tesol QUARTERLY
tesol international association

SYMPOSIUM

This symposium brings together international scholars to discuss issues arising from current practices in the design and use of language assessments for students designated as English language learners in standards-based contexts. We asked each of the contributors to comment on a particular aspect of language assessment, with a particular focus on validity. The authors use England and the United States as two case studies to demonstrate ways in which this topic is highly relevant to current and future scholarship and policy in TESOL.

Symposium: Language Assessment in Standards-Based Education Reform

KATE MENKEN
City University of New York
New York, United States

THOM HUDSON
University of Hawai'i at Manoa
Honolulu, Hawai'i, United States

CONSTANT LEUNG
King's College London
London, England

This symposium article, to which three authors contribute distinct parts, presents the rationale for standards-based language assessment and examines both the uses and misuses of language assessments in English-speaking countries that are engaged in standards-based education reform. Specifically, they focus on the assessment of emergent bilinguals (also referred to as English language learners or English as an additional language students). The first part lays out the intentions and challenges of standards-based language assessment for emergent bilinguals, focusing on validity concerns. The second part describes classroom-based teacher-led assessments of emergent bilinguals in England, which carry high stakes along with the national standardized tests. This contrasts with what is happening in the United States where, as the third part describes, the main focus is on high-stakes standardized testing for purposes of accountability. In

addition to the challenges inherent in attempts to measure language in meaningful ways, a thread cutting across the authors' accounts is the widespread practice of high-stakes standardized testing. The U.S. and English cases show how issues of validity arise when emergent bilinguals are simply included into assessments intended for English monolinguals without appropriate differentiation, and when an assessment is used for purposes beyond what it was designed to do. As all of the authors of this symposium article contend, assessments —particularly when standardized—hold the potential to dominate standards-based education reform efforts when they are ultimately summative and attached to severe consequences.

doi: 10.1002/tesq.180

■ In educational systems globally, standards-based reforms have galvanized renewed interest in assessment as a means to ensure students are indeed attaining established standards. In this symposium article, we examine uses and misuses of language assessments in English-speaking countries engaged in standards-based reform. Specifically, we focus on assessment of emergent bilinguals[1] (also referred to as English language learners and English as an additional language students by this article's respective authors).

This area is of pressing concern for several reasons. First, immigrants' knowledge of the dominant language has historically played a central role in integration into the new country, with language a requirement for national identity, citizenship, and societal participation—as asserted through formal and informal language assessments (McNamara & Shohamy, 2008). Schoolchildren are usually expected to study academic content in the new language, and language testing is spreading internationally as a means for emergent bilinguals to demonstrate acquisition of language and literacy—especially the language variety used for academic purposes (Shohamy, 2001).

Additionally, emergent bilinguals are found to underperform in comparison to their peers on language and content assessments. A comparison between the performance of immigrants and nonimmigrants on the Programme for International Students Assessment (PISA) in 17 countries of the Organisation for Economic Co-operation and Development (OECD) shows immigrant students perform at lower levels not only on language and literacy tests but also on tests of other subjects (OECD, 2006); these findings are supported by data from

[1] García, Kleifgen, and Falchi (2008) use the phrase *emergent bilinguals:* "That is, through school and through acquiring English, these children become *bilingual*" (p. 6). The term is used here and by Menken below.

many countries (see Menken, 2013). This is not surprising given that content proficiency is a positive function of language proficiency on exams administered in the new language, yet closing this "achievement gap" is nonetheless seen as cause for concern (Crawford, 2008; OECD, 2006).

In the first part of this article, Thom Hudson explains intentions and challenges of standards-based language assessment for emergent bilinguals, providing an example of English-proficiency testing in the United States. Leung and Menken then localize the discussion, examining language assessments in England and the United States, respectively, both countries that are implementing standards-based reforms. Leung describes classroom-based, teacher-led assessments of emergent bilinguals in England, carrying high stakes along with national standardized tests. This contrasts with the situation in the United States, where, as Menken describes, the focus is on high-stakes standardized testing for accountability. A common thread between the countries is how issues of validity arise when emergent bilinguals are included into assessments intended for English monolinguals without appropriate differentiation, and when assessments are used for purposes beyond their design (see also Mislevy & Durán, 2014).

OVERVIEW OF STANDARDS-BASED LANGUAGE ASSESSMENT FOR ENGLISH LANGUAGE LEARNERS (BY THOM HUDSON)

Overview of Standards in Language Assessment

Teachers and evaluators make decisions about success or lack thereof based on standards, explicitly or implicitly defined. Standards-based efforts are intended to base curricular, assessment, and teacher decisions on statements regarding students' expected skills. Standards provide benchmarks for that evaluation. When done correctly, standards spell out expectations for students, engage stakeholders, ensure continuity across levels, and involve school community members in improving education.

However, large societal disagreements exist about roles and functions of standards, particularly effects of standards-based assessment. Primary areas of contention relate to how constrained standards are and how they are used. Some worry that the standards movement has been transformed into a testing movement (Popham, 2001; Thompson, 2001). Value interpretations are directly related to uses for which standards are applied. This must be acknowledged because all

assessment is designed for decision making. To the extent that standards-based assessments are used for multiple and disparate (often unvalidated) decisions, they become increasingly contentious. Whether stakeholders (teachers, administrators, students, or parents) believe standards are being used for making appropriate inferences about appropriate decisions is key. In short, discussion centers on validity issues. The current discussion offers a brief overview of large-scale, primarily standardized, summative language assessments.

Language standards have been generated by myriad institutions for multiple purposes. Standards come from transnational organizations (e.g., Council of Europe, 2001); government agencies (e.g., Ministry of Education, Singapore, 2001; U.S. Defense Language Institute [Wilds, 1975]); state departments of education (e.g., California Department of Education, 1999), state consortia (e.g., National Governors Association, 2010), and educational organizations (e.g., Educational Testing Service [ETS]; American Council on the Teaching of Foreign Languages [ACTFL, 2012]; CASAS, 2012), among others. Standards may be general, such as "Orally communicate basic personal needs and desires" (California Department of Education, 1999, p. 6) or quite specific, as in "Can write short simple personal letters expressing thanks and apology" (Council of Europe, 2001, p. 19). Standards vary in specificity for many reasons involving institutional contexts. Some institutions are governmental with clearly defined curricular objectives (e.g., school districts), while others are extragovernmental (e.g., ACTFL; Council of Europe), addressing instructional settings without common curricula. Standards' specificity level can affect how many standards are generated and assessed. This issue of standards proliferation will be addressed below.

Validity and Standards-Based Assessment

Messick (1980, 1989) developed a validity framework foregrounding the need to see test interpretation and use both in terms of test construct and relevance, as well as consequences of test application. As such, validity concerns have moved beyond psychometric and internal construct criteria to evidence-based assessment and social consequences of assessment more broadly. Messick thus moved concerns beyond a focus on test instruments, rubrics, and specifications, to consequences for curriculum, instruction, training, and accountability. While validity in assessment is generally important regardless of consequences, this inclusion of consequences should provide background for subsequent discussions of standards-based assessment. While procedural identification, selection, and operationalization of standards is

essential to maximizing outcomes, evaluation of standards-based assessment must also address its effects on students' actual learning and future opportunities. Messick (1989) states:

> A construct-centered approach [to assessment design] would begin by asking what complex of knowledge, skills, or other attribute should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? (p. 17)

Mislevy, Almond, and Lukas (2003) maintain:

> This perspective is valuable because it helps organize thinking about assessments for all kinds of purposes, using all kinds of data, task types, scoring methods, and statistical models. We can ask of a simulation task, for example, just what knowledge and skills is it meant to reveal?
>
> Do the scoring methods pick up the clues that are present in performances? How is this evidence synthesized across multiple tasks, or compared when different students attempt different tasks? (pp. 3–4)

Thus, all decisions about assessment of English language learners (ELLs) can be examined in terms of impact of including standards in the assessment and the forms those standards take. The appropriateness of score-based generalizations about examinees' skills can be evaluated. Evidence can indicate decisions made about the nature of language (and language learning) and appropriateness of assessment form.

Specific Issues With Standards-Based Assessment

Standards-based education can guide instruction, help in articulation across grade levels, and focus assessment. General progression occurs through Standards → Curriculum → Instruction → Tests. Within this process, four areas should be embedded in assessment design, use, and interpretation: types, proliferation, prescriptiveness, and outcomes of standards (Popham, 2001), all reflecting concerns with validity and answering the question: What are the consequences of standards-based assessment?

Types. Standards' function should be clearly maintained, as either content or performance standards. Content standards state what students should know, while performance standards indicate performance levels expected. Discussions of standards must not conflate issues of learning standards content with performance standards assessment, as

the two require separate validation processes. Content standards' acceptance does not itself determine assessment form, either as large-scale standardized assessment or as teacher-developed classroom assessment. However, formal performance standards, when applied to content standards in large-scale educational contexts, often imply standardized assessment. This is particularly true when legislation establishing standards requires assessment (e.g., No Child Left Behind Act [NCLB, 2001], as further detailed by Menken below). Performance standards require operationalizing content standards into some format allowing demonstration of the required mastery level. Performance standards provide performance descriptors and expectations, as well as example assessment tasks. However, performance standards should be established empirically. There are several approaches for doing this (Berk, 1984; Brown & Hudson, 2002; Cizek, 2001; Popham, 1978). Fundamentally, it involves contrasting what individuals known as masters can do with what those known not to be masters can do. Too often, performance standards are either artifacts of historical norms (e.g., passing is 60%), or results of wishful thinking. For example, NCLB set a standard that "all students in each group ... described in subparagraph (C)(v) will meet or exceed the State's proficient level of academic achievement on the State assessments" by 2013–2014. Thus, the standard is that all children (100%!) will be grade-level proficient. While this may be desirable, it is unrealistic and non-empirically based.

Proliferation. Selecting standards for inclusion in assessment is complex, requiring attention to several concerns. The primary requirement is for the assessment to truly represent the construct being measured. The construct should not be underrepresented, yet the assessment must only require reasonable time and resources. Thus, it is important to have a limited number of standards. This is often difficult because curriculum designers want completeness in developing content standards. Frequently, too many standards are developed to be assessed, and the standards are often too specific. In instructional settings, broad standards such as, "Determine an author's point of view or purpose in a text and analyze how an author uses rhetoric to advance that point of view or purpose" (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010, p. 40) are often broken into a set of narrow standards for pedagogical purposes. It is tempting to test these pedagogical standards, but that becomes unwieldy. It is also tempting to include only easily tested standards, avoiding such areas as using language for appreciation, debate, or in novel applications. This reductionism can lead to construct underrepresentation.

Prescriptiveness. Generalizability and prescriptiveness must be addressed in determining standards' appropriateness across linguistic contexts. English-language users are increasingly diverse in terms of language background, needs, and contexts. As noted previously, standards establish expectations regarding examinee performance. Expectations can become problematic when standards developed for one context are adopted for a context reflecting greater diversity. English language learners, English lingua franca users, and indigenous English variety users imply differing evaluations of appropriate language. For example, ACTFL Proficiency Guidelines ACTFL, 2012) contain evaluation facets such as the following:

> Speakers at the Intermediate Mid sublevel are able to handle successfully a variety of uncomplicated communicative tasks in straightforward social situations. Conversation is generally limited to those predictable and concrete exchanges necessary for survival in the target culture . . . In spite of the limitations in their vocabulary and/or pronunciation and/or grammar and/or syntax, Intermediate Mid speakers are generally understood by sympathetic interlocutors accustomed to dealing with non-natives.

The focus on *survival in the target culture* may be irrelevant for English users internationally since their language goals may not involve interest in the target culture. Likewise, *sympathetic interlocutors accustomed to dealing with non-natives* may not identify actual interlocutors in many English contexts. This raises the question: Whose English do the standards model? Standards must be examined for generalizability to the target audience, possibly including ELLs in an English-dominant country, an international context, children, or adults.

Outcomes. It is important to examine broad results of standards implementation, directly related to Messick's (1989) notions of value implications and social consequences in the validation process. In the best of all worlds, English language standards congruent with appropriate goals and contexts will bring about learning that meets all stakeholders' needs, uses, and desires. Successful learning would be valued, and learners would be rewarded for hard work. Unfortunately, this has not always been the case. Widespread application of standards in U.S. schools, for example, has often led to instruction being interpreted entirely in terms of test scores. As Popham (2001) notes, too often educators get caught in score-boosting frenzies, not considering whether scores reflect learning. Schools and teachers are evaluated based on scores, often without examining personal, social, and economic variables affecting them. This assignment of values

such that high scores indicate "good" schools and low scores mean "poor" schools, has also placed evaluation almost entirely on single-shot summative tests rather than formative evaluation more likely to promote learning. Additionally, value ascribed to high scores has frequently led to classroom instruction narrowed to the areas targeted by tests. This can lead administrators to devalue teacher skills, resulting in a restrictive curriculum and narrowing the scope of education.

Current State English Language Proficiency Assessment in the United States

One of the most salient developments since 2010 in standards-based assessment has been adoption of the K–12 Common Core State Standards (CCSS) by states throughout the United States. The standards were developed as a framework articulating needs for college and career readiness, covering mathematics as well as English language arts and literacy in history/social studies, technical subjects, and science (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010), along with the Next Generation Science Standards, or NGSS (NGSS Lead States, 2013).

Before the CCSS's development and adoption, the federal government urged states to develop/adopt English language proficiency (ELP) standards linked to the states' content standards to address the diverse needs of ELLs throughout the United States. Large states such as California and New York developed their own ELP standards. As an alternative approach, a consortium of 33 states and territories formed the World-Class Instructional Design and Assessment (WIDA), which developed ELP standards and assessments beginning in 2004, initially to address NCLB requirements. The standards reflect a theoretical view of language learning focused on integrating language with content and communication, incorporating English for social and instructional purposes, language arts, mathematics, science, and social studies. The framework reflects five proficiency levels, including performance indicators (PIs) for each level on each of the four language domains (listening, speaking, reading, and writing).

ACCESS for ELLs® is the WIDA ELP framework's test operationalization. The ELP standards are the basis for test item specifications. Kenyon (2006) notes:

> The standards provide guidance to both the content of the items on the ACCESS as well as to their difficulty level. The model performance indicators (PIs) and the performance levels described in the standards

informed the development of generative specifications of the field test items. (p. 174)

Over 400 PIs were designed for the initial test, spread across four grade-level groupings, the five standards (e.g., language arts), the four language domains, and the five proficiency levels.

The fundamental focus on language for interaction and content learning provides the WIDA model a theoretically sound foundation. However, the approach is open to the validity issues raised above. The standards avoid some of the tautological issues of conflating content and performance standards by keeping the two standards types distinct until the development of PIs. Thus, a PI addressing *main ideas and details* in reading might for one level involve *find identifying information*, while a higher-level PI would involve *interpret text to identify main ideas* (WIDA, 2007). Standards proliferation is a general potential problem but is controlled through careful assessment sampling and test development at different grade bands. However, achieving construct representation of five language standards across each test administration while controlling test length is demanding. In implementation, the standards focus almost exclusively on summative evaluation and accountability. Initial WIDA documents incorporate a framework for classroom instruction, but focus on ACCESS has detracted from focus on formative evaluation.

WIDA has now largely committed to incorporating the CCSS (WIDA, 2013), bringing both opportunities and challenges. The CCSS have been designed to establish common-language skills across contents but do not incorporate ELLs' English development. Along with content, they focus on socialization into school discourse and college and career preparation, areas often difficult to assess. However, CCSS implementation includes a call for formative assessment through the ASSETS assessment system being developed by WIDA (see http://www.wida.us/assessment/assets.aspx).

Concluding Remarks

This standards-based language assessment overview has attempted to contextualize subsequent discussions. Explanation has focused on broader assessment validity issues. Although considerable threats are emerging from inferences sometimes drawn from standards-based language assessment, there are also considerable strengths to be pursued and usefully applied. One clear potential advantage is more effective teaching. However, that generally only happens when focus is on

standards promoting effective teaching and learning, with less preoccupation with summative standards assessment.

"COMMON" ASSESSMENT FOR ALL? FORMATIVE ASSESSMENT ISSUES FROM THE NATIONAL CURRICULUM IN ENGLAND (BY CONSTANT LEUNG)

Curriculum and Assessment Context

Students in England are becoming increasingly diverse. Over 1 million students had a language background other than English in 2013, 18.1% and 13.6% of elementary and secondary school populations respectively (see http://www.naldic.org.uk/research-and-information/eal-statistics). Over 50% of secondary school students in London have ethnic-minority backgrounds (Hamnet, 2011). Over 30% of London secondary school students and 40% of elementary students are considered English as an additional language (EAL) learners (von Ahn, Lupton, Greenwood, & Wiggins, 2010).

For 30 years, particularly since the implementation of the National Curriculum in 1992, school policies in England have consistently followed principles of common entitlement and inclusive access for all students. In practice this means all students, irrespective of ethnic and language backgrounds, are expected to follow age/grade-related curriculum and assessment arrangements, which must conform to statutory requirements in the National Curriculum (e.g., Department for Education [DfE], 2013a, 2013b; also see Leung, 2009, 2010, 2012a for policy analyses). The National Curriculum (revised version to be introduced in September 2014) delineates content (called programmes of study) of subjects such as English, mathematics, science, and history to be covered in elementary and secondary schools.

English as an Additional/Second Language is not itself considered a subject; therefore, there is no programme of study for EAL. The official curriculum states:

> Teachers must ... take account of the needs of pupils whose first language is not English. Monitoring of progress should take account of the pupils' age, length of time in this country, previous educational experience and abilities in other languages.... Teachers should plan teaching opportunities to help pupils develop their English and should aim to provide the support pupils need to take part in all subjects.
>
> (DfE, 2013a, 2013b, p. 8)

EAL learners are expected to develop capacities to use English through study of the subjects in the medium of English in age-graded classes (including the subject English, the content of which is normed on the general English as a first language student population). Teachers of all subjects are expected to support EAL development in classroom work, using language-minority students' first/preferred language to promote understanding where possible. EAL is not a subject offered in initial teacher-education programmes, so individual teachers' EAL-related repertoires often vary in accordance with their experiences and interests. There is no statutory requirement for teachers to have formal training in EAL teaching.

- Regarding content specification, the National Curriculum delineates statutory requirements for subjects and grade levels. For instance, for the subject English, the spoken language component requires elementary pupils (first six years of schooling) be taught to, *inter alia*, listen and respond appropriately to adults and their peers; articulate and justify answers, arguments and opinions; and select and use appropriate registers for effective communication. (DfE, 2013a, 2013b, p. 17)
- Teachers are advised this content "should be taught at a level appropriate to the age of the pupils" (DfE, 2013b, p. 17). For the reading component, statutory requirements include two sub-components, word reading and comprehension, each carrying grade-specific requirements. For example, for word reading first grade (Year 1, aged 5 to 6), teachers are instructed to teach students to, *inter alia*, apply phonic knowledge and skills as the route to decode words, read accurately by blending sounds in unfamiliar words containing GPCs [grapheme-phoneme correspondences] that have been taught, and read aloud accurately books that are consistent with their developing phonic knowledge and that do not require them to use other strategies to work out words. (DfE, 2013a, 2013b, p. 20)

These requirements comprise both broad and specific "standards" statements. But the National Curriculum will not have published rating/attainment scales with level attainment descriptors from 2014/15 (DfE, 2013). Schools are expected to apply such grade/age-related statutory requirements universally to students of all language backgrounds, irrespective of their English proficiency and previous schooling.

Regarding statutory assessment the National Curriculum framework is currently comprised of two elements: (1) national tests following elementary schooling Year 6 (aged 11) and school-leaving examinations following secondary school (aged 16); and (2) classroom-based teacher-led assessment (called *teacher assessment* throughout) at the end

of Year 2 (aged 7) and Years 7, 8, and 9 (aged 12, 13, and 14). Statutory assessment results are published and used by the government to evaluate school performance (with funding and other implications), and parents often choose schools based on published school performance *league tables*, rendering statutory assessment high stakes. For teacher assessment required for summative reporting within the statutory framework (e.g., teacher conducting an end-of-year assessment), schools are asked to use published (past) assessment tasks. A main reason why statutory assessment includes teacher assessment is that heavy reliance on national tests in the 1990s and 2000s to boost achievement did not produce expected outcomes. The "test-led" push was also costly and received persistent criticism from educators and communities (see Leung & Rea-Dickins, 2007). Given this symposium article's focus, I will concentrate on teacher assessment.

Summative teacher scores are supported by published rubrics and external moderation (e.g., DfE, 2103a, Section 7). Teachers are advised that in summative judgments of students' work, they should use consistent assessment criteria for all students:

> Summative assessment for bilingual pupils [students from minority language backgrounds, including EAL learners], as for all pupils, should be based on national curriculum measures. . . . It is not recommended that additional locally developed scales of fluency are used for summative purposes.
>
> (Department for Education and Skills, 2005, p. 6)

These statements clearly indicate that all students' performances should be rated against common criteria, with performances of native English-speaking students and EAL learners in all subjects judged against the same expectations. For instance, for English content after Year 9 (aged 14) all students should speak confidently and effectively, including through using Standard English confidently in a range of formal and informal contexts, including classroom discussion (DfE, 2013b, p. 17). Teachers are expected to interpret this (and all other statutory criteria) in an "inclusive" way; EAL students are assessed and rated the same ways as native English-speaking students. Like U.S. policy (NCLB), setting the unrealistic goal that all students receive proficient scores regardless of ELL status, applying assessment criteria in a nondifferentiated way has not been empirically validated; its expected outcomes raise validity concerns, as Hudson presented above.

In addition to the statutory summative assessment, schools are encouraged to develop formative assessments, also known within the National Curriculum as Assessment for Learning (AfL) (Department for Children, Families and Schools, 2008). The following broad

definition of AfL from the Assessment Reform Group has articulated this approach's value:

> Assessment for Learning [formative assessment] is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there.
>
> (Assessment Reform Group, 2002)

Presently, formative assessment/AfL is largely understood as teacher-led assessment integrated into everyday teaching activities. It has received renewed support in official curriculum discourse: "Schools [are] to introduce their own approaches to formative assessment, to support pupil attainment and progression" (DfE, 2013c). Current official formative assessment guidance does not cover EAL issues (see below).

Teacher Assessment: Research Perspective

From a research perspective, teacher assessment in England's schools raises interesting questions. I address two connected issues concerning teacher professional knowledge and practice. First, most teachers in England recognize that many EAL learners face English-related challenges across the curriculum; for some teachers, the advice to use only mainstream assessment criteria does not fit with "tak[ing] account of the needs of pupils whose first language is not English" (DfE, 2013a, 2013b, p. 8). Referencing Hudson's discussion, assessment policy in England is prescriptive in this regard, in that evaluating students by the same criteria dictates homogeneity despite the diverse population. It is thus important to ask: How far do teachers assess EAL students using only official "common" criteria (as recommended by curriculum authorities)?

Leung's (2012b) study explored the criteria adopted by a group of English subject teachers while marking EAL learners' written class assignments. These teachers worked in a London school where 75% of the students came from EAL backgrounds. The teachers had between 3 and 25 years of professional experience. The assignments were by six students in Years 9 and 10 (14- and 15-year-olds). Analysis of the teachers' audiorecorded think-aloud data suggests they invoked a range of considerations, some of which are recognisable in terms of established teaching approaches, such as writing as technical skills and conventions (e.g., spelling, punctuation), writing as creativity (e.g., showing poetic quality), writing as process (e.g., drafting, proof-read-

ing), and writing in genre (e.g., following conventionalised discourse moves). Additionally, teachers also appeared to attend to person-oriented considerations, such as student effort (e.g., "[the student is trying to use the possessive form] but doesn't get that at all ... but she has worked very hard and I'm going to tell her on this that she is on the right lines" [Leung, 2012b, p. 5]). Another concern was EALness (e.g., one of the teachers remarked, "I have left footmarks [uncorrected] which we would normally say footprints instead of because there's a limit to how much you can take on board at this stage in a language acquisition" [p. 5]). Findings from this exploratory study suggest the participant teachers exercised multifaceted judgments. Incidentally, they often displayed both summative and formative concerns in think-aloud comments.

The second issue concerns the pedagogic basis of teacher assessment, particularly in relation to formative assessment. Popular professional literature often gives teachers advice on open-ended initiation-response-evaluation, longer wait time, learning-oriented feedback, and other techniques. While these are certainly helpful techniques, their formative value should be studied against "hidden" teacher-assessment aspects related to disciplinary perspectives and teaching and learning theories (Black & Wiliam, 1998; Black, Harrison, Lee, Marshall, & Wiliam, 2003).

Teachers are not neutral, classroom-procedure operators; they have opinions on their subject and pedagogic and educational principles more generally. So the questions they ask and feedback they offer will depend, at least partially, on their disciplinary beliefs and pedagogic principles. For instance, a teacher who believes that learning English is primarily about learning vocabulary and grammar and that learning is achieved through practice and memorization, will likely ask different questions, and provide different feedback than someone who believes language learning is achieved through student-led meaningful communication in real-life contexts. If we believe that formative assessment's ultimate purpose is enabling students to learn, then teachers' repertoires should be comprised of not just techniques like asking open-ended questions but also capacity to transcend their preferred disciplinary values by considering alternative ways of addressing students' issues in their feedback. A related issue is how much the educational environment encourages teacher initiative and "flair." Cowie (2009) suggests teachers and students need considerable pedagogic space and time to negotiate their way to shared knowledge and understanding. In an environment where learning outcomes and the content-cum-process of learning are prespecified (e.g., phonics instruction in the National Curriculum), knowledge may well be seen as "given" and transparent, and teaching-learning may be menu-driven, what

Crossouard and Pryor (2012) call representational epistemology. How far formative assessment can move beyond prescribed curriculum parameters in these circumstances is a moot point.

Concluding Remarks

The inclusive approach to school curriculum and assessment in England has ensured common participation in public education for students irrespective of language background. However, *the same English for all* principle has created considerable complexity for assessment (and pedagogy more generally), particularly in terms of appropriateness and validity of universal "standards" for linguistic minority students at various stages of English learning. The laissez faire approach to EAL teaching offers teachers no guidance for effective assessment, raising serious reliability and validity issues. The case for more student-sensitive "standards" and pedagogic practice would be strengthened by paying attention to relevant professional experience and research.

STANDARDS-BASED LANGUAGE ASSESSMENT IN THE UNITED STATES: A CASE OF TESTING MISUSE (BY KATE MENKEN)

This part of the article focuses on a widespread form of language assessment in the United States, now also carrying extremely high stakes for students, their teachers, and their schools—using standardized tests to assess emergent bilinguals' language proficiency and content knowledge. With rapid immigration, the U.S. population has grown increasingly diverse in recent years, such that approximately 20% of those older than five come from homes where languages other than English are spoken, and about 13% were born abroad (U.S. Census Bureau, 2013). Ten percent of U.S. public school students, or about 4.7 million students, are emergent bilinguals—an increase of about 600,000 from just 8 years ago (National Center for Education Statistics, 2013). They predominantly speak Spanish (approximately 3.6 million students), and state data indicates the following languages are most common after Spanish (their language categories): Vietnamese, Chinese, Arabic, Hmong, Haitian, Tagalog, Somali, and Navajo (National Clearinghouse for English Language Acquisition, 2011).

As the school population has gained diversity, I document how U.S. high-stakes standardized testing exemplifies many of the validity threats

Hudson outlined above, particularly regarding outcomes and social consequences. While standards-based assessment can be teacher-led and rooted in classroom practices, as noted by Leung regarding England's curriculum, this is not the main focus of current U.S. reforms.[2]

Assessment and Accountability Policies for U.S. Emergent Bilinguals

The Elementary and Secondary Education Act's (ESEA) 1994 reauthorization mandated the creation and adoption of academic standards and corresponding assessment systems inclusive of emergent bilinguals (U.S. Department of Education, 1994). Because the U.S. education system is decentralized, each state developed its own standards and assessments, according to its definition of designated ELLs (Linquanti & Cook, 2013). The 2001 ESEA reauthorization, entitled No Child Left Behind (NCLB) (NCLB Act, 2001), remains in effect, emphasizing assessment for accountability purposes so that the federal government can ensure their public school funding results in student performance gains. The law's assessment and accountability requirements have resulted in nationwide use of high-stakes standardized achievement tests, typically administered in English only.

Specifically, all students—including emergent bilinguals—are required to make *adequate yearly progress* toward achieving *annual measurable achievement objectives* with the unreasonable expectation, as noted above, that all students score "proficient" on reading and math tests by 2014 (NCLB Act, 2001).[3] Faced with the mandate that all students be assessed, states began requiring that emergent bilinguals take the same tests as those taken by (and developed for) English monolinguals, with a set of accommodations that vary across states to *level the playing field* (Abedi, Hofstetter, & Lord, 2004; Solórzano, 2008). Emergent bilinguals have the additional requirement of *demonstrated improvement* in learning English, based in most states on a standardized English proficiency exam like WIDA's ACCESS, described above. Each school receives state-calculated achievement objectives; repeated failure results in sanctions, possibly even restructuring or closure. Adding to these high stakes, many states also use scores to determine individual students' grade promotion and

[2] This may change, because states adopting consortium assessments are required to implement and report formative assessments (Bunch, 2011).

[3] Since 2012 about 42 states have received waivers from some stringent NCLB aspects, for example, that all students score proficient on state tests, often granting additional time to meet this goal (*Education Week*, 2013).

graduation and to evaluate teacher effectiveness (Menken & Solorza, 2014).

The Obama administration in 2010 gave selected states funding through Race to the Top, a grants competition program advancing standards and assessments, systems for gathering and analyzing data measuring student progress, teacher effectiveness, and improvement of failing schools—all areas for which states rely upon standardized testing (Menken & Solorza, 2014). Seeking commonality among standards across states, this program also encouraged states to adopt the Common Core State Standards (CCSS). Forty-six states have adopted the CCSS, signifying agreement for at least 85% of state standards in mathematics and English language arts (CCSS Initiative, 2013). The CCSS were recently implemented in schools, and CCSS assessments are to begin in 2014–2015.

Two state consortia, the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC), received federal grants of approximately $185 million each to develop new CCSS-aligned assessments (SBAC, 2012). Thus far, 45 states have joined them (Gewertz, 2013). The U.S. Department of Education (2013a) recently published a report critiquing both consortia for failure to focus on emergent bilinguals. Both PARCC and SBAC have since outlined testing supports for emergent bilinguals, using the same accommodations paradigm as NCLB (wherein students receive state-determined accommodations, such as extra time, bilingual dictionaries, and/or test translations) (Maxwell, 2013). As the United States begins assessing CCSS, no fundamental changes are being proposed regarding assessment of emergent bilinguals, positioning U.S. policy to repeat mistakes made under NCLB.

Moreover, assessment of English proficiency and academic content operates from a monolingual, monoglossic (García, 2009) paradigm, setting expectations that bilinguals should learn and perform on assessments the same ways monolinguals do (see also Flores & Schissel, 2014). Yet recent research highlights how bilinguals' cognitive and linguistic practices differ from monolinguals' (Blackledge & Creese, 2010; Brutt-Griffler & Varghese, 2004; Cook, 2012; García, 2009). Following Grosjean's work (1989), Brutt-Griffler and Varghese (2004) propose: "Far from being monolinguals in two languages, as it were, they carve out their own space as *bilinguals"* (p. 93). Thus bilinguals' distinctive qualities must be understood and evaluated independently of monolinguals, or they will always be positioned as failures (Cook, 2012). Referring to Hudson's discussion above, there is some prescriptiveness here, compromising validity for an assessment's failure to adequately acknowledge test-takers' diversity.

English Language Proficiency (ELP) Testing Practices and Effects

One type of high-stakes tests emergent bilinguals must take measures ELP to evaluate students' reading, writing, speaking, listening, and comprehension under NCLB. NCLB requires further that states implement an ELP assessment aligned to their content standards, measuring academic-language proficiency, and using results to track a state's attainment of annual measurable achievement objectives (Abedi, 2007; NCLB Act, 2001). ELP tests are used to classify students as ELLs, and predict their ability to perform at the level of their English monolingual peers in English-medium classrooms and on English-administered content tests (Solórzano, 2008).

ELP tests vary by state. In response to NCLB, states are implementing locally developed tests, using commercially available tests, or implementing one of the following four consortium-developed tests developed with NCLB funding: WIDA's ACCESS for ELLs (described by Hudson), Comprehensive English Language Learner Assessment (CELLA), English Language Development Assessment (ELDA), or Mountain West Assessment (MWA) (Bunch, 2011). Of these, the MWA is no longer being used, and others will likely change as states assess CCSS, as the following two consortia have received funds to develop new ELP assessments: the Assessment Services Supporting ELs through Technology Systems (ASSETS) and the English Language Proficiency for the 21st Century (ELPA21) Consortium.

Serious validity and reliability concerns have been raised regarding the ELP tests being used across the United States. As Solórzano (2008) writes,

> [T]he most widely used tests define English proficiency differently, are technically suspect, classify and thus reclassify ELLs differently and inappropriately, and rarely predict academic success in English classrooms. (p. 288)

Wolf et al. (2008) reinforce concerns in a thorough literature review, noting how little evidence there is of validity particularly on newly developed tests, given a dearth of empirical evaluations. Questionable ELP tests have serious consequences, including provision of inappropriate educational services, poor predictions of English-classroom success or ability to take English-administered content assessments, and unfair school penalties under current accountability practices.

One issue is that there is no agreed-upon definition of language proficiency or academic language. For instance, Valdés (2004) shows how different U.S. professional communities define academic

language differently, and Wolf et al. (2008) highlight challenges of defining academic language in measurable ways, drawing into question the very construct ELP tests aim to evaluate. ELP exams are thus often poor predictors; for example, although 52% of emergent bilinguals passed California's ELP test, earning reclassification as English proficient, only 10% actually passed its English Language Arts exam (Rumberger & Gándara, 2005). In a contrasting predictive shortcoming, Menken (2008) documented in New York numerous cases of students passing the English language arts exam yet failing the ELP test. Bailey and Huang (2011) attribute these gaps to test developers' lack of understanding about academic language, resulting in ELP tests of little predictive value. Regardless of such concerns, ELP scores are factored into schools' accountability ratings in certain states, placing schools under pressure to expedite language learning and somehow increase the numbers of students scoring proficient on ELP assessments (Deville & Chalhoub-Deville, 2011). While U.S. states have begun aligning ELP standards to the CCSS, concerns outlined here regarding ELP assessment have not been addressed, and existing accountability systems remain, as states move into the assessment phase of CCSS implementation.

Emergent Bilinguals' Performance on High-Stakes Content Tests

There are also problems of test validity and (mis)use related to extremely high-stakes content tests across the United States. The reality is that, even when intended to assess academic content knowledge, any English-medium test is actually a language proficiency exam for an English learner, because proficiency mediates test performance (Menken, 2008; Mislevy & Durán, 2014). While certain accommodations appear more effective than others, language proficiency level is still correlated to content test performance (see Cook, Linquanti, Chinen, & Jung, 2012); overall research on the effectiveness of accommodations for emergent bilinguals remains inconclusive (Schissel, 2010; Solórzano, 2008). What is clear is that emergent bilinguals consistently underperform in comparison to monolingual peers, making these students, their teachers, and schools serving them disproportionately likely to be penalized. The Government Accountability Office (GAO) (2006) reports that nearly two-thirds of states had percentages of emergent bilinguals scoring proficient lower than the state's annual progress goals.

I analyzed passing rates of emergent bilinguals in each state on reading and math tests reported to the federal government for accountability under NCLB (U.S. Department of Education, 2013b)

and used these to calculate national passing rates and compare emergent bilinguals to English monolingual peers. To be clear, I present this "achievement gap" data to highlight the inadequacy of currently available accommodations and problematize how language is a liability for emergent bilinguals, for whom testing is primarily punitive in outcomes (see Figures 1 and 2 for results). Figure 1 shows that on average U.S. emergent bilinguals score approximately 30 percentage points below their peers on fourth-grade reading tests, 40 percentage points below on eighth-grade reading tests, and 46 percentage points below on high school reading tests. That emergent bilinguals do not perform as well as English monolingual peers does not necessarily mean they are failing to acquire English or academic content but rather reinforces that these students are in fact language learners (Menken, 2008).

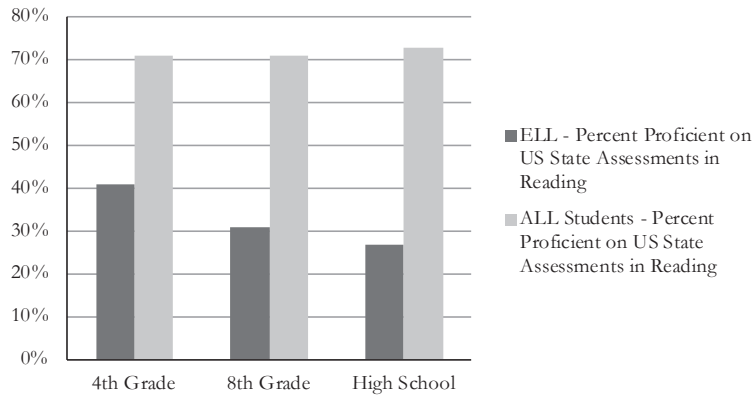

FIGURE 1. Percent of U.S. students receiving proficiant scores on reading assessments, 2011–2012: Average of state results.

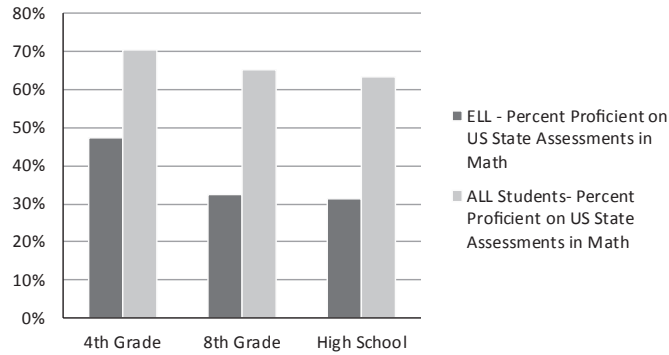

FIGURE 2. Percent of U.S. students receiving proficiant scores on math assessments, 2011–2012: Average of state results.

There is also an "achievement gap" in math (see Figure 2). Figure 2 shows that emergent bilinguals nationally score approximately 24 percentage points below peers on fourth-grade mathematics tests, 33 percentage points below on eighth-grade math tests, and 31 percentage points below in high school. Research indicates it is impossible to divorce language from content in assessments, as exemplified by these results (Abedi et al., 2004). Even in states where students may take exams in home languages,[4] translations, although helpful, cannot fully even out results because U.S. emergent bilinguals receive instruction at least partially in English, while researchers maintain language of instruction must match the language of the test for scores to be valid (Abedi et al., 2004; Menken, 2008; Solórzano, 2008). While data is less available for science and social studies, as these areas are not required for NCLB accountability, similar achievement gaps have been found in these subjects across different states (Laorenza, Pacheco, & Shah, 2012; O'Connor, Abedi, & Tung, 2012).

The student population for which a test is designed is likewise extremely important, as it affects the test's integrity and that of any decisions based on results (Solórzano, 2008). Consequently, numerous researchers have questioned the validity of using results of states' content assessments for high-stakes decisions such as grade promotion, graduation, or school evaluation (see Solórzano, 2008). While new CCSS assessments are being developed to replace current tests, states are speeding ahead before making significant efforts to redress validity threats, and concerns for emergent bilinguals are left at the periphery of reforms.

Effects of Test-Based Accountability on Emergent Bilinguals and Their Schools

Despite these validity concerns, high-stakes standardized testing is widespread in the United States, and pressures of test-based accountability are unforgiving of emergent bilinguals, their schools, and their teachers. For example, about half of states require that students pass high school exit exams to receive a diploma. Emergent bilinguals disproportionately fail these exit exams, barring them from graduation. Of 24 states with exit exams participating in a Center on Education Policy (2007) survey, 18 identified the performance gap between

[4] The GAO (2006) reported that 16 states offer translated versions of exams in some grades for certain emergent bilinguals, yet few emergent bilinguals in those states may actually take translated exams due to lack of availability in all students' languages and because some students have not developed home-language literacy skills.

emergent bilinguals and others as the most difficult to close. Based on California data, Salazar (2007, cited in Solórzano, 2008, p. 313) reports that "depending on which formula one uses (state or federal), English language learners graduate at either a 48.8% or 22.6% rate, respectively—both undesirable outcomes." In New York, which uses exit exams, graduation rates are lower for emergent bilinguals than other students (34% as compared to 74% [New York State Education Department, 2013]) and dropout rates are higher (33% as compared to 17%); the dropout rate for emergent bilinguals has increased by 14 percentage points in New York since testing requirements began (Menken, 2013).

Another way policymakers use state content-area tests is to determine grade retention, another area where emergent bilinguals are more vulnerable. This occurs despite research demonstrating retention does not necessarily improve achievement and increases likelihood that an emergent bilingual will drop out (Menken, 2008; Solórzano, 2008). Similarly, tests are being used in the evaluation of teachers. In New York City, for example, student scores on ELP and academic-content tests are substantially comprised of teacher evaluations recently introduced by local policymakers to improve teacher effectiveness. Teachers are now ranked individually and, accordingly, the city's "worst teacher" was an ESL teacher in 2012, about whom a disparaging article appeared in the *New York Post* with her photo (Roberts, 2012), prompting debate over fairness of such ratings when applied to teachers of emergent bilinguals.

Emergent bilinguals' schools are also negatively affected by student scores. As noted above, most U.S. schools serving emergent bilinguals fail to achieve the progress goals set for these students, making them more likely to face federal sanctions. In New York City, most of the 35 schools on the city's list of failing schools served above-average numbers of emergent bilinguals (Menken, 2009). Statewide, 49% of what the state terms *persistently lowest-achieving* schools—the schools at greatest risk of facing failure-related sanctions—serve emergent-bilingual populations of at least 10% (above the state average of 8%); 20% of these schools serve emergent-bilingual populations of at least 20% (New York State Education Department, 2011). Schools serving emergent bilinguals are clearly more likely to be restructured or closed under current policies.

Another testing byproduct attached to high-stakes consequences is tests' becoming *de facto* language policies in schools, determining what is taught, how, and in what language(s) (Menken, 2008; Shohamy, 2001). Testing *washbac*k refers to tests' effects on teaching and learning (Cheng & Watanabe, 2004). In U.S. schools, narrowing curriculum to focus on tested subjects and topics has been well documented, with reading and math receiving instructional time at the expense of other

areas, and with emergent bilinguals among those most likely to experience "teaching to the test" (Amrein & Berliner, 2002; Menken, 2008; Shepard, 2010). In schools under pressure to improve NCLB accountability reports, particularly on English-administered tests, high-stakes testing of emergent bilinguals has likewise prompted increased English-only instruction and elimination of bilingual programs, despite documented benefits of instruction using students' home languages (Gándara & Baca, 2008; Menken & Solorza, 2014). Despite deleterious effects of current high-stakes testing on emergent bilinguals, states will begin CCSS assessment before having resolved these problems.

Concluding Remarks

The U.S. case exemplifies testing misuse, when political agendas push a test's uses beyond intended purposes and when major decisions are based on questionable data. As the country moves into a new phase of standards-based reforms, most expect the new CCSS assessments to be even more challenging in their ELP and content-learning expectations. New York offers an example—in spring 2013, it implemented statewide CCSS tests in Grades 3–8; emergent bilinguals' passing rate dropped to a paltry 3.2% on the English language arts test and 9.8% on the math test,[5] while accountability policies remained unchanged. Problems that arose under NCLB in testing emergent bilinguals have not been addressed. It is worrying that the CCSS assessment approaches currently being developed by individual states, PARCC, and SBAC continue operating from the same accommodations paradigm, already proven ineffective and detrimental for emergent bilinguals under NCLB, leaving students disproportionately likely to fail high-stakes tests and face consequences (Menken, 2008; Schissel, 2010). As the United States moves into assessment of new standards, it should be of grave concern that efforts have not been made to redress the overreliance on standardized testing for accountability and the validity issues that have arisen.

CONCLUSION

This symposium presents rationales for standards-based language assessment and describes practices in the United States and England.

[5] This is down from emergent-bilinguals' passing rate of about 13% on English language arts tests and 30% on math tests for those same grades (New York State Education Department, 2011).

In addition to challenges inherent in attempts to meaningfully measure language, a unifying theme across authors is widespread practices of high-stakes standardized testing. Although teacher-led assessments in England might appear flexible, their "one-size-fits-all" approach ultimately mirrors standardized testing, limiting capacity to assess emergent bilinguals. In his critique of the international exporting and importing of standards-based reform, or what he terms *transnational standardization*, Luke (2011) writes:

> We now live in an era when schooling and education, teaching and learning have undergone a whole-scale redefinition by reference to a culture of accountability, performance, and measurability. (p. 367)

As this symposium's authors contend, assessment—particularly standardized testing—can entirely dominate standards-based reform efforts when assessments are ultimately summative and attached to high-stakes consequences. Given the power of testing, assessments rather than standards can drive instruction and school change, often in ways counter to reforms' stated aims. Moreover, when assessments are invalid or inappropriate for emergent bilinguals in the name of standardization, and when results carry serious negative consequences, they can cause these students more harm than good.

THE AUTHORS

Kate Menken is an associate professor of Linguistics at Queens College of the City University of New York (CUNY), and a research fellow at the Research Institute for the Study of Language in Urban Society at the CUNY Graduate Center.

Thom Hudson is a professor of second language studies at the University of Hawai'i at Manoa, and co-editor of the electronic journal *Reading in a Foreign Language*. His research has concentrated on second language testing, standards based assessment, reading, language for specific purposes, and curriculum and program development.

Constant Leung is a professor of educational linguistics at King's College London. His research interests include additional/second language curriculum development and language assessment. He is editor of research issues for *TESOL Quarterly* and senior associate editor for *Language Assessment Quarterly*. He is an Academician of Social Sciences (UK).

REFERENCES

Abedi, J. (2007). *English language proficiency assessment in the nation: Current status and future practice*. Title III Report. Davis: The Regents of the University of California.

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28. doi:10.3102/00346543074001001

American Council on the Teaching of Foreign Languages (ACTFL). (2012). ACTFL proficiency guidelines 2012. Retrieved from http://actflproficiencyguidelines2012.org/

Amrein, A., & Berliner, D. (2002). *An analysis of some unintended and negative consequences of high-stakes testing*. Tempe: Education Policy Research Unit, Arizona State University. Retrieved from http://www.asu.edu/educ/epsl/EPRU/epru_2002_Research_Writing.htm

Assessment Reform Group. (2002). Assessment for learning: 10 principles. Retrieved from http://assessmentreformgroup.files.wordpress.com/2012/01/10principles_english.pdf

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. doi:10.1177/0265532211404187

Berk, R. A. (Ed.). (1984). *A guide to criterion-referenced test construction*. Baltimore, MD: The Johns Hopkins University Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead, England: Open University Press.

Black, P., & Wiliam, D. (1998). *Inside the black box*. London, England: King's College.

Blackledge, A., & Creese, A. (2010). *Multilingualism: A critical perspective*. London, England: Continuum.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, England: Cambridge University Press.

Brutt-Griffler, J., & Varghese, M. (2004). Introduction. *International Journal of Bilingual Education and Bilingualism, 7*, 93–101. doi:10.1080/13670050408667803

Bunch, M. B. (2011). Testing English language learners under No Child Left Behind. *Language Testing, 28*, 323–341. doi:10.1177/0265532211404186

California Department of Education. (1999). English-language development standards for California public schools: Kindergarten through grade twelve. Retrieved from www.cde.ca.gov/be/st/ss/documents/englangdevstnd.pdf.

CASAS. (2012). CASAS basic skills content standards. Retrieved from https://www.casas.org/product-overviews/curriculum-management-instruction/casas-basic-skills-content-standards

Center on Education Policy. (2007). *State high school exit exams: Working to raise test scores*. Washington, DC: Author.

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

Cizek, G. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, N.J: Lawrence Erlbaum.

Common Core State Standards (CCSS) Initiative. (2013). *In the states*. Washington, DC: Author. Retrieved from http://www.corestandards.org/in-the-states

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

Cook, V. J. (2012). Some issues for SLA research. In L. Pedrazzini & A. Nava (Eds.), *Learning and teaching English: Insights from research* (pp. 39–68). Monza, Italy: Polimetrica.

Council of Europe. (2001). *Common European framework of reference for language: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
Cowie, B. (2009). Teacher formative assessment decision making: A consideration of principles and consequences. *Assessment Matters, 1*, 47–63.
Crawford, J. (2008). *Advocating for English learners: Selected essays*. Bristol, England: Multilingual Matters.
Crossouard, B., & Pryor, J. (2012). How theory matters: Formative assessment theory and practices and their different relations to education. *Studies in Philosophy and Education, 31*, 251–263. doi:10.1007/s11217-012-9296-5
Deville, C., & Chalhoub-Deville, M. (2011). Accountability-assessment under No Child Left Behind: Agenda, practice, and future. *Language Testing, 28*, 307–321. doi:10.1177/0265532211400876
Department for Children, Families and Schools. (2008). *The assessment for learning strategy*. Nottingham, England: DCFS Publications.
Department for Education. (2013a). *The national curriculum in England: Key stages 1 and 2 framework document*. London, England: Author. Retrieved from www.gov.uk/dfe/nationalcurriculum
Department for Education. (2013b). *The national curriculum in England: Key stages 3 and 4 framework document*. London, England: Author. Retrieved from www.gov.uk/dfe/nationalcurriculum
Department for Education. (2013c). *Assessing without levels*. London, England: Author. Retrieved from http://www.education.gov.uk/schools/teachingandlearning/curriculum/nationalcurriculum2014/a00225864/assessing-without-levels
Department for Education and Skills. (2005). *Aiming high: Guidance on assessment of pupils learning English as an additional language*. Nottingham, England: DfES.
Education Week. (2013, October 2). NCLB waivers: A state-by-state breakdown. Retrieved from http://www.edweek.org/ew/section/infographics/nclbwaivers.html
Flores, N., & Schissel, J. L. (2014). Dynamic bilingualism as the norm: Envisioning a heteroglossic approach to standards-based reform. *TESOL Quarterly, 48*(3), 454–479. doi:10.1002/tesq.182
Gándara, P., & Baca, G. (2008). NCLB and California's English language learners: The perfect storm. *Language Policy, 7*(3), 1–16. doi:10.1007/s10993-008-9097-4
García, O. (2009). *Bilingual education in the 21st century: A global perspective*. Malden, MA: Wiley-Blackwell.
García, O., Kleifgen, J., & Falchi, L. (2008). From English language learners to emergent bilinguals. *Equity Matters: Research Review No. 1*. New York: Campaign for Educational Equity.
Gewertz, C. (2013, October 21). Assessment consortia: Who's in and who's out? Curriculum Matters Blog. *Education Week*.
Government Accountability Office (GAO). (2006). *No Child Left Behind Act: Assistance from education could help states better measure progress of students with limited English proficiency*. Washington, DC: Author.
Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language, 36*, 3–15. doi:10.1016/0093-934X(89)90048-5
Hamnet, C. (2011). Concentration or diffusion? The changing geography of ethnic minority pupils in English secondary schools, 1999–2009. *Urban Studies* (online version), doi:10.1177/0042098011422573
Kenyon, D. (2006). *Development and field test of ACCESS for ELLs®*. Madison, WI: WIDA.

Laorenza, E., Pacheco, M., & Shah, H. (2012). *STEM inequity: New England's ethnic poverty and ELL achievement gaps*. Providence, RI: The Education Alliance at Brown University, New England Equity Assistance Center.

Leung, C. (2009). Mainstreaming: Language policies and pedagogies. In I. Gogolin & U. Neumann (Eds.), *Streitfall Zweisprachigkeit—The bilingualism controversy* (pp. 215–231). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Leung, C. (2010). English as an additional language: Learning and participating in mainstream classrooms. In P. Seedhouse, S. Walsh, & C. Jenks (Eds.), *Conceptualising learning in applied linguistics* (pp. 182–205). Basingstoke, England: Palgrave Macmillan.

Leung, C. (2012a). English as an additional language policy—Rendered theory and classroom interaction. In S. Gardner & M. Martin-Jones (Eds.), *Multilingualism, discourse, and ethnography* (pp. 222–240). Abingdon, England: Routledge.

Leung, C. (2012b). Qualitative research in language assessment. In *Encyclopedia of Applied Linguistics*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0979/pdf

Leung, C., & Rea-Dickins, P. (2007). Teacher assessment as policy instrument: Contradictions and capacities. *Language Assessment Quarterly*, *4*, 6–36. doi:10.1080/15434300701348318

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.

Luke, A. (2011). Generalizing across cultural borders: Policy and the limits of educational science. *Educational Researcher*, *40*, 367–377. doi:10.3102/0013189X11424314

Maxwell, L. (2013, June 26). PARCC approves testing policies for English-language learners. Learning the Language Blog. *Education Week*.

McNamara, T., & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics*, *18*, 89–95. doi:10.1111/j.1473-4192.2008.00191.x

Menken, K. (2008). *English learners left behind: Standardized testing as language policy*. Clevedon, England: Multilingual Matters.

Menken, K. (2009). "No child left behind" and its effects on language policy. *Annual Review of Applied Linguistics*, *29*, 103–117. doi:10.1017/S0267190509090096

Menken, K. (2013). Emergent bilingual students in secondary school: Along the academic language and literacy continuum. *Language Teaching*, *46*, 438–476. doi:10.1017/S0261444813000281

Menken, K., & Solorza, C. (2014). No child left bilingual: Accountability and the elimination of bilingual education programs in New York City schools. *Educational Policy*, *28*, 96–125. doi:10.1177/0895904812468228

Messick, S. (1980). Test validation and the ethics of assessment. *American Psychologist*, *35*, 1012–1027. doi:10.1037/0003-066X.35.11.1012

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.

Ministry of Education, Singapore. (2001). *English language syllabus 2001 for primary and secondary schools*. Retrieved from http://moe.gov.sg/education/syllabuses/english-language-and-literature/

Mislevy, R. J., & Durán, R. P. (2014). A sociocognitive perspective on assessing EL students in the age of Common Core and Next Generation Science Standards. *TESOL Quarterly*, *48*(3), 560–585. doi:10.1002/tesq.177

Mislevy, R., Almond, R., & Lukas, J. (2003). *A brief introduction to evidence-centered design.* Research Report RR-03-16. Princeton, NJ: Educational Testing Service.
National Center for Education Statistics. (2013). *English language learners.* Alexandria, VA: U.S. Department of Education. Retrieved from http://nces.ed.gov/programs/coe/indicator_cgf.asp
National Clearinghouse for English Language Acquisition (NCELA). (2011). *What languages do English learners speak?* NCELA fact sheet. Washington, DC: Author. Retrieved from http://www.ncela.gwu.edu/files/uploads/NCELAFactsheets/EL_Languages_2011.pdf
National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards.* Washington, DC: Author. Retrieved from http://www.corestandards.org/the-standards
New York State Education Department. (2011). *Raising the achievement of English language learners.* Memorandum from State Commissioner King to the P-12 Education Committee. Retrieved from http://www.regents.nysed.gov/meetings/2010Meetings/November2010/1110p12d3.pdf
New York State Education Department. (2013). Graduation rates: Students who started 9th grade in 2004, 2005, 2006, 2007, and 2008 supplemental packet. Albany, NY: Author. Retrieved from http://www.p12.nysed.gov/irs/pressRelease/20130617/home.html
Next Generation Science Standards (NGSS) Lead States. (2013). *Next Generation Science Standards: For states, by states.* Washington, DC: The National Academies Press.
No Child Left Behind (NCLB) Act. (2001). Pub. L. No 107-110 Sec. 1111 (b)(F). Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf
O'Connor, R., Abedi, J., & Tung, S. (2012). *A descriptive analysis of enrollment and achievement among English language learner students in Delaware* (Issues and Answers Report, REL 2012–No. 132). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs
Organisation for Economic Co-operation and Development (OECD), Programme for International Students Assessment (PISA). (2006). *Where immigrant students succeed: A comparative review of performance and engagement in PISA 2003.* Paris, France: OECD.
Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.
Popham, W. J. (2001). *The truth about testing: An educator's call to action.* Alexandria, VA: Association for Supervision and Curriculum Development.
Roberts, G. (2012, February 26). Queens parents demand answers following teacher's low grades. *New York Post.*
Rumberger, R., & Gándara, P. (2005). How well are California's English learners mastering English? *University of California Linguistic Minority Research Institute Newsletter, 14*(2), 1–2.
Schissel, J. (2010). Critical issues surrounding test accommodations: A language planning and policy perspective.*Working Papers in Educational Linguistics, 25* (1),17–35.
Shepard, L. A. (2010). What the marketplace has brought us: Item-by-item teaching with little instructional insight. *Peabody Journal of Education, 85*, 246–257. doi:10.1080/01619561003708445
Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests.* London, England: Longman/Pearson Education.

Smarter Balanced Assessment Consortium (SBAC) (2012). *Smarter Balanced and PARCC to launch new technology readiness tool to support transition to online assessments*. San Francisco, CA: Author.

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

Thompson, S. (2001). The authentic standards movement and its evil twin. *Phi Delta Kappan, 82*, 358–362.

U.S. Census Bureau. (2013). *State and county quickfacts*. Washington, DC: Author. Retrieved from http://quickfacts.census.gov/qfd/states/00000.html

U.S. Department of Education. (1994). *The improving America's schools act of 1994: Summary sheets*. Washington, DC: Author.

U.S. Department of Education. (2013a). *Race to the top technical review*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop-assessment/performance.html

U.S. Department of Education. (2013b). Consolidated state performance report, 2011–12. Washington, DC: Author. Retrieved from http://www2.ed.gov/admins/lead/account/consolidated/index.html

Valdés, G. (2004). Between support and marginalisation: The development of academic language in linguistic minority children. *International Journal of Bilingual Education and Bilingualism, 7*(2–3), 102–132. doi:10.1080/13670050408667804

von Ahn, M., Lupton, R., Greenwood, C., & Wiggins, D. (2010). *Languages, ethnicity and education in London*. London, England: Institute of Education.

Wilds, C. P. (1975). The oral interview test. In R. L. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp. 29–44). Arlington, VA: Center for Applied Linguistics.

Wolf, M., Kao, J., Herman, J., Bachman, L., Bailey, A., & Bachman, P., . . . Chang, S. (2008). *Issues in assessing English language learners: English language proficiency measures and accommodation uses: Literature Review*. CRESST Report 731. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation, University of California, Los Angeles.

World-Class Instructional Design and Assessment (WIDA). (2007). *English language development (ELD) standards*. Retrieved from http://www.wida.us/standards/eld.aspx#2007

World-Class Instructional Design and Assessment (WIDA). (2013). *Essential actions: A handbook for implementing WIDA's framework for English language development standards*. Retrieved from http://www.wida.us/standards/eld.aspx