Language Testing in Asia

REVIEW Open Access

Interconnection between constructs and consequences: a key validity consideration in K–12 English language proficiency assessments

Mikyung Kim Wolf*

*Correspondence:
mkwolf@ets.org

Center for Language Education and Assessment Research, Educational Testing Service (ETS), 660 Rosedale Rd, Princeton, NJ 08541, USA

Abstract

In the US K–12 public school settings, each state's annual English language proficiency (ELP) assessment, as part of an educational accountability system, has tremendous impact on English learner (EL) students' academic paths and success. It is thus critical to establish a robust validation framework and empirical evidence to ensure that states' ELP assessments are both appropriately interpreted and justifiably used for their intended purposes. The present article, as a perspective piece, highlights two key interrelated areas of validity concern: the construct and the consequences of ELP assessments. This article describes how the ELP construct has been redefined and operationalized in recent K–12 ELP standards and assessments in the US K–12 education settings, reflecting the current US educational reform and policy impact. Then, the article presents the ramifications these new ELP assessments have for making high-stake decisions about EL students and how construct validity issues are closely tied to consequential validity. A set of pivotal research areas pertaining to construct and consequential validity is proposed with implications for practice and policies to support EL students' needs. Implications from interconnection between the construct and the consequences of ELP assessments and research areas suggested in this article can be applicable in other countries where language assessments are part of educational reform or accountability systems.

Keywords: Accountability, Consequential validity, Construct validity, English learners, K–12 English language proficiency assessments

Introduction

Many countries perform education reforms in order to improve their educational systems, thereby equipping all their students with appropriate knowledge and skills to realize their potential in societies (Sahlberg, 2006). In the context of the US K-12 public education, accountability has been a centerpiece of education reform with the intent of holding educational agencies (states' department of education, districts, schools) accountable for all students' equitable learning and reducing achievement gap (Linn, 2000). In particular, academic standards and standardized assessments have been playing major parts in accountability and education reform for the past five decades in the USA. Standard-based accountability and test-based accountability are the frequent terminologies used to encapsulate the US K-12 education reform policies (e.g., Deville & Chalhoub-Deville, 2011; Hamilton et al., 2012; Lane, 2020; O'Day & Smith, 2019). The changes of standards and uses of standardized assessments have made direct impacts on what is taught and how it is taught in the US K-12 public education.

This article deals with standardized English language proficiency (ELP) assessments that play increasingly crucial roles in accountability in the US K-12 public education. ELP assessments serve multiple purposes such as (a) identifying English learner (EL) students[1] who are in need of language supports and services, (b) measuring students' ELP levels to determine appropriate instructional types, (c) informing EL exit decisions (i.e., determining students' proficiency to move out of EL status), and (d) reporting the number of EL students progressing toward ELP attainment and accountability that evaluates school performance and appropriate interventions for schools. Considering these substantial stakes implicated for individual students, educators, and schools, validity for ELP assessment uses is an important topic.

Amongst various validity issues, this article focuses on the construct of K–12 ELP assessments and their relation to potential consequences as a key issue in strengthening a validity argument as well as justifications of accountability testing in the context of the US K-12 public education. Messick (1989, 1996) explicitly draws attention to consequences as part of his comprehensive concept of construct validity. He argues that construct representation and construct-irrelevant sources of variance are the traceable aspects of assessments to which positive or negative consequences can be attributed (Messick, 1996). For example, clearly defined and well-represented constructs in an assessment can influence a teacher's decision about what to teach. In a similar vein, Bachman and Palmer (2010), in their Assessment Use Argument validity framework, put a greater emphasis on the consequences of assessment use while making explicit links among the construct, assessment development, use, and consequences. Extending these ideas, Chalhoub-Deville (2016, 2020) calls for special attention to the societal dimension of consequences when evaluating validity for accountability testing that impacts educational reform and policies.

Aligned with these views, this article contends the significance of the interrelation between the construct and consequences when making validity arguments for accountability testing. In doing so, this article details how the ELP construct has been redefined and operationalized in recent K–12 ELP standards and assessments in the US K–12 education settings. Then, it discusses what ramifications these new ELP assessments have for making high-stake decisions about EL students and how construct validity issues are closely tied to consequential validity. A set of pivotal research areas pertaining to construct and consequential validity is proposed with implications for practice and policies to support EL students' needs. Although this article focuses on the USA contexts, the

[1] An "English learners (EL)" is the term used in the US federal and state government documents to refer to a student who is in need of linguistic support to meaningfully participate in K-12 school settings due to their developing English language proficiency. EL is considered an educational classification term. Thus, I used EL *students* to refer to individual students instead of ELs.

validity issues discussed here would be applicable in any country where language assessments are used for accountability as well as for high-stake decisions.

Contexts: English learners and accountability policy on English language proficiency in U.S. K-12 public education

According to a recent US census report for the 2017–2018 school year, over 5 million students are officially classified as *English learners* (ELs), constituting approximately 10% of the total enrollment in K–12 public schools in the USA (U.S. Department of Education, Office of English Language Acquisition, 2021). With the growing number of children whose home language is not English, schools are mandated to identify EL students who are in need of linguistic support due to their developing English language proficiency. Once formally designated as ELs, by federal law, these students are entitled to receive appropriate services and instructional support including bilingual or ESL programs and language-related accommodations during instruction and assessment.

EL students' achievement gap has been substantial, raising a concern regarding equity in the US K–12 public education. For example, in the 2019 National Assessment of Education Progress (NAEP) Grade 4 mathematics assessment, only 16% of EL students performed at or above their expected level of proficiency compared to 44% of non-EL students who met their proficiency level (US Department of Education, n.d.). This gap typically becomes larger in higher grade levels, as more challenging academic content and increased language demands are introduced in the upper grades. Researchers have increasingly used the term "opportunity gap" rather than "achievement gap" to describe this persistent disparity, as it has mainly resulted from inequitable opportunities that EL students experience (Callahan & Shifrer, 2016; Umansky, 2016). It is evident that helping EL students develop appropriate English language skills needed in school settings is crucial to address this opportunity gap.

The ELP assessment of EL students has now become a significant component in educational accountability for U.S. K–12 public schools. That is, states and schools are held accountable for EL students' ELP attainment. States are mandated to annually assess EL students' ELP and report the progress of these students' ELP attainment. This federal-level policy requirement has substantially influenced the assessment of EL students, spawning large-scale, standard-based ELP assessments. These ELP assessments involve high-stake uses both for individual students and school programs. The results of ELP assessments are used to indicate the types of services individual students need as well as the time students should exit out of EL status (National Research Council, 2011a; Wolf & Farnsworth, 2014). While states apply various criteria to make EL-status exit decisions, ELP assessments are used as an essential criterion in all states (Linquanti & Cook, 2015). ELP assessment results are also used to evaluate the quality of programs and to determine resource/funding allocations (Tanenbaum et al., 2012). Hence, in order to ensure that states' ELP assessments are both appropriately interpreted and justifiably used for their intended purposes, establishing a robust validity argument backed by empirical evidence is crucial.

As key contextual information for the evolution of ELP assessments, it is important to understand both prior and current accountability policies regarding the assessment of English language proficiency. The federal educational law governing K–12 education policies and practice in the USA, the Elementary and Secondary Education Act (ESEA), has been continuously reauthorized since first passed in 1965. The reauthorization of ESEA in 2001, known as No Child Left Behind (NCLB), greatly influenced the assessment of EL students. Under the Title I section of the law, states were required to include EL students in statewide assessments and report the testing results by subgroups. Under the Title III section of the law, NCLB stipulated that states must develop or adopt ELP standards and annually administer ELP assessments based on these standards. Prior to NCLB, EL students were often excluded from statewide assessments and were not monitored for their ELP progress in a standardized manner (Abedi, 2008). The policy within NCLB created the first generation of standard-based K–12 ELP assessments. However, there was little guidance available when developing ELP standards and assessments, leading to a considerable degree of variability of the content of ELP standards and thus the construct of ELP assessments (Bailey & Huang, 2011). This point is further described in the next section.

The latest reauthorization of ESEA took place in 2015 with the name of the Every Student Succeeds Act (ESSA), replacing NCLB. ESSA continued to focus on standard-based accountability. Regarding ELP assessments, it says:

> *"(i) IN GENERAL. — Each State plan shall demonstrate that local educational agencies in the state will provide for an annual assessment of English proficiency of all English learners in the schools served by the state educational agency.*
> *(ii) ALIGNMENT.—The assessments described in clause (i) shall be aligned with the state's English language proficiency standards described in paragraph (1)(F)." (ESSA, 2015, Section 1111(b)(2)(G), pp. 1830–1831)*

This stipulation underscores the assessment-based accountability for states, districts (i.e., local educational agencies), and schools to ensure the appropriate monitoring of EL students' ELP development. It also reinforces the importance of the alignment between assessments and standards. Notably, ESSA specifies that states' ELP standards must be aligned with states' academic content-area standards such as English language arts, mathematics, and science standards (ESSA, 2015, Section 1111(b)(2)(F)). This statement resulted partly from a body of research on academic language conducted during the NCLB period. For instance, a number of researchers asserted that the explicit instruction of academic language is critical to address EL students' needs (Bailey, 2007; Butler et al., 2004; Pereira & de Oliveira, 2015; Schleppegrell, 2012). Butler et al. (2004)'s study provided clear evidence of the mismatch of language skills measured in previous ELP assessments and those needed for students to engage in various disciplinary areas (e.g., tasks from mathematics and science textbooks). Based on the findings, the researchers called for the reconceptualization of the ELP assessment construct.

In the context of standard-based education reform, another significant impetus for the alignment between academic content and ELP standards is attributed to the advent of the Common Core State Standards. The tenet of standard-based education is to ensure the quality and equity of education for all students by setting the same expectations for all students (Hamilton et al., 2012). With a clearly documented set of expected knowledge and skills for students to achieve (i.e., standards), educational agencies (states, districts, and schools), and teachers can have a common goal and be clear about what they are accountable for. Yet, the content and variability of the academic standards across states were criticized not only due to the quality of the standards but also due to the low performance of American students on international assessments (National Research Council, 2011b). The initiative to have a core set of more rigorous and challenging academic standards for students to be ready for college and the workplace led to the development of the Common Core State Standards in (a) English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects and (b) Mathematics (National Governors Association Center for Best Practices [NGA], Council of Chief State School Officers [CCSSO], 2010). Similarly, the next-generation science standards were produced to serve as a new set of standards in science with more rigorous content, aligned with the Common Core State Standards (NGSS Lead States, 2013). Currently, almost all states have adopted the content of the common core in their state academic standards since 2014. Subsequently, there has been a need to develop new ELP standards or modify existing ELP standards to reflect the language demands manifested in the new academic content standards. This evolving change of ELP standards and the reconceptualization of the ELP construct is discussed in the next section.

Historically, ELP regulations were stipulated in different sections of the educational law (i.e., ESEA). However, ESSA placed regulations on ELP assessments under Title I of the law along with regulations on content-area assessments (e.g., ELA, mathematics), moving them from Title III. Since the federal government requires states to submit technical and validity evidence for their accountability assessment systems under Title I, this movement raised state stakeholders' attention to the quality of their ELP assessments (Hakuta & Pompa, 2017). This federal process to monitor the quality and validity of state assessment systems is called *peer review* (US Department of Education, 2018). The federal peer review guidance document employs a validity framework adopted from the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association,, & National Council on Measurement in Education [AERA, APA,, & NCME], 2014). This framework describes types of validity evidence based on test content, response processes, internal structure, and relations to other variables. In particular, states must submit evidence of alignment between assessment and standards as well as between ELP standards and content-area standards, as part of the validity evidence based on test content. This context indicates heightened attention to the appropriate ELP assessment development and valid interpretations/uses of assessment scores, heavily influenced by educational policies.

ELP standards and constructs for accountability assessments

Inevitably under standard-based education reform, the changes to academic and ELP standards exert a tremendous impact on the construct and content of ELP assessments. As briefly noted earlier, pre-NCLB ELP assessments primarily measured basic English language skills in interpersonal and social contexts (Butler et al., 2004; Schrank et al., 1996). Moreover, there was no requirement to assess all four of the language skills (listening, reading, speaking, writing) despite the assessments' use for identifying EL students to provide appropriate instructional support for them. Due to the limited construct being measured, the results of the traditional ELP assessments were criticized for not reflecting whether an EL student is at the level of readiness or competency to perform in an academic setting (Solórzano, 2008).

The enactment of NCLB led many states to rush to develop or adopt ELP standards and assessments with little guidance on defining the ELP construct (Boals et al., 2015; Wolf et al., 2008). At the time the NCLB-era ELP standards and assessments were developed, the construct of academic English language, or the language of school, had not been effectively defined (DiCerbo et al., 2014). As a result, the ELP construct was represented variously in different standards and assessments; for example, some existing ELP standards and assessments embodied different approaches to representing academic vs. social language, and discrete vs. integrated language skills (Bailey & Huang, 2011; Forte et al., 2012; Wolf & Faulkner-Bond, 2016). In their examination of NCLB-era ELP assessments, Wolf and Faulkner-Bond (2016) found that three states' ELP assessments included different types and degrees of academic and social language proficiency. For instance, one state's ELP assessment contained more technical academic language contexts than the two other ELP assessments. The representation of general academic, technical academic, and social language contexts was also varied across the four language domains within each ELP assessment.

Many states underwent another wave of changes in ELP standards and assessments as a result of new college and career readiness standards, such as the Common Core State Standards and the next-generation science standards. As mentioned earlier, ESSA reinforced the alignment of ELP standards and academic content-area standards. The federal peer review also required states to submit evidence to demonstrate how their content-areas (e.g., language arts, mathematics, science) and ELP standards were aligned to each other (US Department of Education, 2018).

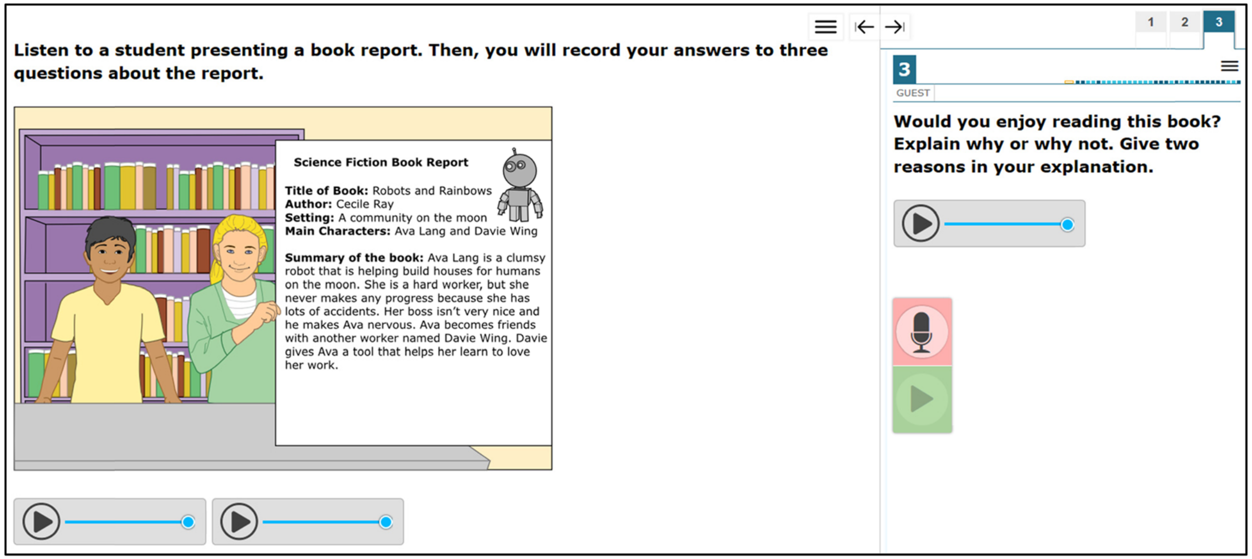
To address the challenges that EL students would face with the implementation of these college and career readiness standards, a number of researchers attempted to unpack the language demands embedded in these standards (e.g., Bailey & Wolf, 2020; Bunch, 2013; Hakuta et al., 2013; Lee, 2017; Moschkovich, 2012, Wolf et al., 2022). Porter et al. (2011)'s alignment study indicated that the Common Core State Standards, in fact, contained more cognitively complex and academically rigorous expectations for students to achieve compared to states' previous academic content standards. Sophisticated and increased language demands to meet the common core have been noted for the design and development of ELP assessments. For instance, Bunch (2013) characterizes the language and literacy demands in the common core as engaging complex informational texts from a variety of sources (reading standards), constructing arguments with evidence in writing and research (writing standards), working collaboratively while understanding multiple perspectives and presenting ideas (speaking and listening standards), and developing linguistic resources to do the abovementioned tasks effectively (language standards). For the common core-mathematics, Bunch describes how the high language demands include defining problems, explaining procedures, justifying conclusions, and creating evidence-based arguments, to name a few.

To support EL students in meeting these challenging academic standards, states and consortia of states endeavored to reflect the language demands of academic standards in developing new ELP standards or modifying existing ELP standards. While general consensus emerged on the close interconnection between language and content implied within the common core, different approaches to operationalizing the ELP construct in ELP standards and assessments were formulated (Wolf et al., 2016). In the current ESSA period, two consortia of multiple states named the English Language Proficiency Assessment for the 21st Century (ELPA21) and WIDA combined together to serve over 40 out of 50 states with their respective ELP standards and assessments.

Broadly put, WIDA's ELP standards describe the social, instructional, and academic language that students need to engage in school (WIDA, 2014, 2020). Academic language is represented as the language of language arts, mathematics, science, and social studies. WIDA modified its existing ELP standards to augment the correspondence of language demands between the common core and WIDA's ELP standards (WIDA, 2014). Recently, WIDA (2020) released its 2020 edition of the standards, further specifying the integration of language and content while taking a more functional approach to language development (e.g., focusing on key language use and functions such as narrate, inform, argue, and explain across multiple content areas) (Molle & Wilfrid, 2021). To operationalize the standards in ELP assessments[2], for example, sample WIDA listening items in Grades 6–8 contain teacher talk on how to measure the area of a table in a mathematics class. These items, then, assess students' understanding of the mathematical procedures and terminology explained in the teacher talk (see the WIDA website, https://wida.wisc.edu/assess/access/preparing-students/practice for sample items).

On the other hand, ELPA21 created brand-new ELP standards, adopting an approach to identifying common language practices described across disciplinary-area standards (Stage et al., 2013). ELPA21's standards explicitly state that they attempted to include "the *critical language, knowledge about language*, and *skills using language* that are in college-and-career-ready standards and that are necessary for English language learners (ELLs) to be successful in schools." (CCSSO, 2014, p. 1, italics in the original text). This approach enforced a strong presence of general academic language skills in ELPA21's ELP assessments. For instance, one of the 10 ELPA21 standards in Grades 4–5 says "An ELL can construct grade appropriate oral and written claims and support them with reasoning and evidence" (CCSSO, 2014, p. 19). This standard is tightly aligned with one of the speaking and listening standards (under the Presentation of Knowledge and Ideas) in Common Core State Standards-English Language Arts (ELA). This Grade 5 ELA standard expects students to be able to "report on a topic or text or present an opinion, sequencing ideas logically and using appropriate facts and relevant, descriptive details to support main ideas or themes; speak clearly at an understandable pace" (NGA & CCSSO, 2010, p. 24). This standard about constructing arguments or claims with reasoning and evidence also resonates with a set of key practices delineated in the common core-mathematics and next-generation science standards. Figure 1 presents a sample ELPA21 item in order to illustrate how this ELPA21 standard is assessed in the ELPA21 speaking section. This item is intended to primarily cover the ELPA21 standard mentioned above, assessing a student's communicative ability to construct an opinion with reasoning and evidence. It is worth noting that the context of student presentations on a book report is provided, reflecting disciplinary classroom contexts with their

[2] The sample WIDA items described in this article are aligned to the 2012 WIDA standards. Although the new WIDA standards were released in 2020, ELP assessments aligned to the 2020 WIDA standards have not yet been developed.



Listen to a student presenting a book report. Then, you will record your answers to three questions about the report.

Science Fiction Book Report

Title of Book: Robots and Rainbows
Author: Cecile Ray
Setting: A community on the moon
Main Characters: Ava Lang and Davie Wing

Summary of the book: Ava Lang is a clumsy robot that is helping build houses for humans on the moon. She is a hard worker, but she never makes any progress because she has lots of accidents. Her boss isn't very nice and he makes Ava nervous. Ava becomes friends with another worker named Davie Wing. Davie gives Ava a tool that helps her learn to love her work.

Would you enjoy reading this book? Explain why or why not. Give two reasons in your explanation.

Fig. 1 A released sample ELPA21 assessment item, Grades 4–5. Copyright © 2021 by the English Language Proficiency Assessment for the 21st century (ELPA21). Reprinted with permission

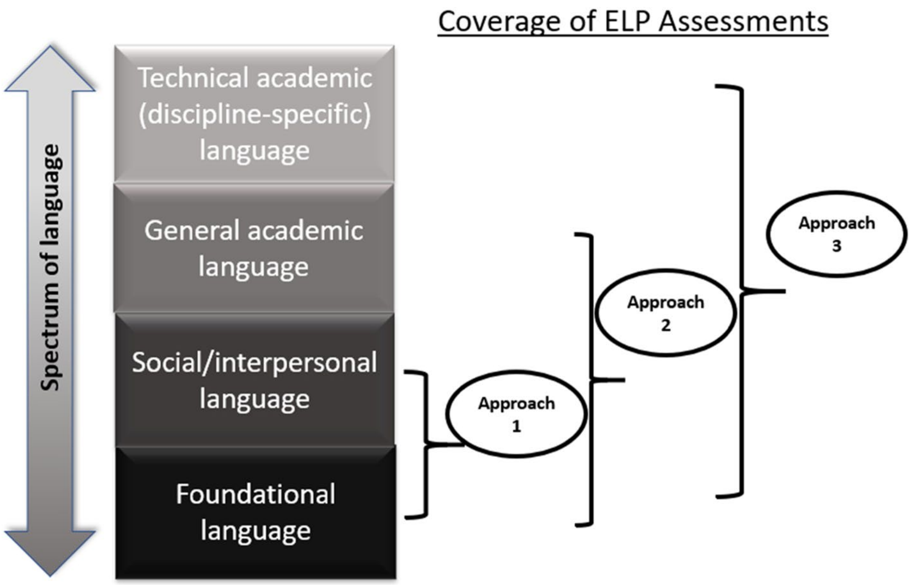

Coverage of ELP Assessments
Spectrum of language
Technical academic (discipline-specific) language
General academic language
Social/interpersonal language
Foundational language
Approach 1
Approach 2
Approach 3

Fig. 2 Different approaches to cover the ELP construct across ELP assessments

integration of listening, reading, and speaking skills. It is also worth noting that these skills are now being assessed for relatively young students (i.e., Grades 4–5). Both the WIDA and ELPA21 examples demonstrate the sophisticated language demands in current ELP standards and assessments that have resulted from an increased rigor in academic content standards.

Variability of ELP constructs and potential consequences

The changes in the constructs and content of ELP assessments described in the previous section entail differential ramifications at various levels and for stakeholders who make various decisions based on ELP assessment results. To discuss the interconnections between the construct and consequences, Fig. 2 summarizes different approaches taken to operationalize the ELP construct across ELP assessments and over the course of standard-based reform in the US K–12 public education. Figure 2 also represents the spectrum view of language (Snow, 2010), moving away from the traditional dichotomous view of social vs. academic language skills.

Prior to NCLB, ELP assessments employed Approach 1 in which foundational and social/interpersonal language skills were predominantly covered in their construct. Currently, ELP assessments use either the second (covering from foundational to general academic language skills) or the third (from foundational to technical academic language skills) approaches, to represent the language skills described in academic content standards. In the third approach, the ways to include the discipline-specific language skills (e.g., explaining mathematical procedures; making conjectures from a science experiment) across ELP assessments can also differ, as exemplified by ELPA21's and WIDA's ELP assessments. In addition to these two different approaches by ELPA21 and WIDA, large EL-populated states such as Arizona, California, New York, and Texas have implemented their own state-developed ELP assessments with their own approaches (Wolf et al., 2016). This landscape illustrates how the ELP constructs in current ELP assessments remain as varied as in the NCLB period.

When considering the possible consequences associated with this construct variability, the comparability of ELP assessments and in turn fair accountability across schools has come to be an important validity concern. The score interpretations and inferences made about a student's ability may be quite different depending on which ELP assessment the student takes. Moreover, it may take longer for a student to exit from EL status depending on the types of ELP assessments that were administered, given previous literature suggesting that academic language skills take longer to develop compared to social language skills (Hakuta et al., 2000; Cummins, 2008).

Taking the third approach in Fig. 2 for example, one of the unintended consequences from a challenging ELP construct can be "late" EL exit decisions. This raises considerable fairness concern in that EL students must take both academic content-area and ELP assessments measuring challenging academic language skills whereas non-EL students take only content-area assessments that have no impact on individual students' academic path (e.g., course selection). Since EL students must meet the proficiency level based on ELP assessments to exit out of the EL designation, challenging ELP assessments contribute to a preponderance of long-term ELs—students who remain designated as ELs and stay in EL programs for 6 years or more. These EL students experience barriers to educational opportunities such as the limited opportunity to access the rigorous courses that are available to their non-EL peers (Umansky, 2016).

The third approach also raises a question about the construct-irrelevant source of variance pertaining to the content knowledge inadvertently measured in ELP assessments where content knowledge is not an intended construct. The potential positive consequence of this approach, on the other hand, is to foster a close coordination between ESL/language teachers and content-area teachers to instruct the language skills needed for content learning. ESL instruction can move to include more rigorous language skills involved in disciplinary areas (e.g., constructing arguments, making source-based presentations), in addition to the foundational language skills (e.g., phonetic, morphological, and syntactic formation) EL students need.

There are other possible consequences potentially resulting from the specific construct of ELP assessments. Related to accountability, the number of EL students who meet the proficiency level in ELP assessments can be partly a function of the challenging construct measured in the assessments. Professional development (both pre- and in-service) training and instructional materials at the teacher and school/district levels will also be impacted by the ELP construct covered in specific ELP assessments of use.

It is inarguably important to include the academic language construct in ELP assessments (approaches 2 and 3) since the assessment scores and the levels associated with the scores should indicate that the student possesses ELP to handle academic materials and tasks in school settings. However, the best practice to operationalize the academic language construct in ELP assessments for the current purposes warrants continued investigation. Defining the ELP construct of current accountability testing should be a balanced act, considering not only the theories of L2 development but also the consequences implicated at the level of individual students, teachers, schools, and policymakers. This effort should also be accompanied by empirical validation research for providing evidence-based guidance for accountability testing.

Validation research areas related to the interconnections between construct and consequences

Thus far, I have described the major shifts of the ELP construct in the US K–12 ELP assessments due to standard-based accountability policies, along with a brief account of the potential consequences resulting from the construct shifts for various stakeholders and at the different educational system levels. In this section, I discuss imminent research areas to support ELP testing for accountability, particularly pertaining to the intersection between the ELP assessment constructs and consequences. I propose specific research directions for each area.

Area 1: expanding ELP alignment investigation

Validation efforts for accountability testing have traditionally centered on the technical qualities of assessment instruments ensuring that score interpretations and inferences made from assessment results for various decisions are defensible. However, considering the intended effects of successful reform and students' learning outcomes, validity arguments for accountability testing must encompass research on the assessment's consequences (Bennett, 2010; Chalhoub-Deville, 2016, 2020; Lane, 2020). The traditional focus on the technical properties of assessments is still evident in the federal peer review process of different states' accountability systems. As described earlier, the US peer review regulatory guidance (US Department of Education, 2018) specifies that states submit validity evidence based on assessment content, response processes, internal structure, and relations to other variables, following the framework laid out in the *Standards for Educational and Psychological Testing*. Farnsworth (2020) points out that the peer review guidance neglects consequential validity despite its prominence in *Standards for Educational and Psychological Testing* and other well-established validity theories (e.g., Kane, 2013; Messick, 1996) as one of the major types of validity evidence.

Concerning evidence based on assessment content, states are only required to submit evidence on alignment between their ELP assessments and ELP standards. While this content alignment between standards and assessment is one necessary type of validity evidence, alignment evidence must expand to be inclusive of curriculum and instruction, particularly in standard-based accountability. The underlying premise of alignment is that there should be tight and transparent associations among what is taught and learned (objectives, standards), how the content is taught (curricula, instruction), and what is assessed (assessments) to promote students' learning outcomes (Porter, 2002). Lane (2020) also notes that accountability policies are intended to result in positive consequences such as improving student achievement as well as enhancing curriculum and instruction. Assuming that ELP assessments are well-aligned with ELP standards, ELP assessments' construct and task types can serve as a vehicle to instantiate standards for teachers. It is expected that teachers who are familiar with their states' ELP assessment content and results endeavor to align their instruction and curriculum with the construct of their ELP assessments. In the language testing field, washback research has yielded ample evidence of instructional changes resulting from high-stake language assessment use (e.g., Cheng et al., 2004; Tsagari & Cheng, 2017).

Future research concerning alignment should address: (a) how ELP standards, assessments, curricula, and instruction are aligned to one another; (b) the extent to which teachers (both language and content-area teachers) are familiar with ELP standards and assessments (e.g., standard coverage, test content, score reports); and (c) whether and what instructional changes have taken place from the use of states' new ELP assessments. Some recent studies examining teachers' understanding about new academic and ELP standards have shown that teachers may have varied interpretations about standards and a somewhat limited understanding of the academic language embodied in standards (Neugebauer & Heineke, 2020; Wolf et al., 2022). These findings raise questions about the extent to which ELP assessments and standards bring about intended consequences for instruction and standard-based accountability. These studies also suggest that more empirical research is needed to examine the types of professional support provided for teachers to understand the core language knowledge and skills embodied in ELP standards and assessments.

More comprehensive alignment research on the construct and consequences can also be beneficial for continuously improving accountability systems, including the ELP standards themselves. While the construct of ELP assessment must be driven by the states' ELP standards, the quality and appropriateness of the content of ELP standards require further research, both theoretically and empirically. Empirical alignment research coupled with a growing body of the academic language literature based on K–12 schooling (e.g., Bailey et al., 2018; Gebhard & Harman, 2011; Haneda, 2014; Uccelli et al., 2014) will offer valuable knowledge and evidence to strengthen the ELP standards and accountability testing.

Area 2: examining the assessment performance of current EL and exited EL students

Validity evidence based on the relation to other variables or measures should also be expanded to shed light on the consequences of ELP accountability testing. To date, only a handful of criterion-related validity studies are available with the US K–12 ELP accountability assessments (Cook et al., 2012; Parker et al., 2009, Wolf & Faulkner-Bond, 2016). Since ELP assessments are used to determine EL students' exit from EL status (i.e., removing EL-related services), these studies utilized content-area (e.g., English language arts, mathematics, science) assessments as criterion measures and examined the relationships between ELP and content assessment performance of EL students. In particular, Cook et al. (2012) argue that research on the relationship between ELP and content assessment performance is useful to determine the point at which EL students' ELP is no longer a major hindrance to their performance on academic content assessments. Using the data of ELP and content (ELA and mathematics) assessment scores from three different states with sizable EL populations, they found a pattern of a diminishing relationship between language and content scores as EL students reach higher levels of ELP. This pattern would suggest that language proficiency reaches a maximum level of prediction of content performance, after which the prediction ceases to increase even as ELP continues to improve. The researchers point out that relating these available data from ELP and academic content assessments could provide empirical evidence to support policymakers in selecting performance ranges for ELP assessment standard setting (e.g., determining cut scores on an ELP assessment to make an EL exit decision). The cut scores indicating the "English proficient" level have considerable impact on students' instruction, school evaluation, and funding allocation for EL education. Thus, empirical investigation of criterion-related evidence for ELP assessments is paramount for ensuring the intended consequences.

Importantly, this line of criterion validation research should be accompanied by the content analysis of ELP assessments as well as of criterion measures to the extent this is possible (i.e., understanding the constructs of the measures of interest). Operational assessment materials may not be accessible to researchers for security reasons. However, publicly available assessment information such as test specifications, blueprints, technical reports, practice tests, and sample items must be critically examined for adequate interpretations and inferences about criterion-related validity evidence for ELP assessments.

Area 3: collecting various stakeholders' practice and perspectives on ELP accountability

The impacts of ELP assessments in accountability contexts are far-ranging at the individual, institutional (schools, districts, states' educational agencies), societal, and policy levels. The washback literature in the language-testing field has often employed rich qualitative investigation to analyze the impacts of high-stake assessment use based on various stakeholders' perspectives (e.g., Cheng et al., 2004). However, there is a paucity of empirical research that examines relevant stakeholders' perspectives on the impacts of ELP assessments in the US K–12 accountability contexts. Past research on collecting stakeholders' views on the US K–12 accountability programs and testing has primarily focused on general education and mainstream teachers (i.e., content areas). This area of research has provided valuable insights into both the positive and unintended impacts of accountability policy and testing. For example, Hamilton et al.'s (2007) study examined state-, district-, and school-level stakeholders' perceptions about the changes resulting from the NCLB accountability requirements through surveys, interview, and site visits in three states over 3 years. The study reported the positive, intended impact of all stakeholders' efforts to align curriculum and instruction to state standards. Interestingly, school principals reported that they made efforts to ensure that instruction was also aligned with state assessments, indicating the heightened attention to assessments. Additional notable changes included the use of assessment results for instructional planning and the provision of extra learning opportunities for low-performing students and other subgroup students due to the accountability requirement of disaggregated assessment reporting by subgroups.

Undesirable changes were also reported, particularly by teachers, including a narrowing of curriculum and instruction to focus on the assessments' contents. This study also found discrepant perceptions among superintendents, principals, and teachers regarding the degree of positive impacts of accountability testing and programs. For instance, district/school administrators and teachers perceived the adequacy of test scores reflecting student achievement differently with administrators being more positive about the adequacy. Teachers also noted the inconsistency between state accountability policies and local resources available to support the policy. In addition, they pointed out the lack of support for students' basic skills and expressed concerns about unrealistic expectations of the NCLB goals. These types of careful studies, which have a representative sample and systematic qualitative data collection, offer important empirical evidence to evaluate the consequences of accountability testing and programs and thus further improve accountability policies. It is imperative to conduct similar studies in the realm of K–12 ELP accountability testing.

Recently, in the context of the US K-12 ELP assessment, Kim et al. (2020) investigated how teachers interpreted the terms and information presented in the score reports of WIDA's ELP assessment. Their findings indicate that the need for professional support is evident to enhance the teachers' assessment literacy to adequately interpret the score reports. Their study signals the importance of taking account of teachers' assessment literacy when investigating the stakeholders' perspectives on the consequences of K-12 ELP accountability assessments.

As Chalhoub-Deville (2016, 2020) argues, the validation of accountability testing for successful education reform should involve a broad range of stakeholders in order to investigate the reform processes and impacts. She urges language-testing researchers to be proactive and undertake societal impact assessment analyses to support appropriate policy-making and the valid use of language assessments for accountability purposes. Currently, states are required to develop a theory of action for their accountability systems, delineating the intended consequences and any unintended adverse impacts (Lane, 2020). To conduct such qualitive investigations involving various stakeholders in a principled, systematic way, language-testing researchers may utilize a theory of action as a framework to examine the ELP accountability testing impacts. By doing so, the links between the ELP assessment constructs and their associated consequences will be better understood for informing the areas of improvement.

Conclusion

Standard-based reform coupled with test-based accountability in the U.S. K–12 context has promoted positive consequences such as deliberate efforts of alignment of standards with instruction and assessments among various stakeholders, data-driven instructional planning, and attention to subgroups of students (Lane, 2020; Spurrier et al., 2020). At the same time, unintended adverse consequences have also emerged, including the use of test scores as the basis for teacher evaluation and instruction practices of "teaching to the test." The heavy emphasis on test scores led some states to establish a monolingual instructional policy for EL education. A number of researchers have raised serious concerns about such negative consequences of diminishing the long-term benefits of bi/multilingual education for students (e.g., Menken et al., 2014; Solórzano, 2008).

ELP assessments play a vital role in standard-based accountability in the US K–12 education, concerning millions of EL students and educators. ELP accountability testing can act as a lever to enact positive educational reform and to support EL students' achievement. For instance, the explicit presence of academic language proficiency in standards, assessments, and instruction was one of the intended impacts of ELP accountability policies. Yet, due to the high-stake usage of ELP assessments for individual students, there is an inherent tension in using ELP assessments for accountability. As laid out in the previous sections, there is an urgent need for language-testing and educational researchers to forge collaborative and systematic validation research. I have highlighted the further research areas of expanding alignment investigation, examining both current EL and exited EL students' academic performances over time, and gathering relevant stakeholders' perspectives on ELP accountability. These areas of research are certainly fraught with challenges in collecting adequate data, requiring substantial resources and collaboration among researchers, practitioners, administrators, and policy makers. However, taking on these challenges is essential to ensure the validity and adequacy of current ELP accountability testing and to foster its intended positive impacts for students and other relevant stakeholders.

Abbreviations

CCSSO	Council of Chief State School Officers
EL	English learner
ELA	English language arts
ELP	English language proficiency
ELPA21	English Language Proficiency Assessment for the 21st Century
ESSA	Every Student Succeeds Act

Acknowledgements

The author would like to thank Micheline Chalhoub-Deville (the editor of this special issue on accountability testing) and other reviewers for their helpful comments and suggestions on earlier drafts of this paper.

Author's contributions

Sole author. The author(s) read and approved the final manuscript.

Funding

No funding was received for this paper.

Availability of data and materials

Not applicable.

Declarations

Competing interests

The author declares no competing interests.

Received: 1 March 2022 Accepted: 11 September 2022
Published online: 07 October 2022

References

Abedi, J. (2008). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice*, *27*(3), 17–31. https://doi.org/10.1111/j.1745-3992.2008.00125.x.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Bailey, A., Maher, C., & Wilkinson, L. C. (Eds.) (2018). *Language, literacy, and learning in the STEM disciplines: Language counts for English learners*. Routledge.

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

Bailey, A. L., & Huang, B. H. (2011). Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing*, *28*(3), 343–365. https://doi.org/10.1177/0265532211404187.

Bailey, A. L., & Wolf, M. K. (2020). The construct of English language proficiency in consideration of college and career ready standards. In M. K. Wolf (Ed.), *Assessing English language proficiency in K–12 U.S. schools*, (pp. 36–54). Routledge. https://doi.org/10.4324/9780429491689-3.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives*, *8*, 70–91. https://doi.org/10.1080/15366367.2010.508686.

Boals, T., Kenyon, D. M., Blair, A., Cranley, M. E., Wilmes, C., & Wright, L. J. (2015). Transformation in K–12 English language proficiency assessment changing contexts, changing constructs. *Review of Research in Education*, *39*(1), 122–164. https://doi.org/10.3102/0091732X14556072.

Bunch, G. C. (2013). Pedagogical language knowledge preparing mainstream teachers for English learners in the new standards era. *Review of Research in Education*, *37*(1), 298–341. https://doi.org/10.3102/0091732X12461772.

Butler, F. A., Bailey, A. L., Stevens, R., Huang, B., & Lord, C. (2004). *Academic English in fifth-grade mathematics, science, and social studies textbooks (CSE technical report, 642)*. University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Callahan, R., & Shifrer, D. (2016). Equitable access for secondary English learner students course taking as evidence of EL program effectiveness. *Educational Administration Quarterly*, *52*(3), 463–496. https://doi.org/10.1177/0013161X16648190.

Chalhoub-Deville, M. (2016). Validity theory: reform policies, accountability testing, and consequences. *Language Testing*, *33*(4), 453–472. https://doi.org/10.1177/0265532215593312.

Chalhoub-Deville, M. (2020). Toward a model of validity in accountability testing. In M. K. Wolf (Ed.), *Assessing English language proficiency in K–12 U.S. schools*, (pp. 245–264). Routledge.

Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing: Research contexts and methods*. Lawrence Erlbaum Associates.

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*. U.S. Department of Education.

Council of Chief State School Officers. (2014). *English language proficiency (ELP) standards with correspondences to K–12 English language arts (ELA), mathematics, and science practices, K–12 ELA standards, and 6-12 literacy standards*.

Cummins, J. (2008). BICS and CALP: empirical and theoretical status of the distinction. In B. Street, & N. H. Hornberger (Eds.), *Encyclopedia of language and education*, (vol. 2, pp. 71–83). Springer.

Deville, C., & Chalhoub-Deville, M. (2011). Accountability-assessment under No Child Left Behind: Agenda, practice, and future. *Language Testing*, *28*(3), 307–321. https://doi.org/10.1177/0265532211400876.

DiCerbo, P., Anstrom, K., Baker, L., & Rivera, C. (2014). A review of the literature on teaching academic English to English language learners. *Review of Educational Research*, *84*(3), 446–482. https://doi.org/10.3102/0034654314532695.

Every Student Succeeds Act. (2015). Public Law No. 114-354.

Farnsworth, T. (2020). A review of validity evidence on K–12 English language proficiency assessment: Current state and future direction. In M. K. Wolf (Ed.), *Assessing English language proficiency in K–12 U.S. schools*, (pp. 75–91). Routledge.

Forte, E., Kuti, L., & O'Day, J. (2012). *National evaluation of Title III implementation: A survey of states' English language proficiency standards*. U.S. Department of Education. Retrieved from https://www2.ed.gov/rschstat/eval/title-iii/national-implementation-report.pdf.

Gebhard, M., & Harman, R. (2011). Genre theory in K–12 schools: A response to high-stakes school reforms in the United States. *Journal of Second Language Writing*, *20*(1), 45–55. https://doi.org/10.1016/j.jslw.2010.12.007.

Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* University of California Linguistic Minority Research Institute.

Hakuta, K., & Pompa, D. (2017). *Including English learners in your state Title I accountability plan*. Council of Chief State School Officers. Retrieved from https://files.eric.ed.gov/fulltext/ED580945.pdf.

Hakuta, K., Santos, M., & Fang, Z. (2013). Challenges and opportunities for language learning in the context of the CCSS and the NGSS. *Journal of Adolescent & Adult Literacy*, *56*(6), 451–454. https://doi.org/10.1002/JAAL.164.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russel, J., Naftel, S., & Barney, H. (2007). *Standards-based accountability under No Child Left Behind experiences of teachers and administrators in three states*. Retrieved from https://www.rand.org/pubs/monographs/MG589.html.

Hamilton, L. S., Stecher, B. M., & Yuan, K. (2012). Standards-based accountability in the United States: Lessons learned and future directions. *Education Inquiry*, *3*(2), 149–170. https://doi.org/10.3402/edui.v3i2.22025.

Haneda, M. (2014). From academic language to academic communication: Building on English learners' resources. *Linguistics and Education*, *26*, 126–135. https://doi.org/10.1016/j.linged.2014.01.004.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000.

Kim, A. A., Chapman, M., Kondo, A., & Wilmes, C. (2020). Examining the assessment literacy required for interpreting score reports: a focus on educators of K–12 English learners. *Language Testing*, *37*(1), 54–75. https://doi.org/10.1177/0265532219859881.

Lane, S. (2020). *Test-based accountability systems: The importance of paying attention to consequences (research report no. RR-20-02)*. Educational Testing Service. https://doi.org/10.1002/ets2.12283.

Lee, O. (2017). Common Core State Standards for ELA/literacy and next generation science standards: Convergences and discrepancies using argument as an example. *Educational Researcher*, *46*(2), 90–102. https://doi.org/10.3102/0013189X17699172.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4–16. https://doi.org/10.2307/1177052.

Linquanti, R., & Cook, H. G. (2015). *Re-examining reclassification: Guidance from a national working session on policies and practices for exiting students from English learner status*. Council of Chief State School Officers.

Menken, K., Hudson, T., & Leung, C. (2014). Symposium: Language assessment in standards-based education reform. *TESOL Quarterly*, *48*(3), 586–614. https://doi.org/10.1002/tesq.180.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–103). American Council on Education & Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*, 243–256. https://doi.org/10.1177/026553229601300302.

Molle, D., & Wilfrid, J. (2021). Promoting multilingual students' disciplinary and language learning through the WIDA framework for equitable instruction. *Educational Researcher*, *50*(9), 585–594. https://doi.org/10.3102/0013189X211024592.

Moschkovich, J. (2012). *Mathematics, the Common Core, and language: Recommendations for mathematics instruction for ELs aligned with the Common Core*, *Commissioned paper by the Understanding Language Initiative* (). Stanford University Retrieved from http://ell.stanford.edu/papers/.

National Governors Association Center for Best Practices, Council of Chief State School Officers (2010). *Common core state standards*.

National Research Council (2011a). *Allocating federal funds for state programs for English language learners*. National Academies Press.

National Research Council (2011b). *Incentives and test-based accountability in education*. National Academies Press.

Neugebauer, S. R., & Heineke, A. J. (2020). Unpacking K–12 teachers' understandings of academic language. *Teacher Educational Quarterly*, *47*(2), 158–182.

NGSS Lead States (2013). *Next generation science standards: For states, by states*. National Academies Press.

O'Day, J. A., & Smith, M. S. (2019). *Opportunity for all: A framework for quality and equality in education*. Harvard Education Press.

Parker, C., Louie, J., & O'Dwyer, L. (2009). *New measures of English language proficiency and their relationship to performance on large-scale content assessments. U.S. Department of Education*. Retrieved from https://files.eric.ed.gov/fulltext/ED504060.pdf.

Pereira, N., & de Oliveira, L. C. (2015). Meeting the linguistic needs of high-potential English language learners: What teachers need to know. *Teaching Exceptional Children*, *47*(4), 208–215. https://doi.org/10.1177/0040059915569362.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core Standards: The new U.S. intended curriculum. *Educational Researcher*, *40*(3), 103–116. https://doi.org/10.3102/0013189X11405038.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*(7), 3–14. https://doi.org/10.3102/0013189X031007003.

Sahlberg, P. (2006). Education reform for raising economic competitiveness. *Journal of Educational Change*, *7*(4), 259–287.

Schleppegrell, M. J. (2012). Academic language in teaching and learning: Introduction to special issue. *The Elementary School Journal*, *112*(3), 409–418.

Schrank, F. A., Fletcher, T. V., & Alvarado, C. G. (1996). Comparative validity of three English oral language proficiency tests. *Bilingual Research Journal*, *20*(1), 55–68. https://doi.org/10.1080/15235882.1996.10668620.

Snow, C. E. (2010). Academic language and the challenge of reading for learning. *Science*, *328*(5977), 450–452. https://doi.org/10.1126/science.1182597.

Solórzano, R. W. (2008). High stakes testing: Issues, implications, and remedies for English language learners. *Review of Educational Research*, *78*(2), 260–329. https://doi.org/10.3102/0034654308317845.

Spurrier, A., Alderman, C., Schiess, J., & Rotherham, A. J. (2020). *The impact of standards-based accountability*. Bellwether Education Partners Retrieved from https://files.eric.ed.gov/fulltext/ED606418.pdf.

Stage, E. K., Asturias, H., Cheuk, T., Daro, P. A., & Hampton, S. B. (2013). Opportunities and challenges in next generation standards. *Science*, *340*, 276–277. https://doi.org/10.1126/science.1234011.

Tanenbaum, C., Boyle, A, Soga, K., Carlson, L., Golden, L., et al. (2012) *National evaluation of Title III implementation: Report on state and local implementation*. U.S. Department of Education. Retrieved from https://www2.ed.gov/rschstat/eval/title-iii/state-local-implementation-report.pdf.

Tsagari, D., & Cheng, L. (2017). Washback, impact, and consequences revisited. In E. Shohamy, I. Or, & S. May (Eds.), *Language testing and assessment*, (3rd ed., pp. 359–372). Springer.

U.S. Department of Education (2018). *A state's guide to the U.S. Department of Education's assessment peer review process*. Retrieved from https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf.

U.S. Department of Education, Office of English Language Acquisition [OELA]. (2021). *English learner population by local education agency fact sheet*. Retrieved from https://ncela.ed.gov/files/fast_facts/20210315-FactSheet-ELPopulationbyLEA-508.pdf.

Uccelli, P., Barr, C. D., Dobbs, C. L., Phillips Galloway, E., Meneses, A., & Sánchez, E. (2014). Core academic language skills (CALS): An expanded operational construct and a novel instrument to chart school—relevant language proficiency in pre-adolescent and adolescent learners. *Applied Psycholinguistics*, *36*, 1077–1109. https://doi.org/10.1017/S014271641400006X.

Umansky, I. M. (2016). Leveled and exclusionary tracking: English learners' access to core content in middle school. *American Educational Research Journal*, *53*(6), 1792–1833. https://doi.org/10.3102/0002831216675404.

WIDA (2014). *2012 Amplification of the English language development standards, kindergarten–grade 12 ("WIDA ELD Standards")*. Board of Regents of the University of Wisconsin System.

WIDA (2020). *WIDA English language development standards framework, 2020 edition, Kindergarten–grade 12*. Board of Regents of the University of Wisconsin System.

Wolf, M. K., Bailey, A. L., Ballard, L., Wang, Y., & Pogossian, A. (2022). *Unpacking the language demands in academic content and English language proficiency standards for English learners*. International Multilingual Research Journal. Advance online publication. https://doi.org/10.1080/19313152.2022.2116221.

Wolf, M. K., & Farnsworth, T. (2014). English language proficiency assessments as an exit criterion for English learners. In A. Kunnan (Ed.), *The companion to language assessment*, (vol. 1, pp. 303–317). Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla118.

Wolf, M. K., & Faulkner-Bond, M. (2016). Validating English language proficiency assessment uses for English learners: Academic language proficiency and content assessment performance. *Educational Measurement: Issues and Practice*, *35*(2), 6–18. https://doi.org/10.1111/emip.12105.

Wolf, M. K., Guzman-Orth, D., & Hauck, M. C. (2016). *Next-generation summative English language proficiency assessments for English learners: Priorities for policy and research (ETS research report no. RR-16-08)*. Educational Testing Service. https://doi.org/10.1002/ets2.12091.

Wolf, M. K., Kao, J., Herman, J. L., Bachman, L. F., Bailey, A., Bachman, P. L., … Chang, S. M. (2008). *Issues in assessing English language learners: English language proficiency measures and accommodation uses - literature review (CRESST technical report 731)*. University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Publisher's Note