Language Learners with Disabilities (IVARED).

Cook, H. G. (2006). Aligning English language proficiency tests to English language learning standards. In Council of Chief State School Officers, *Aligning assessment to guide the learning of all students: Six reports* (pp. 131–154). Washington, DC: Council of Chief State School Officers.

Cook, H. G. (2009). Annual Measurable Achievement Objective # 2 (AMAO 2): A Report on a Meeting Comparing the Kentucky Core Content Test (KCCT) and ACCESS for ELLs™ to Establish an AMAO 2 Exit Criterion. Kentucky Department of Education.

Cook, H. G. (2013). *Examining how to establish English proficiency using alternate composite scores in a WIDA state: A logistic regression approach*. Madison, WI: University of Wisconsin–Madison, Wisconsin Center for Education Research. Retrieved from http://wida.us/get.aspx?id=634

Cook, H. G. (2014). *WIDA research report: Examining relationships between Alternate ACCESS and state alternate assessments: Exploring notions of English language proficiency*. Madison, WI: WIDA Consortium. Retrieved from https://www.wida.us/get.aspx?id=750

Cook, H. G., Boals, T., Wilmes, C., & Santos, M. (2008). *Issues in the development of annual measurable achievement objectives for WIDA consortium states* (WCER Working Paper No. 2008-2). Madison, WI: University of Wisconsin–Madison, Wisconsin Center for Education Research. Retrieved from http://wcer-web.ad.education.wisc.edu/docs/working-papers/Working_Paper_No_2008_02.pdf

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

Council of Chief State School Officers (CCSSO). (2012). *Framework for English language proficiency development standards corresponding to the Common Core State Standards and the Next Generation Science Standards*. Washington, DC: Author. Retrieved from http://www.ccsso.org/Documents/2012/ELPD%20Framework%20Booklet-Final%20for%20web.pdf

Council of Chief State School Officers (CCSSO). (2016). *Major provisions of Every Student Succeeds Act (ESSA) related to the education of English learners*. Retrieved from http://www.ccsso.org/Documents/2016/ESSA/CCSSO%20Resource%20on%20ELs%20and%20ESSA.pdf

Council of Chief State School Officers (CCSSO) & Wisconsin Center for Education Research (WCER). (2010). *Surveys of enacted curriculum: Content & instruction in focus.* Washington, DC: Author. Retrieved from http://www.ccsso.org/Documents/2010/SEC_Content_Instruction_2010.pdf

Cumming, A. (2013). Validation of language assessments. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 6006–6015). Hoboken, NJ: Wiley-Blackwell.

Department of Justice & Office of Civil Rights (DOJ & OCR). (2015). *Dear colleague letter.* Washington, DC: Authors.

Douglas, K. M., & Mislevy, R. J. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics, 35*, 280–306.

ELPA21. (2015). *ELPA21 accessibility and accommodations manual, school year 2015–2016*. Retrieved from https://ne-elpa.nextera.questarai.com/Help/Accessibility and Accommodations Manual SY 2015-2016_NE.pdf

Every Student Succeeds Act of 2015 (ESSA), Pub. L. No. 114–95 (2015).

Family Educational Rights and Privacy Act, 20 U.S.C. § 1232g; 34 CFR Part 99 (1974).

Francis, D. J., Tolar, T. D., & Stuebing, K. K. (2012). *The language-achievement connection for English language learners.* Paper presented at the Modern Modeling Methods Conference, University of Connecticut, Storrs, CT.

Gottlieb, M. (2013). *Essential actions: A handbook for implementing WIDA's framework for English language development standards*. Madison, WI: University of Wisconsin–Madison, Wisconsin Center for Education Research.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 47–76). New York, NY: Routledge.

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. §612 (2004).

Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Kenyon, D. M., MacGregor, D., Li, D., & Cook, H. G. (2011). Issues in vertical scaling of a K–12 English language proficiency test. *Language Testing, 28*(3), 383–400.

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

Linquanti, R., & Cook, H. G. (2016). Re-examining reclassification: Guidance from a national working session on policies and practices for exiting students from English learner status. In R. Linquanti, H. G. Cook, A. L. Bailey, & R. MacDonald, *Moving toward a more common definition of English learner: Collected guidance for states and multi-state assessment consortia.* Washington DC: Council of Chief State School Officers. Retrieved from http://www.ccsso.org/Documents/Moving%20Toward%20a%20More%20Common%20Definition%20of%20English%20Learner-Final(0).pdf

Linquanti, R., & George, C. (2007). Establishing and utilizing an NCLB Title III accountability system: California's approach and findings to date. In J. Abedi (Ed.), *English language proficiency assessment in the nation: Current status and future practice* (pp. 105–118). Davis, CA: University of California Press.

Liu, K. K., Goldstone, L. S., Thurlow, M. L., Ward, J. M., Hatten, J., & Christensen, L. L. (2013). *Voices from the field: Making state assessment decisions for English language learners with disabilities.* Minneapolis, MN: University of Minnesota, Improving the Validity of Assessment Results for English Language Learners with Disabilities (IVARED).

Liu, K. K., Ward, J. M., Thurlow, M. L., & Christensen, L. L. (2015). Large-scale assessment and English language learners with disabilities. *Educational Policy,* 1–33. doi: 10.1177/0895904815613443

National Center on Educational Outcomes (NCEO). (2014, July). *Exploring alternate ELP assessments for ELLs with significant cognitive disabilities* (NCEO Brief #10). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

National Research Council Committee on the Assessment of 21st Century Skills. (2011). *Assessing 21st century skills: Summary of a workshop.* Washington, DC: National Academies Press.

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

Olson, J. F., & Fremer, J. (2013). *TILSA test security guidebook.* Washington, DC: Council of Chief State School Officers. Retrieved from http://www.ccsso.org/Documents/TILSATestSecurityGuidebook.pdf

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press. Retrieved from http://www.nap.edu/books/0309072727/html

Popham, W. J., Keller, T., Moulding, B., Pellegrino, J., & Sandifer, P. (2005). Instructionally supportive accountability tests in science: A viable assessment option? *Measurement: Interdisciplinary Research and Perspectives, 3*(3), 121–179.

Powers, S., Williams, N., Keng, L., & Starr, L. (2014). *An example of evidence-based standard setting for an English language proficiency test*. Philadelphia, PA: NCMA. Retrieved from http://researchnetwork.pearson.com/wp-content/uploads/An-Example-of-EBSS-for-an-ELP-Assessment_v2.pdf

Robinson-Cimpian, J. P., & Thompson, K. D. (2015). The effects of changing test-based policies for reclassifying English learners. *Journal of Policy Analysis and Management*. doi: 10.1002/pam.21882

Tanenbaum, C., Boyle, A., Soga, K., Le Floch, K. C., Golden, L., Petroccia, M., Toplitz, M., Taylor, J., & O'Day, J. (2012). *National evaluation of Title III implementation: Report on state and local implementation.* Washington, DC: Office of Planning, Evaluation and Policy Development, U.S. Department of Education. Retrieved from http://files.eric.ed.gov/fulltext/ED531982.pdf

Thurlow, M. L., Liu, K. K., Ward, J. M., & Christensen, L. L. (2013). *Assessment principles and guidelines for ELLs with disabilities.* Minneapolis, MN: University of Minnesota, Improving the Validity of Assessment Results for English Language Learners with Disabilities (IVARED).

U.S. Department of Education. (2014). *Questions and answers regarding inclusion of English learners with disabilities in English language proficiency assessments and title III annual measurable achievement objectives*. Washington, DC: Author. Retrieved from http://www2.ed.gov/policy/speced/guid/idea/memosdcltrs/q-and-a-on-elp-swd.pdf

U.S. Department of Education. (2015). *Addendum to questions and answers regarding inclusion of English learners with disabilities in English language proficiency assessments and title III annual measurable achievement objectives*. Washington, DC: Author.

U.S. Department of Justice & U.S. Department of Education. (2014). *Frequently asked questions on effective communication for students with hearing, vision, or speech disabilities in public elementary and secondary schools*. Retrieved from http://www2.ed.gov/about/offices/list/ocr/docs/dcl-faqs-effective-communication-201411.pdf

Umansky, I. M., & Reardon, S. F. (2014). Reclassification patterns among Latino English learner students in bilingual, dual immersion, and English immersion classrooms. *American Educational Research Journal, 51,* 879–912.

WIDA. (n.d.). *Alternate ACCESS for ELLs.* Retrieved from https://www.wida.us/assessment/alternateaccess.aspx

Working Group on ELL Policy. (2011). *Improving educational outcomes for English language learners: Recommendations for the reauthorization of the Elementary and Secondary Education Act.* Questions and answers. Washington, DC: Author. Retrieved from http://ellpolicy.org/wp-content/uploads/QA.pdf

Additional Resources

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

Betebenner, D. W. (2009). *Growth, standards, and accountability*. Dover, NH: National Center for the Improvement of Educational Assessment.

California Department of Education. (2011). *California English Language Development Test: A comparison study of kindergarten and grade 1 English-fluent students and English learners on the 2010–11 edition of the CELDT.* Sacramento, CA: Author. Retrieved from http://www.cde.ca.gov/be/pn/im/documents/imadaddec11item02a1.pdf

California Department of Education. (2013). *Monitoring instrument item report: 2013 – 14 program instruments. English Learner On-site 2013–14.* Sacramento, CA: Author. Retrieved from http://www.cde.ca.gov/ta/cr/elmonthome.asp

California Department of Education. (2014). *2013–14 Title III accountability report: Information guide.* Sacramento, CA: Author. Retrieved from http://www.cde.ca.gov/ta/ac/t3/documents/infoguide13-14.pdf

Christensen, L. L., Lail, K. E., & Thurlow, M. L. (2007). *Hints and tips for addressing accommodations issues for peer review.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Cook, H. G., & MacDonald, R. (2014). *Reference performance level descriptors: Outcome of a national working session on defining an "English proficient" performance standard.* Working paper series. Washington, DC: Council of Chief State School Officers.

Cook, H. G., & Wilmes, C. (2007). *Alignment between the Kentucky Core Content for Assessment and the WIDA Consortium English Language Proficiency Standards*. Madison, WI: Wisconsin Center for Education Research.

Flowers, C., Wakeman, S., Browder, D., & Karvonen, M. (2007). *Links for academic learning: An alignment protocol for alternate assessments based on alternate achievement standards*. Charlotte, NC: University of North Carolina at Charlotte.

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

Kane, M. (2006). Validation. In R. L. Brennan (Ed.). *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*, 221–240.

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners on large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79,* 1168–1201.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Llosa, L. (2011). Standards-based classroom assessments of English proficiency: A review of issues, current developments, and future directions for research. *Language Testing, 28*, 367–382.

Thissen, D. (2007). Linking assessments based on aggregate reporting: Background and issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 287–312). New York, NY: Springer.

U.S. Department of Education. (2008). Title III of the Elementary and Secondary Education Act of 1965 (ESEA), as amended by the No Child Left Behind Act of 2001. (NCLB). Notice of final interpretations. 73 Fed. Reg. 61828, 61837-61838 (October 17, 2008).

Appendix A. Panelists, Staff, and Partners

Panelists

Content-Area Assessments

Content Area	Panelists
Psychometricians	Brian Gong (Center for Assessment) Phoebe Winter (Consultant in Assessment Research and Development)
State Policy	Ed Roeber (Michigan State University)
State Representatives on Assessment	Lily Roberts (California Department of Education) Margaret Ho (Council of Chief State School Officers, Program Director for English Language Proficiency Assessment 21 Sustainability)

Note. Ed Roeber contributed to the Title III literature review.

English Language Proficiency Assessments

Content Area	Panelists
Psychometricians	Lyle Bachman (University of California at Los Angeles) David Francis (University of Houston)
English Language Development	Donna Christian (Center for Applied Linguistics)
Standards Development	Pam Spycher (WestEd)
Special Education	Martha Thurlow (National Center on Educational Outcomes)
Title III Directors	Angelica Infante-Green (New York State Department of Education) David Nieto (Illinois State Board of Education)
English as a Second Language Teacher/English Language Proficiency Assessment	Frances Melecio (Orange County Public Schools, Florida)

Note. Martha Thurlow contributed to the Title III literature review.

Project Staff and Partners

Organization	Project Staff
American Institutes for Research	Diane August, Rachel Slama
WestEd	Robert Linquanti
Wisconsin Center for Education Research	Gary Cook
University of California, Davis	Jamal Abedi
University of California, Los Angeles	Li Cai, Mark Hansen

Note. Diane August, Rachel Slama, Robert Linquanti, Gary Cook, and Jamal Abedi contributed to the Title III literature review.

Appendix B. Characteristics of Reviewed Studies

Citation	Publication Type[a]	Study Type[b]	Unit of Analysis[c]	Section or Sections Addressed by This Resource
Abedi, 2008	Peer-reviewed journal article	Review of ELP standards, assessment, PLD, assessment and accountability system	Test	3
Abedi & Ewers, 2013	Technical report	Review of test accommodation	Accommodation	3
Albus & Thurlow, 2008	Peer-reviewed journal article	Review of test accommodation	State	2, 3
American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 2014	Other	Other	Other	2, 3
Bailey & Huang, 2011	Peer-reviewed journal article	Review of ELP standards, assessment, PLD, assessment and accountability system	ELP standards	[d]
Bailey et al., 2007	Technical report	Alignment study	Content standards	1
Betebenner, 2009	Center report	Policy and/or practice report (e.g., CCSSO)	Federal	[d]
California Department of Education, 2011	State report	Empirical analysis of state and district data	Student	[d]
California Department of Education, 2012	State report	Policy and/or practice report (e.g., CCSSO)	ELP standards	1, 4
California Department of Education, 2013	State report	Policy and/or practice report (e.g., CCSSO)	Other	[d]
California Department of Education, 2014	State report	Policy and/or practice report (e.g., CCSSO)	District	[d]
Camilli et al., 2001	Chapter in an edited volume	Other	Content standards	2

Citation	Publication Type[a]	Study Type[b]	Unit of Analysis[c]	Section or Sections Addressed by This Resource
CCSSO, 2012	Policy paper	Policy and/or practice report (e.g., CCSSO)	ELP standards	1
CCSSO & Wisconsin Center for Education Research, 2010	Other	Other	Teacher	1
Christensen et al., 2013	Center report	Policy and/or practice report (e.g., CCSSO)	Accommodation	3
Christensen, Lail, & Thurlow, 2007	Center report	Policy and/or practice report (e.g., CCSSO)	Accommodation	[d]
Cook & MacDonald, 2014	Technical report	Policy and/or practice report (e.g., CCSSO)	ELP standards	[d]
Cook, 2013	Policy paper	Policy and/or practice report (e.g., CCSSO)	Test	3
Cook, 2006	Policy paper	Policy and/or practice report (e.g., CCSSO)	Test	2
Cook, Boals, Wilmes, & Santos, 2008	Center report	Policy and/or practice report (e.g., CCSSO)	Other	[d]
Cook & Wilmes, 2007	Center report	Alignment study	ELP standards	[d]
Cumming, 2013	Chapter in an edited volume	Review of ELP standards, assessment, PLD, assessment and accountability system	Test	[d]
Douglas & Mislevy, 2010	Peer-reviewed journal article	Review of ELP standards, assessment, PLD, assessment and accountability system	Other	3
Flowers, Wakeman, Browder, & Karvonen, 2007	Technical report	Alignment study	Test	[d]
Francis & Rivera, 2007	Chapter in an edited volume	Review of ELP standards, assessment, PLD, assessment and accountability system	Test	[d]

Citation	Publication Type[a]	Study Type[b]	Unit of Analysis[c]	Section or Sections Addressed by This Resource
Francis, Tolar, & Stuebing, 2012	Peer-reviewed paper or conference presentation	Review of ELP standards, assessment, PLD, assessment and accountability system	Test	3
Gottlieb, 2013	Center report	Policy and/or practice report (e.g., CCSSO)	ELP standards	2, 3, 4
Hambleton, 2001	Chapter in an edited volume	Review of ELP standards, assessment, PLD, assessment and accountability system	Performance level descriptors	1, 2
Hambleton, Pitoniak, & Copella, 2012	Chapter in an edited volume	Review of ELP standards, assessment, PLD, assessment and accountability system	Performance level descriptors	[d]
Kane, 2001	Chapter in an edited volume	Review of ELP standards, assessment, PLD, assessment and accountability system	Performance level descriptors	2
Kane, 2006	Chapter in an edited volume	Review of ELP standards, assessment, PLD, assessment and accountability system	Other	[d]
Kane & Case, 2004	Peer-reviewed journal article	Review of ELP standards, assessment, PLD, assessment and accountability system	Other	[d]
Kenyon, MacGregor, Li, & Cook, 2011	Peer-reviewed journal article	Review of ELP standards, assessment, PLD, assessment and accountability system	Test	1, 2, 3, 4
Kieffer, Lesaux, Rivera, & Francis, 2009	Peer-reviewed journal article	Meta-analysis	Accommodation	[d]
Koretz & Hamilton, 2006	Chapter in an edited volume	Review of ELP standards, assessment, PLD, assessment and accountability system	Test	[d]
Linquanti & Cook, 2013	Technical report	Policy and/or practice report (e.g., CCSSO)	Student	3

Citation	Publication Type[a]	Study Type[b]	Unit of Analysis[c]	Section or Sections Addressed by This Resource
Linquanti & George, 2007	Chapter in an edited volume	Review of ELP standards, assessment, PLD, assessment and accountability system	State	4
Liu et al., 2013	Technical report	Policy and/or practice report (e.g., CCSSO)	Teacher	3
Liu, Ward, Thurlow, & Christensen, 2015	Peer-reviewed journal article	Other	Other	3
Llosa, 2011	Peer-reviewed journal article	Review of ELP standards, assessment, PLD, assessment and accountability system	Test	[d]
National Center on Educational Outcomes, 2014	Center report	Policy and/or practice report (e.g., CCSSO)	Other	3
Olson & Fremer, 2013	Technical report	Policy and/or practice report (e.g., CCSSO)	Test	3
Pellegrino, Chudowsky, & Glaser, 2001	Policy paper	Policy and/or practice report (e.g., CCSSO)	Test	3
Popham, Keller, Moulding, Pellegrino, & Sandifer, 2005	Peer-reviewed journal article	Review of ELP standards, assessment, PLD, assessment and accountability system	Test	3
Tanenbaum et al., 2012	Policy paper	Policy and/or practice report (e.g., CCSSO)	Federal	1, 2, 4
Thissen, 2007	Chapter in an edited volume	Review of ELP standards, assessment, PLD, assessment and accountability system	Other	[d]
Thurlow, Liu, Ward, & Christensen, 2013	Technical report	Policy and/or practice report (e.g., CCSSO)	Test	3, 4
Working Group on ELL Policy, 2011	Policy paper	Policy and/or practice report (e.g., CCSSO)	Federal	2, 4

Note. CCSSO is Council of Chief State School Officers. ELP is English language proficiency. PLD is proficiency level descriptors. .

[a] Publication type codes include: center report, chapter in an edited volume, dissertation, local education agency report, peer-reviewed journal article, peer-reviewed paper or conference presentation, policy paper, state report, technical report, other.

[b] Study type codes include: alignment study, case study, empirical analysis of state and district data, meta-analysis, policy and/or practice report (e.g., CCSSO), quasi-experimental design, randomized controlled trial, review of ELP standards/assessment/PLD/assessment and accountability system, review of test accommodation, other.

[c] Unit of analysis codes include: student, teacher, school, district, state, federal, test, test item, content standards, English language proficiency standards, performance-level descriptors, accommodation, other.

[d] These citations provided background to help frame the report but were not directly cited in the report.

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