Understanding Language | Language, Literacy, and Learning in the Content Areas

Issues and Opportunities in Improving the Quality of Large Scale Assessment Systems for English Language Learners

Jamal Abedi, *University of California, Davis*
Robert Linquanti, *WestEd*

Large-scale academic content assessments primarily developed for and field tested on native speakers of English and those proficient in academic English may not produce reliable and valid outcomes for English Learner (ELL) students due to several extraneous factors. Key among these factors are ELL students' current level of English language proficiency (ELP); the unnecessary linguistic complexity of assessment items relative to the construct(s) being measured; and the validity of accommodations provided to improve accessibility of content-based assessments for ELLs. Next-generation assessment systems aligned to Common Core State Standards (CCSS) currently being developed by the two multistate Race-To-the-Top assessment consortia must engage these challenges and develop assessment systems that are accessible to ELLs.

This paper briefly summarizes some fundamental concepts in assessing ELLs, reviews issues that threaten the validity of interpretation of academic content assessments for ELL students, and provides recommendations on how to address such threats. It also highlights ELL-relevant assessment innovations on the horizon, and briefly discusses their promise and potential pitfalls. Finally, it suggests ways to strengthen connections between the academic assessment system development work of the PARCC and SBAC assessment consortia, and the work of next-generation ELP assessment developers and consortia, all with an eye to building a more coherent overall assessment system
for ELLs.

Fundamental Considerations

Unlike all other subgroup memberships for students, English Learner as a status is meant to be *temporary*, and ELLs are expected to leave the category as a result of effective, specialized language instruction and academic support services that they are legally required to receive.[i] ELL status is operationalized typically using both linguistic and academic performance standards, so the most linguistically and academically accomplished students exit the ELL category over time, while those not making sufficient progress remain and are joined by newly-entering ELLs, who are by definition at low ELP levels (Kim-Wolf et al., 2008; National Research Council, 2011; Working Group on ELL Policy, 2010). While assessment and accountability systems usually treat the ELL category as binary (a student is ELL or not), ELLs are very diverse and exhibit a wide range of language and academic competencies, both in English and their primary language (Capps et al., 2005; Taylor, Stecher, O'Day, Naftel, & LeFlock, 2010).

An ELL's ELP level clearly affects her ability to learn academic content in English and to demonstrate academic knowledge and skills on assessment events carried out in English. For most English Learners to learn academic English skills they need to effectively handle grade-level content demands, it takes 4 to 7 years depending upon several factors including their initial English language proficiency, age/grade on entry, and prior educational experiences (Cook,

Linquanti, Chinen & Jung, 2012; Hakuta, Butler, & Witt, 2000; Linquanti and George, 2007). Therefore academic assessments that fail to take account of ELLs' English language proficiency will likely inadequately measure their content area knowledge and skills.[ii] If an ELL student performs poorly on a content assessment, educators and policymakers need to better understand whether this is due to: 1) insufficient English language proficiency to demonstrate content knowledge; 2) a lack of content knowledge or opportunity to learn content; 3) construct-irrelevant interference (e.g., unnecessarily complex language in the assessment); or 4) other sources of bias or error (e.g., cultural distance, rater misinterpretation).

Language Demands of CCSS and the Role of Comprehensive Assessment Systems

The CCSS specify to an unprecedented degree the kinds of academic language competencies that students need in order to perform content area tasks and demonstrate subject matter mastery. In addition to explicitly defining K-12 listening, speaking, reading, and writing standards in English Language Arts (ELA), the CCSS in ELA also define literacy standards for history/social studies, science, and technical subjects at the secondary level. Across these different content areas and including mathematics, all students will now be expected to engage with more complex texts and to carry out more language-rich tasks (e.g., obtaining information, demonstrating understanding, constructing explanations, engaging in arguments, etc.) in discipline-appropriate ways during both learning and assessment situations.

As states adopt and implement the CCSS, many are also collaborating in ELP assessment consortia to revise their existing ELP standards to better correspond to the academic language demands reflected in the CCSS. The greater language-explicitness of the CCSS creates opportunities to better signal both general and discipline-specific academic language uses that all teachers need to foster and all students master within given content areas (Wong-Fillmore and Snow, 2002). Indeed, language is integral to these next-generation content standards and some content standards may need to be assessed in part by measuring such language uses within the content assessment. Nevertheless, assessment developers still need to carefully distinguish what language is content-related (construct-relevant) in order to ensure that language that is unrelated to the focal construct (construct-irrelevant) is not confounded with the content being measured.

The comprehensive systems of both assessment consortia (SBAC and PARCC) need to strike a judicious balance among the three key dimensions of assessment: *summative assessments* of cognitively complex knowledge and behaviors for program review and accountability purposes; *interim benchmark assessments* at key intervals during the school year to predict outcomes and guide interventions; and *formative assessment* practices, processes and tools to directly inform, support, and enhance teacher pedagogy and student learning. While assessment developers usually focus the least attention and resources on the last of these, this form of assessment (*for* and *as* learning) is critically important to get right for ELLs because it is the most *instructionally relevant*. Indeed, formative assessment processes, when seen within a teaching and learning paradigm (versus a measurement paradigm), can be used productively by teachers with *all* students (Heritage, 2010). The inequitable distribution of instructional resources to appropriately support ELLs' learning and the substantial need for better preparation, coaching, and ongoing professional development of all teachers of ELLs, make it all the more important to develop ELL-relevant formative assessment processes and practices that can provide feedback and

guide next steps in teaching and learning for linguistic and academic growth (Gandara, Rumberger, Maxwell-Jolly, & Callahan, 2003; Taylor et al., 2010).[iii]

Formative assessment tools and practices can also be particularly useful in measuring progress in ELL students' academic English development. As suggested above, ELL students who are instructed and assessed in English need to advance toward a level of proficiency in English that allows them to increasingly participate in academic activities using English. Formative assessment outcomes, or other interim assessments of students' English proficiency used formatively, can also help teachers make more informed decisions about their ELL students' degree of readiness to participate in assessment events delivered in English, and about what accommodations may be appropriate to facilitate participation (see below).

A key concern about interim/benchmark assessments meriting note – especially if these are used summatively – involves the language demands that correspond to particular content and the timing of interim assessments measuring such content. Since ELLs' language competencies develop throughout the school year, differential opportunities to learn and demonstrate subject matter knowledge may occur *within* the school year.[iv] In particular, the outcomes of interim assessments administered earlier in the academic year may unfairly represent ELL student results if the language competencies needed to display such academic knowledge are targeted and learned later in the school year. It is therefore critical to determine key target language uses corresponding to the curricular material to be taught, and to ensure that ELLs receive language instruction that addresses these target uses in the time period covered by the interim academic assessment. This implies the need for careful articulation of language instruction and content instruction goals.

Issues in Assessing College and Career Readiness

Another issue challenging next-generation assessments is the conceptualization and measurement of "college and career readiness." While the CCSS's goal of identifying and fostering skills needed for success in college and career is laudable, creating tests to measure such skills poses serious content and psychometric challenges. Under current practice, college entrance examinations such as the SAT and ACT are often used as external validity criteria for predicting students' success in college from their academic performance in high school. Yet such assessments suffer from the same construct-irrelevant language complexity (discussed below) that clearly threatens their validity in predicting ELLs' college and career readiness. Unless such issues are systematically addressed through careful attention to the language used in new academic consortia assessments, biases against ELLs in the interpretation of their college and career readiness will likely be perpetuated.

The Impact of Construct-Irrelevant Factors on Assessment Outcomes for ELL Students

Assessment content and questions should address only the focal construct, or the construct the assessment claims to measure. However, many factors unrelated to the focal construct impact assessment outcomes. Some factors are considered random measurement error (e.g., error due to inconsistencies in scoring open-ended questions).[v] Other extraneous variables *systematically* affect measurement outcomes and their effects cannot be removed from assessment outcomes even with unlimited repeated measurements. Linguistic complexity that is

unrelated to the focal construct represents such an unnecessary, "construct-irrelevant" source of systematic measurement error and may seriously affect ELL measurement outcomes.[vi]

Research has clearly demonstrated the impact of language factors on the assessment of ELL students both on Title I and Title III assessments (Kopriva, 2008; Solano-Flores, 2006). Language that is unrelated, unnecessary, or irrelevant to the construct can cause ELLs difficulty in understanding and responding to assessment questions. Examples of language difficulty include unfamiliar vocabulary, complex grammatical structures, nominalization, multiple embedded clauses, and passive voice constructions. In reading/language arts and ELP assessments, language is so inherent to the focal construct that the concept of unnecessary linguistic complexity may not apply. (Even in these areas, excessive linguistic complexity can still be avoided.) However, in mathematics and science, test items may have complex linguistic structures unrelated to the focal construct that unnecessarily add to cognitive load and slow the reader down (Abedi, Lord, & Plummer, 1997; Abedi, 2006; Abedi, et al (in press); Kopriva, 2008; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006; Solano-Flores, 2006).

Distinguishing Language that is Relevant and Irrelevant to the Focal Construct: A Simple Example

Distinguishing language that is related versus unrelated to the construct poses serious challenges and requires both content and language experts to carefully develop and review test items and tasks and to determine such distinctions. Even in content areas that are not commonly understood as language-heavy in assessment events (e.g., mathematics and science), language plays an essential role both to set the context and to define the content. Based on this premise, decisions are made (explicitly or implicitly) about what language is necessary in assessing content and how to help students understand the content knowledge and skills being elicited, versus what language is irrelevant and causes the student unjustified burden and confusion.

In a study on the impact of language factors on the assessment of ELL students, Abedi and Lord (2001) compared the performance of Grade 8 students who received an original NAEP mathematics version of the test items with the performance of those students who received a linguistically-revised version of the items. The authors found significant improvements on the performance of ELL students taking the mathematics test version which was linguistically modified to reduce the level of unnecessary linguistic complexity. The revised version was prepared by a team of experts in such a way as to not alter or modify any content-related language. Before administration of the two test forms, two mathematics content experts independently compared the original and the linguistically modified versions to make sure the math content was not altered. Subsequent studies (e.g., Abedi, 2006) have confirmed these findings and suggest that reducing unnecessary linguistic complexity improves the validity of content-based assessments for ELL students.

To illustrate the process of linguistic modification that addresses language irrelevant to the focal construct, we present a grade 4 released mathematics test item prompt in its original form and a linguistically-revised form, and then elaborate on what linguistic modifications were conducted on the item.[1]

[1] Retrieved from the California Department of Education website at http://www.cde.ca.gov/ta/tg/sr/documents/cstrtqmath4.pdf. The authors thank Nancy Ewers for providing this example.

Original Prompt: A cookie factory can bake 62 trays of cookies in the morning and 53 trays of cookies in the afternoon. If each tray holds 12 cookies, how many cookies can be baked in 1 day?

Revised Prompt with reduced linguistic complexity: A bakery bakes 62 trays of bread in the morning and 53 trays in the afternoon. Each tray holds 12 loaves of bread. How many loaves did they bake in one day?

The linguistic complexity is reduced by making the following changes:

1. Replace *cookie factory* with the accurate word for such a facility, *bakery.*
2. Replace *cookies* with the more cross-culturally familiar word *bread.*
3. Replace the modal verb phrase *can bake* with simple present tense.
4. Replace the subordinate clause *if each tray holds 12 cookies*, with a simple declarative sentence.
5. Eliminate the modal with passive voice *can be baked* with a past tense interrogative sentence.

A comparison of the linguistic structure of the two versions of this item reveals that the modifications target unnecessary linguistic complexity and do not alter any language related to the content. This makes the reduced linguistic complexity math item potentially more accessible to ELL students.

Leveraging Accommodations for Incremental Improvement: What Have We Learned?

Federal law requires that ELLs be provided with accommodations to make assessments more accessible to them. However, accommodations used by ELLs must be *effective* in improving accessibility and must also be *valid* (demonstrate an absence of advantage for non-ELLs). Decisions on the number and type of accommodations to be used with ELLs are left to each state (Kieffer, Lesaux, Rivera & Francis, 2009; Shafer Willner, Rivera & Acosta, 2008; Solano-Flores, 2006; Young et al., 2008).[vii]

Among the most important criteria for selecting appropriate accommodations for ELL students is relevance to specific need. Unlike students with disabilities with different needs, ELLs share a common need for assistance with the language of assessments. Research evidence supports the value of accommodations offering direct and indirect linguistic support *when appropriately tailored to ELL characteristics and testing conditions* (Pennock-Roman & Rivera, 2011; Shafer Willner et al., 2008). For example, English customized dictionaries/ glossaries that clarify key vocabulary not directly tied to the construct being measured are effective in paper-and-pencil versions when ELLs are given extra time. Pop-up English glossaries (via computer-based delivery) are more helpful under restricted time constraints. Students instructed bilingually may need to be accommodated based on their ELP level as well as on the goals of instructional program. For example, primary-language (L1) versions of test items are promising for L1 speakers at low-ELP levels and those being instructed in L1 while learning English. This accommodation is less effective for ELLs at intermediate ELP levels and those receiving content

instruction in English. Dual-language formats (parallel bilingual versions and bilingual glossaries) also show promise if there are generous time limits to help students with differing capacities in both languages. Finally, plain-English accommodations, though promising, have yielded mixed results with effectiveness, though not with validity when used in rigorous experimental research designs (Duran, 2008; Kieffer et al., 2009).[viii] Since the CCSS calls forth many academic language skills that are inextricably related to more complex content knowledge and are central to many focal constructs, such necessary language complexity may be ineligible for simplification.

Clearly, not all accommodations are appropriate for all ELLs since ELLs are heterogeneous in ways that can measurably influence the effectiveness of particular accommodations. Using a decision algorithm to assign configurations of accommodations tailored to ELLs' linguistic and socio-cultural characteristics shows promise in yielding better performance outcomes than providing all available accommodations or no accommodations (Kopriva, Emick, Hipolito-Delgado & Cameron, 2007). While such an algorithm depends on the capacity of states and districts to collect relevant data at the student level, new data systems coming online in states and districts may soon make this a
feasible option.

Computer Adaptive Tests, Automated Scoring, & Language-Minimizing Accommodations

Several innovations on the horizon hold promise for improving the assessment system's responsiveness to ELLs but these must be pursued carefully. *Computer adaptive testing*, or online testing formats presenting students with test questions of a level of difficulty continually adjusted based on how the student has answered previous questions, may more accurately estimate ELLs' content knowledge while also increasing testing efficiency and reducing stigma and demoralization (see Reckase, 2011). In order for CAT to work for ELLs, the optimization algorithms that assign test items must be sensitive to an ELL's ELP level (particularly in literacy domains) so that items of equivalent construct difficulty, but with differing levels of linguistic complexity, can be assigned to ELLs of different ELP levels. Given that both consortia are expected to develop tens of thousands of items, it will be crucial to categorize the language demands inherent in test items and tasks by ELP performance level descriptors, even if such benchmarking and anchoring must be done in terms broad enough to be comparable across different ELP assessments (e.g., beginning, intermediate, advanced). Likewise, *automated scoring routines* that enable computerized scoring of short essays and constructed responses may need to be specially programmed to recognize common "trans-language" features of ELL writing. This would require training such artificial intelligence engines with exemplars that include grammatical, vocabulary, and discourse features of ELLs at various stages of second language acquisition, in order to provide a more careful analysis and meaningful judgment of performance by ELP level.

Finally, substantial work is being done using online formats that increase access by conveying information to and receiving information from ELLs at the lowest ELP levels (Kopriva, 2011). Multi-semiotic approaches in particular appear promising in accessing conceptual and procedural knowledge in science and math of ELLs at the lowest ELP levels.[ix] Such methods can provide more accurate and valid information on these ELLs and also signal to educators that ELLs at all stages of language development can learn and be assessed in the academic content. However, these "language-minimizing" accommodations must be understood and

utilized as temporary strategies to measure students' knowledge while students develop language competencies required by CCSS. If not, such efforts could unintentionally suggest to teachers that ELLs' language development is not essential to learning and demonstrating academic content knowledge, and contribute to their instructional marginalization.

Implications for Moving Forward

As the nation implements more rigorous, language-rich academic content standards in English language arts, mathematics, and science and moves towards more comprehensive academic and ELP assessment systems, educators and assessment developers have a clear opportunity and obligation to ensure that the growing ELL student population has meaningful access to rigorous instruction and valid, useful assessments to measure language and content knowledge, skills, and abilities.

There have been intriguing suggestions that, just as ELL students reaching a "threshold" level of English language proficiency must be effectively supported in developing their interpersonal, presentational, and interpretive uses of language by content area teachers (Walqui & Heritage, 2012), so too such discipline-specific social and academic language competencies delineated in next-generation standards might be called out, measured and reported for *all students* – English learners, standard English learners, and monolingual standard English speakers – *as part of* the Race To the Top academic content assessments. While it would respond to a steadily growing call to operationalize and measure academic language uses for all students, this approach also raises several significant challenges. These include ensuring the validity of the language competencies postulated within the content standards as integral to demonstrating mastery of subject matter content; avoiding unnecessary linguistic complexity in assessing necessary linguistic competencies; and aligning the assessment infrastructure to clearly and coherently articulate where the "threshold" is crossed from ESL/ELD precursor language progressions to language constructs found in the content standards. (See Bailey & Wolf [2012] for further discussion of ELP standards.)

Given the demonstrated impact of English language proficiency on ELLs' opportunity to learn and on their assessment outcomes, states and consortia first need to develop a coherent framework to ensure that the breadth, depth, and complexity of academic language uses reflected in CCSS are adequately captured in new ELP standards. Academic assessment consortia and ELP assessment developers should also strengthen communication, data collection and analysis, experimentation, and prototyping of ELP and academic assessment tasks to yield more aligned and coherent information across assessment systems. They can also invest heavily in formative assessment processes and practices, tools and tasks that carefully map key academic language competencies and target language uses, and ensure these language competencies are articulated in language learning progressions reflected in ELP standards and ELP assessment specifications. Moreover, academic assessment developers can use these ELP performance standards to evaluate the language demands of different content assessment items and tasks, which will be critical in the adaptive assignment of the tens of thousands of test items to ELLs of different ELP levels.

The PARCC and SBAC assessment consortia can also incorporate lessons from research on ELL assessment and accommodations into their test development processes. For example, they can: 1) Examine different interpretations of test scores by subgroups of students, including ELLs at different ELP levels, to identify possible threats to valid interpretation of assessment

outcomes; 2) Identify possible construct-irrelevant sources in assessment items and tasks by conducting cognitive labs and think-aloud procedures on ELLs at different ELP levels; 3) Have content, language, and assessment experts identify unnecessary linguistic complexity; 4) Distinguish linguistic structures that are related and unrelated to the focal construct; even construct-relevant language can be made more accessible for ELLs by, for example, avoiding long and complex reading comprehension passages when such complexities are not required by CCSS standards being measured; 5) Specify accommodations for ELLs based on student characteristics, testing conditions, and instructional services provided; avoiding accommodations that are ineffective or irrelevant for ELLs, or that alter the focal construct; and 6) Provide evidence to substantiate selection and delivery of accommodations for ELLs.

The two academic assessment consortia have access to the insights of rigorous, current research on ELL academic assessment. They also have colleagues working in parallel on next-generation ELP standards and assessments articulating the language demands of CCSS. Where gaps in knowledge exist, enormous opportunities also exist to collaboratively prototype, field test, study, and advance understanding of ELL-responsive assessment tasks and strategies. Such resources and opportunities can and must inform the development of next-generation assessment systems that are more accessible, valid, and useful to ELLs.

---

[i] English learners are language minority students not sufficiently proficient in English to be able to benefit adequately from regular mainstream instruction and demonstrate their knowledge and abilities using English.

[ii] Note that issues of measurement are distinct from issues of accountability. From a measurement perspective, knowing an ELL's ELP level (particularly with respect to literacy) is essential to judging the validity of the inferences from the assessment. With respect to educator accountability however, there may still be a rationale for including such results to determine school or district effectiveness, particularly if ELLs have not been supported to progress in their English-language proficiency over time (see Cook et al., 2012).

[iii] ELL-focused formative approaches are slowly evolving and include pilot academic content learning progressions and associated language learning targets; prototyped performance tasks and instructional supports linked to those tasks; and professional development models that systematically build teachers' capacities to evaluate ELLs' access to and accomplishment of language and content objectives that indicate progress toward larger instructional goals. See FLARE; WestEd's Quality Teaching for English Learners Program; Bailey & Heritage, 2010; Heritage, 2008.

[iv] See Wise (2011) for discussion of issues related to different aggregation methods of interim/benchmark assessment results.

[v] Random measurement error affects the observed score (X) but its impact reduces by averaging over repeated observation. The observed score (X) becomes closer to the true score (T) as the number of measurements increases. In classical test theory (Thorndike, 2005), the correlation between the true score and error score is assumed to be zero. For example, in scoring open-ended questions, some judges might be lenient and provide higher scores for everyone whereas some other raters may be less generous in their rating and assign lower ratings to everyone. Averaging over a number of ratings from the two groups of raters should control for this source of measurement error. Accordingly, in a generalizability approach (Shavelson & Webb, 1991), the number of levels within each facet of measurement is increased in order to reduce measurement error and improve score dependability.

[vi] Test items that are culturally biased may also pose difficulty for test takers with different cultural backgrounds. Such difficulties could systematically distort the assessment outcomes and reduce students' test scores significantly.

[vii] For a more elaborated summary of this and the following section, see Linquanti (2011).

[viii] Possible explanations include heightened sensitivity of test developers in recent years to avoid unnecessarily linguistically-complex items, which may reduce gains produced by plain-English versions; and the presence of comparison groups of non-ELLs that include former ELLs with ongoing linguistic needs who also benefit from the accommodation, which may affect comparison statistics.

[ix] See www.onpar.us for examples of this strategy.

References

Abedi, J. (2006). Language Issues in Item-Development. In Downing, S. M. and Haladyna, T. M. (Ed.), *Handbook of Test Development* (pp. 377-398). New Jersey: Lawrence Erlbaum Associates, Publishers.

Abedi, J.; Bayley, R.; Ewers, N.; Mundhenk, K. Leon, S. & Kao, J. (in press). Accessible Reading Assessments for Students with Disabilities. *International Journal of Disability, Development and Education, 59*(1).

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219-234.

Abedi, J., Lord, C., & Plummer, J. (1997). Language background as a variable in NAEP mathematics performance (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Bailey, A. & Heritage, M. (2010). *English-language proficiency assessment foundations: External judgments of adequacy*. Evaluating the validity of English-language proficiency assessments. (An enhanced assessment grant). Washington DC: EVEA Project.

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

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: US Department of Education, Office of Planning, Evaluation and Policy Development.

Duran, R. (2008). Assessing English-language learners' achievement. *Review of Research in Education, 32*, 292-327

FLARE (Formative Language Assessment Records for English-Language Learners) project. http://flareassessment.org

Gándara, P., Rumberger, R., Maxwell-Jolly, J., & Callahan, R. (2003). English learners in California Schools: Unequal Resources; Unequal Outcomes. *Educational Policy Analysis Archives*. http://.epaa.asu.edu/epaa/v11n36/

Hakuta, K., Butler, Y., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* University of California Linguistic Minority Research Institute Policy Report 2000-1. Santa Barbara: UC-LMRI.

Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* (p. 9). Washington DC: Council of Chief State School Officers.

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington DC: Council of Chief State School Officers.

Kieffer, M., Lesaux, N., Rivera, M. & Francis, D. (2009). Accommodations for English-language learners taking large-scale assessments: A meta-analyses on effectiveness and validity. *Review of Educational Research, 79*(3), 1168-1201.

Kopriva, R., (2011, April 9). The promise of demonstration-based interactive test task environments for struggling readers and English learners. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Kopriva R. (Ed.). (2008). *Improving testing for English Language learners*. New York: Routledge Press.

Kopriva, R., Emick, J., Hipolito-Delgado, C., & Cameron, C. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice, 26*(3), 11-20.

Linquanti, R. (2011). Strengthening assessment for English learner success: How can the promise of the common core state standards and innovative assessment systems be realized? In D. Plank and J. Norton (Eds.), *The road ahead for state assessments* (pp. 13-25). Palo Alto & Cambridge: Policy Analysis for California Education and Rennie Center for Education Research & Policy.

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

National Research Council (2011). Allocating federal funds for state programs for English-language learners. Panel to review alternative sources for the limited English proficiency allocation formula under Title III, Part A., *Elementary and Secondary Education Act*, Committee on National Statistics and Board Testing and Assessment. Washington DC: National Academies Press.

Pennock-Roman, M. and Rivera, C. (2011), Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice*, 30: 10–28.

Reckase, M.D. (2011). Computerized adaptive assessment (CAA): The way forward. In D. Plank and J. Norton (Eds.), *The road ahead for state assessments* (pp. 1-11). Palo Alto & Cambridge: Policy Analysis for California Education and Rennie Center for Education Research & Policy.

Shafer Willner, L., Rivera, C., & Acosta, B. (2008). Descriptive study of state assessment policies for accommodating English language learners. Prepared for the LEP Partnership, U.S. Department of Education. Arlington, VA: The George Washington University Center for Equity and Excellence in Education.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105-126.

Shavelson & Webb (1991). *Generalizability theory, a primer.* NewBury Park: Sage Publications.

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *Teachers College Record (108)*11, pp. 2354-2379.

Taylor, J., Stecher, R., O'Day, J., Naftel, S. & LeFloch, K.C. (2010). *State and Local Implementation of the No Child Left Behind Act, Volume IX – Accountability under NCLB: Final Report.* Washington, DC: U.S. Department of Education.

Thorndike, R. M. (2005). *Measurement and Evaluation in Psychology and Education*. Upper Saddle River, NJ: Pearson, Merrill.

WestEd (2011). Quality teaching for English learners program. http://www.wested.org/cs/tqip/print/docs/qt/home.htm

Wise, L. (2011). *Picking up the pieces: Aggregating results from through-course assessments*. Princeton: Center for K-12 Assessment & Performance Management at ETS.

Wolf, M. K., Herman, J. L., Kim, J., Abedi, J., Leon, S., Griffin, N., Shin, H. (2008). *Providing Validity Evidence to Improve the Assessment of English Language Learners. (CSE Tech. Report No. 738)*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Wong-Fillmore, L., & Snow, C. E. (2002). What teachers need to know about language. In C. T. Adger, C. E. Snow, & D. Christian (Eds.), *What teachers need to know about language* (pp. 7-54). Washington DC: Center for Applied Linguistics.

Working Group on ELL Policy (2010). *Improving educational outcomes for English language learners: Recommendations for ESEA reauthorization*. Palo Alto: Author. Available at: http://ellpolicy.org

Young, J. W., & King, T. C. (2008). *Testing accommodations for English language learners: A review of state and district policies*: New York: The College Board. Report # 2008-6.

The Understanding Language Initiative would like to thank the Carnegie Corporation of New York and the Bill and Melinda Gates Foundation for making this work possible. For more information about this paper, please contact **UnderstandingLanguage@stanford.edu**

Understanding Language
Stanford University School of Education
485 Lasuen Mall
Stanford, CA 94305-3096
ell.stanford.edu