Assessment of Young English Language Learners in Arizona:

Questioning the Validity of the State Measure of English Proficiency

Eugene E. Garcia, Kerry Lawton and Eduardo H. Diniz de Figueiredo
Arizona State University

July 2010

The Civil Rights Project
Proyecto Derechos Civiles

Abstract

This study analyzes the Arizona policy of utilizing a single assessment of English proficiency to determine if students should be exited from the ELL program, which is ostensibly designed to make it possible for them to succeed in the mainstream classroom without any further language support. The study examines the predictive validity of this assessment instrument on ELLs' performance on state required academic achievement tests at three grade levels. It finds that at subsequent grade levels after redesignation, the "one-test" AZELLA becomes less predictive of academic achievement, That is, the test over predicts student achievement, suggesting that many students may be under-served due to their scores the test. .This finding calls into question Arizona's "one-test" procedure for redesignating ELLs to a non-service category. Given the large and increasing size of the ELL student population in Arizona, the current focus on testing and accountability, and the documented problems in current assessment practices, improvement in instruments and procedures is critical. These improvements are necessary at all phases of the assessment process, but as this study indicates, the present policy is likely denying services these student need and violating the rights of these students to an equal educational opportunity.

ASSESSMENT OF YOUNG ENGLISH LANGUAGE LEARNERS IN ARIZONA: QUESTIONING THE VALIDITY OF THE STAE MEASURE OF ENGLISH PROFICIENCY

Introduction

The Challenge of Assessing ELLs

The increasing demand for evaluation, assessment, and accountability in education comes at a time when the fastest growing student population in Arizona is children whose home language is not English. This presents several challenges to practitioners and school systems generally when they lack familiarity with important concepts such as second language acquisition, acculturation, and the role of socioeconomic background as they relate to test development, administration, and interpretation. Because assessment is key in developing and implementing effective curricular and instructional services that are required to promote student learning, English language learner (ELL) children have the right to be assessed to determine their educational needs. Through individual assessments, teachers can personalize instruction, make adjustments to classroom activities, assign children to appropriate program placements, and have more informed communication with parents. They can also identify learning problems that may require additional outside assistance. And educational systems need to know how ELLs are performing in order to make proper adjustments to their programs and to effect necessary policy changes. Notwithstanding the increasing need, states have struggled to develop strong assessment programs with appropriate instruments for use with young ELLs.

Although hundreds of languages are represented in schools in the United States, Spanish is the most common; nationally almost 80% of all ELL students speak Spanish as their first language (Gándara & Rumberger, 2009); in Arizona the percentage is even higher – closer to 85% (xxx), and while some Spanish language tests exist, most lack the technical qualities required of high-quality assessment tools, or the specifications to serve as the accountability purposes of NCLB.. Additionally, there is a shortage of bilingual professionals with the skills necessary to evaluate these children, (Arias, 2009). The intent of this article is to describe the challenges inherent in assessing young English language learners, to review important principles associated with the development of such assessment, and to present an empirical analysis that indicates that Arizona's current assessment instrument does not yield valid inferences about ELLs' readiness to undertake English only instruction.

Young English Language Learners: Who Are They?

Several terms are used in the literature to describe children from diverse language backgrounds in the United States. A general term describing children whose native language is other than English, the mainstream societal language in the US, is *language minority*. This term is applied to non-native English speakers regardless of their current level of English proficiency. Other common terms are *English language learner* (ELL), English learner (EL), and *limited English proficient* (LEP). These terms are used interchangeably to refer to students whose native language is other than English, and whose English proficiency is not yet developed to a point where they can profit fully from English instruction or communication. In this article, the term *English language learner* and its respective abbreviation is preferred as it places an emphasis on students' learning and progress rather than their limitations.

Young ELLs (generally considered to be between 0 – 8 years) have been the fastest growing student population in the country over the past few decades, due primarily to increased rates in immigration. Currently, one in four school-age children in Arizona has a foreign-born parent (Capps et al., 2005), and many of these children learn English as a second language, though not all. Overall, the population of children speaking a non-English native language in Arizona rose from 16 percent in 1979 to 27 percent in 1999 (NCELA, 2006) and the number of language minority students in K-12 schools has recently been estimated to be over 120,000 (August, 2006).

Assessing the development of ELLs demands an understanding of who these children are in terms of their linguistic and cognitive development, as well as the social and cultural contexts in which they are raised. The key distinguishing feature of these children is their non-English language background. In addition to linguistic background, other important attributes of ELL children include their ethnic, immigrant, and socioeconomic histories (Abedi, Hofstetter, & Lord, 2004; Capps et al., 2005; Figueroa & Hernandez, 2000; Hernandez, 2006). Though diverse in their origins, ELL students, on average, are more likely than their native English-speaking peers to have an immigrant parent, to live in low-income families, and to be raised in cultural contexts that do not reflect mainstream norms in the US (Capps et al., 2005; Hernandez, 2006).

Decades of research support the notion that children can competently acquire two or more languages (Garcia, 2005). Relationships of linguistic properties between languages are complex, and several theories have been presented over the years to explain how language develops for young bilingual children. Among the major theoretical approaches, available empirical evidence suggests that transfer theory best explains the language development of young children managing two or more languages (Genesee, Geva, Dressler, & Kamil, 2006). This theoretical position asserts that certain linguistic skills from the native language transfer to the second. In like manner, errors or interference in second language production occurs when grammatical differences between the two languages are present. Language that is contextually-embedded and cognitively undemanding—or automatic, over-learned communication—does not lend itself well to transfer. Contextually-reduced and cognitively demanding language skills, on the other hand, tend to transfer more easily between languages. Higher order cognitive skills relevant to academic content are more developmentally interdependent and, therefore, amenable to transfer (Genesee, Geva, Dressler, & Kamil, 2006). In the process of cross-linguistic transfer, it is normal for children to mix (or "code-switch") between languages. Mixing vocabulary, syntax, phonology, morphology, and pragmatic rules serves as a way for young bilingual children to enhance meaning. Because language use is context-driven, the bilingual child's choice of language depends on characteristics of and the particular relationship with the addressee(s) as well as the cultural identity and attitudinal features of the child, and overall comfort with the language.

Assessment Issues

Although many young ELLs have immigrant parents or caregivers, the vast majority of these students are native born US citizens and have been legally granted the same rights to education as their native English-speaking peers. Benefiting from valid educational assessment is included in these rights. While the current knowledge base and legal and ethical standards governing ELL assessment are limited, they are sufficient to provide guidance for the development of appropriate and valid assessment. Making improvements on existing assessments will require commitments from policymakers and practitioners to develop and implement appropriate assessment tools and procedures, to link assessment results to improved practices, and to utilize trained staff capable of carrying out these tasks. Researchers and scholars can facilitate the improvement of assessment practices by continuing to evaluate implementation strategies in schools, and by developing systematic assessments of contextual factors relevant to linguistic and cognitive development. Assessments of contextual processes will be necessary if current assessment strategies, which largely focus on the individual, are to improve classroom instruction, curricular content, and, therefore, student learning (Rueda, 2007; Rueda & Yaden, 2006).

Reasons to assess

Several skills and developmental abilities of young children are assessed in early education programs, including preschool and the first few elementary school years. Sensing an increase in demand for greater accountability and enhanced educational performance of young children, the National Education Goals Panel developed a list of principles to guide early educators through appropriate and scientifically-sound assessment practices (Shepard, Kagan, & Wurtz, 1998). Moreover, the panel presented four purposes for assessing young children. Pertinent as well to the assessment of young ELL children, the purposes were a) to promote children's learning and development, b) to identify children for health and special services, c) to monitor trends and evaluate programs and services, and d) to assess academic achievement to hold individual students, teachers, and schools accountable (i.e., high stakes testing) (Committee on the Foundations of Assessment, 2001; National Research Council, 2008; Shepard, Kagan, & Wurtz, 1998). Embedded within each of these purposes are important considerations for practice so as to preserve assessment accuracy and support interpretations of results that lead to increased educational opportunity for the student.

Legal and ethical precedent

The impetus for appropriate and responsive assessment practices of young ELLs is supported by a number of legal requirements and ethical guidelines, which have developed over time. Case law, public law, and ethical codes from professional organizations support the use of sound assessment tools, practices, and test interpretations. A widely cited set of testing standards are found in a recent publication from the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME) entitled *Standards for Educational and Psychological Testing* (1999). Revised from the 1985 version, in its fourth edition, this volume offers a number of ethical standards for assessing the psychological and educational development of children in schools, including guidelines on test development and application. Included is a chapter on testing children from diverse linguistic backgrounds, which discusses the irrelevance of many psychoeducational tests developed for and normed with monolingual, English-speaking children. Caution is given to parties involved in translating such tests without evaluating construct and content validity and developing norms with new and relevant samples. It also discusses accommodation recommendations, linguistic and cultural factors important in testing, and important attributes of the tester. Similar, though less detailed provisions exist in the *Professional Conduct Manual* published by the National Association of School Psychologists (2000).

It has been argued that the standards presented by APA, AERA and NCME have outpaced present policy, practice, and test development (Figueroa & Hernandez, 2000). However, the federal Individuals with Disabilities Education Act (IDEA 2004) does provide particular requirements related to the assessment of ELLs. It requires, for example, the involvement of parents/guardians in the assessment process as well as a consideration of the child's native language in assessment. Unlike ethical guidelines, which often represent professional aspirations and are not necessarily enforceable, public law requires compliance. The Office for Civil Right (OCR) is given the charge to evaluate compliance to federal law and, where necessary, audit public programs engaged in assessment practices and interpretations of test performance by ELLs and other minority children.

Assessment practice: use and misuse

Several domains of development are assessed during the early childhood years. These include cognitive (or intellectual), linguistic, socioemotional (or behavioral), motor, and adaptive (or daily living) skills, as well as hearing, vision, and health factors. Educational settings are primarily concerned, however, with the cognitive, academic, and linguistic development of children. Other domains are of interest insofar as they impact students' educational well-being, as stated in IDEA (2004). This section focuses primarily on these areas not because others are irrelevant, but because they are given the most emphasis and importance in schools. Developing appropriate assessment measures and practices, however, transcends developmental domains and is considered important for the assessment of culturally and linguistically diverse children in general.

In addition to the concerns that attend the assessment of all children, there are central issues inherent in the assessment of young children from non-English language backgrounds. Implementation research suggests that assessment practices with young ELLs continue to lag behind established legal requirements and ethical standards set forth by APA, AERA and NCME. In part, this is because of a lack of available instruments normed on representative samples of English language learners, because of inadequate professional development and training, and partly because of insufficient research to inform best practice. Such is the case for the assessment of language, cognitive skills, academic achievement, and other areas. Each of these areas is visited briefly.

English Proficiency Assessment in Arizona

Language is the key distinguishing feature for ELLs. Therefore, assessments of language in early and elementary school settings are used to determine oral English proficiency, to determine first- and second-language vocabulary skills, to predict literacy performance , and ultimately to identify and place students into programs (including special education) (Garcia, McKoon, & August, 2006). Prior to the 2004-2005 school year, Arizona's procedures for reclassifying English language learners (ELLs) to non-ELL (or FEP) status were based on multiple measures related to student language proficiency and academic achievement. In that year, however, Arizona adopted the Stanford English Language Proficiency Test (SELP[1]), a measure that provides an indication of language proficiency but not academic attainment, and began using it as the sole criterion to reclassify ELLs to English proficient status.

Mahoney, Haladyna and MacSwan (2009) investigated the appropriateness of relying on a single measure to reclassify English learners (ELs) to non-EL (or FEP) status in Arizona. According to these researchers, Arizona's change in reclassification procedures was problematic, in that it disregarded the view shared by AERA, APA, and NCME and others that that relying on a single testing measure for high stakes educational decisions is inappropriate. Moreover, citing observations of several state teachers and administrators, the researchers suggest Arizona's new single test re-classification procedure removes many ELLs from language services before they are ready to succeed in mainstream classes.

Therefore, with the objective of examining the effectiveness of the SELP as a reclassification tool, Mahoney and colleagues proposed two research questions. The first one sought to answer whether SELP-reclassified ELs develop the necessary English language skills to be successful in the English language curriculum. The second question focused on how the SELP differed from tests that had been previously used for reclassification purposes in the state.

In order to answer their first question, Mahoney and colleagues analyzed the performance of SELP-reclassified students from a Phoenix metropolitan area school district in grades 3-8 in the Arizona's Instrument to Measure Standards test (AIMS) in 2005. Results from this group were statistically compared to those of a control group, which consisted of students who had been reclassified in 2004 through multiple measures (in this case, the use of the Woodcock-Muñoz test together with the Reading Comprehension subtest of the Stanford Achievement Test – SAT-9). Results showed that the control group outperformed the SELP-reclassified students on all parts of the AIMS, and the percentage of reclassified students was higher in 2005 than in 2004. The research team interpreted these results as evidence of premature reclassification by the SELP, a fact that could jeopardize reclassified ELs' performance in mainstream classrooms.

In order to answer their second question (how the SELP differs from tests that were previously used for reclassification purposes in the state), Mahoney and colleagues compared the consistency of pass/fail decisions and the passing rates of the SELP to those of the Language Assessment Scales test (LAS), which was one of the tests that had been used in previous years in Arizona. Both tests were administered to a group of 288 students from one Phoenix metropolitan area school within a short period of time. Results showed that 17% of the students were not classified consistently by the two tests, and the SELP passing rate was statistically higher than that of the LAS.

Mahoney, Haladyna and MacSwan's conclusion was that that SELP is probably over-reclassifying ELs into FEP status, as teachers and administrators had already perceived. They emphasize the need for reclassification tools that rely on multiple measures, as recommended by the measurement community, and suggest that states must adopt procedures that have a language component as well as an academic achievement indicator for reclassification decisions. The present study extends the work by Mahoney, et. al., (2009), with the goal of evaluating the strength of the relationship between AZELLA (the test that replaced the SELP) and AIMS subtest scores for students identified as English Language Learners in Arizona.

Methods

This study examined the relationship between performance on the AZELLA and performance on the state's NCLB-mandated standardized achievement test –the AIMS, in a sample of ELL students. The study sought to answer whether the relationship between the two tests is consistent across grade levels.

Participants

All participants attended elementary or middle schools within a mid-size, urban school district in the Southwest United States. During the 2008-2009 school year, this district served approximately 8,500 students who were enrolled in 17 schools. Archival performance data for participants were provided by the school district and there was no direct contact between the researchers and the participants in this study.

Only children who were administered and received valid scores on the Arizona Instrument to Measure Standards (AIMS) and the Arizona English Language Learner Assessment (AZELLA) during the 2008-2009 school year were included in the study. For each participant, the anonymous identification number assigned by the district was used to match student performance on the two tests. If participants were administered the AZELLA more than once, only scores for the first administration were included. Additional sampling criteria were: (a) enrollment in the third, fifth, or eighth grades and (b) English Language Learner classification. These criteria yielded a sample of 710 students (see Table 1). Of the participants, 378 (53%) of the participants were male and 332 (47%) were female. Ninety-three percent of the participants were identified as Hispanic and Spanish was the primary language spoken (84%).

Table 1. *Sample size by grade level*

Grade Level	*n*
3	349
5	187
8	174

Instruments

Arizona Instrument to Measure Standards (AIMS). AIMS is a standardized achievement measure designed to assess student performance in three academic categories: mathematics, reading, and writing (ADE: Azella Technical Manual). Reliability of the 2009 AIMS reading and math subtests was estimated with Cronbach's measure of internal consistency. For English-language learners in the grades targeted in this study, Alpha coefficients ranged from .82 to .91. Internal consistency was generally higher for mathematics than reading, and higher for lower grades than higher grades.

AIMS tests contain embedded items from the Terranova making it possible to derive both criterion-referenced (AIMS CRT) and norm-referenced (AIMS NRT) scale scores. As the AIMS CRT and NRT do not contain the same items, inter-correlations between the two forms were provided as evidence for construct validity. The developers report high correlations between the two forms when assessing the same construct and lower correlations among dissimilar constructs (e.g., Reading and Mathematics). For the current study, AIMS CRT scores on the reading and mathematics subtests were analyzed.

Arizona English Language Learner Assessment (AZELLA). AZELLA is a criterion-referenced test used by the state of Arizona to assess English proficiency for the purposes of determining whether students receive ELL services. Developed alongside Arizona's K-12 English Language Proficiency standards, AZELLA was intended to augment the Stanford English Language Proficiency (SELP) test. The technical manual estimates alignment to state standards to be 85%. Depending on grade level, several forms of AZELLA are administered. The Elementary form is used for students in grades 3, 4, and 5. The Middle Grades form is administered to students in grades 6, 7, and 8. Both tests contain similar item types (i.e., multiple-choice; extended response) and yield scores on four subtests: Speaking, Listening, Reading, and Writing. Subtest scores are combined to form a Total Composite score. Evidence for the reliability of AZELLA is provided with Cronbach's Alpha. Coefficients for targeted grades were high, ranging from .90 to .97. Inter-correlations among subtest scores were rational, providing evidence for criterion-related validity. Composite scores on the AZELLA Elementary and Middle Grades forms will be used in this study.

Data Analysis

Attenuation-corrected Pearson correlation coefficients were calculated to investigate the relationship among AIMS subtest scores and AZELLA composite scores. Scatter plots were examined for outliers and to rule out non-linear relationships among the dependent variables (Green & Salkind, 2005). Alpha was set at .05 and Bonferroni methods were used to correct error rate across the multiple correlations. However, due to the large sample sizes and the nature of the variables being compared, it was assumed all correlations would be statistically significant. The square of *r* was calculated and served as the measure of effect size. To evaluate the consistency of the relationship among AIMS and AZELLA performance across grade levels, a Fischer's *z* transformation was performed and 95% confidence limits were obtained using methods described by Zou (2007) Due to the requirement that participants must be administered both AIMS and AZELLA, very few cases were missing values on one or more variables. Two cases were missing scores on the AIMS Reading subtest and were deleted from the eighth grade analysis listwise.

Results

The results of the correlation analysis are provided in Table 2. As expected, all correlations were significant after controlling for Type 1 error at .05. For reading, the strongest association between AZELLA and AIMS Reading occurred in the sample of third graders $r(347) = .71$, $p < .001$, with 50% of the variance in AIMS Reading accounted for by its linear relationship to AZELLA performance. The correlation for fifth graders was also large, but less so than for third graders. The magnitude of the correlation for the eighth grade sample was moderate, with AZELLA performance only accounting for 11% of the variance in AIMS Reading.

Similar trends were found in the correlations between AZELLA and AIMS math although the correlations for each grade were slightly lower than those for reading. As with reading, the correlation for third graders was high, $r(347)= .61$, $p < .001$, $r^2= .37$, and larger in magnitude than those associated with fifth graders or eighth graders. Overall, the results of the correlation analysis suggest that students who perform well on AZELLA also tend to perform well on AIMS, although this relationship is slightly stronger for reading than math, and much stronger for third graders.

To further investigate the relationship between grade level and the association between AZELLA and AIMS, 95% confidence intervals were calculated for between-grade differences for both content areas using Fisher's *z* transformation. As shown in Table 3, the hypothesis that the strength of association between AZELLA performance and AIMS Reading performance decreases as grade level increases is supported. With 95% confidence, the correlation between AZELLA with AIMS Reading is .08 to .41 larger for third graders than fifth graders. The confidence intervals for the remaining grade comparisons for third graders suggest that the strength of association is .13 to .46 larger for third graders than eight graders and .02 to .41 larger for fifth graders than eighth graders. Interpretations of these last two comparisons (third vs. eighth; fifth vs. eighth) need to consider that the two groups were administered similar, but different forms of AZELLA. Despite this, it does seem that overall, the correlation between AZELLA and AIMS math performance decreases significantly as grade level increases.

Table 2. *Correlations among AZELLA composite scores and AIMS subtest by grade*

	AIMS M SS	AIMS R SS
	Third Grade	
Pearson's *r*	.61	.71
R-squared	.37	.50
N	349	349
	Fifth Grade	
Pearson's *r*	.49	.53
R-squared	.24	.28
N	187	187
	Eighth Grade	
Pearson's *r*	.30	.33
R-squared	.09	.11
N	172	174

The results do not support our hypothesis as strongly in math. The 95% intervals for the difference between 3rd and 5th graders and the difference between 5th and eighth graders contained zero, indicating that we do not have sufficient evidence to conclude statistically significant differences between these groups. However, while statistical significance cannot be concluded, the lower bounds are very close to zero, indicating that despite the lack of statistical significance, there may be considerable differences in the strength of association among grades, with a decreasing trend as grade level increases. Although these results are suggestive, one limitation of the current study was an inability to control for immigration status and length of time in ELD services. Including these variables in future research would help to clarify the differences in predictive ability described in this study.

Table 3. 95% CI for Differences in Pearson's *r*

	Mathematics		Reading	
	Lower	Upper	Lower	Upper
3rd and 5th	-.03	.32	.08	.41
3rd and 8th	.10	.44	.13	.46
5th and 8th	.00	.39	.02	.41

Conclusions

Findings and Principles in the Assessment of ELLs

This study finds that at higher grade levels, the "one-test" AZELLA becomes less predictive of academic achievement. In doing so, the use of the AZELLA over predicts the transitioned student's capacity to succeed academically in the regular classroom and places a critical barrier to obtaining an equal education in Arizona. This finding calls into question Arizona's "one-test" procedure for identifying ELL students and their transition to a non-service category, particularly transitioning into the English-only educational curriculum. Hence, the gap between current *practice* in the assessment of English language learners in Arizona and the *standards* set forth through research, policy, and ethics is largely a function of the gap between practical and optimal realities. Due to the many demands and constraints placed on teachers and schools from local, state, and federal governments, including budgeting responsibilities and the many programs implemented each school year, it can be extremely challenging to keep pace with best practices and ethical standards. However, given the large and increasing size of the young ELL child population in Arizona, the current focus on testing and accountability, and the documented problems in current assessment practices, improvements are critical. These improvements are necessary at all phases of the assessment process, including pre-assessment and assessment planning, conducting the assessment, analyzing and interpreting the results, reporting the results (in written and oral format), and determining eligibility and monitoring.

Researchers and organizational bodies have offered principles for practitioners engaged in the assessment of ELLs. Among the most comprehensive comes a list from the National Association for the Education of Young Children (NAEYC; Clifford et al., 2005). Included as a supplement to the NAEYC position statement on early childhood curriculum, assessment and program evaluation, Clifford et al. present detailed recommendations "to increase the probability that all English language learners will have the benefit of appropriate, effective assessment of their learning and development" (p.1). The last of these recommendations concerns further needs (i.e., research and practice) in the field. Because these recommendations—presented here as *principles*—materialized as a collaborative effort from a committee comprised of over a dozen researchers in the field, they are also representative of recommendations found in the literature.

First, *assessment instruments and procedures should be used for appropriate purposes*. Assessments should be used fundamentally to support learning, including language and academic learning. For evaluation and accountability purposes, ELLs should be included in assessments and provided with appropriate tests and accommodations.

Second, *assessments should be linguistically and culturally appropriate*. This means assessment tools and procedures should be aligned with cultural and linguistic characteristics of the child. Tests should be culturally and linguistically validated to verify the relevance of the content (i.e., content validity) and the construct purported to be measured (i.e., construct validity). Moreover, in the case of normed-based tests, the characteristics of children included in the normative sample should reflect the linguistic, ethnic, and socioeconomic characteristics of the child.

Third, *the primary purpose of assessment should be to improve instruction*. The assessment of student outcomes using appropriate tools and procedures should be linked closely to classroom processes. This means relying on multiple methods and measures, evaluating outcomes over time, and using collaborative assessment teams, including the teacher, who is a critical agent for improved learning and development. Assessment that systematically informs improved curriculum and instruction is the most useful.

Fourth, *caution ought to be used when developing and interpreting standardized formal assessments*. Standardized assessments are used for at least three purposes—to determine program eligibility, to monitor and improve learning, and for accountability purposes. It is important ELLs are included in large-scale assessments, and that these instruments continue to be used to improve educational practices and placements. However, those administering and interpreting these tests ought to use caution. Test development issues must be scrutinized, and evidence-based accommodations ought to be provided during accountability assessments.

Finally, *families should play critical roles in the assessment process*. Under federal law, parents have the right to be included in the decision making process regarding the educational placement for their child. Moreover, the educational benefit of the assessment process for a given child is optimal when parents' wishes are voiced and considered throughout. Although family members should not administer formal assessments, they are encouraged to be involved in the selection of assessments and the interpretation of results. The process and results of assessment should be explained to parents in a way that is meaningful and easily understandable.

Future Directions for Practice in Arizona

As mentioned, there is a gap between current assessment practice of young ELLs and what the research and the legal and ethical standards suggest is best practice. It is important, therefore, that research and practice continue an ongoing dialogue to improve this scenario. There are three ways in which researchers and scholars will be able to engage assessment scholarship to this end. Support and necessary funding should be provided by policy makers, institutions of higher education, and other research programs to pursue this course.

First, the field needs more tests developed and normed especially for English language learners. This will require a bottom-up approach, meaning assessment tools, procedures, and factor analytic structures are aligned with cultural and linguistic characteristics of ELL children, as opposed to top-down approaches where, for example, test items are simply translated from their original language to the native languages of young ELLs. Normed-based tests should also take into account important characteristics of the child, including their linguistic, ethnic, and socioeconomic histories.

Second, it is time conceptual and empirical work on student assessment move beyond the student level. That is, the majority of the present discussion reflects the extant literature which has focused heavily on the assessment of processes and outcomes within the student—assessing language and academic learning. With this knowledge-base teachers and schools are expected to adjust aspects of the environment to improve learning. It has become clear that processes outside the student—including within the classroom (e.g., teacher-student interactions, peer to peer interactions), the home (e.g., frequency of words spoken, amount of books), and within the school (e.g., language instruction policies)—affect learning, the field presently lacks conceptual frameworks and the measures necessary to move this research forward to systematically improve student learning. Preliminary research on the role of context in learning suggests that variations environmental factors can increase student engagement and participation (Christenson, 2004; Goldenberg, Rueda, & August, 2006), which, in turn can lead to increased learning—and that the influence of contextual contingencies on learning outcomes is mediated by children's motivation to learn (Rueda, 2007; Rueda, MacGillivray, Monzó & Arzubiaga, 2001; Rueda & Yaden, 2006). Conceptual frameworks should account for the multilevel nature of contexts, including the nesting of individuals within classrooms and families, classrooms within schools, and schools within school districts, communities, and institutions. Moreover, the role of culture and the feasibility of cultural congruence across within- and out-of-school contexts will be important to this work. Meaningful empirical work in this area will require the convergence of research methods (e.g., multi-level statistics and the mixing of qualitative approaches with quasi-experimental designs) and social science disciplines (e.g., cognitive psychology, educational anthropology, sociology of education).

Finally, as the population of young ELLs continues to grow, more serious psychometric work is needed so as to better serve these students in ways in which they will profit from the "right" to be assessed reliably and validly so they might be served effectively. Arizona is presently failing its ELL students in this regard.

References

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.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16-26.

ADE (Arizona Department of Education) & Harcourt (2007). AZELLA Technical Manual. AZ-1. Phoenix: ADE: http://www.ade.state.az.us/oelas/AZELLA/AZELLAAZ-1TechnicalManual.pdf

AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education). (1999). *Testing and assessment: The standards for educational and psychological testing.* Washington, DC: AERA. Online: www.apa.org/science/standards.html.

August, D. (2006). Demographic overview. In D. August & T. Shanahan (Eds.), *Report of the national literacy panel on language minority youth and children.* Mahwah, NJ: Lawrence Erlbaum Associates.

Bainter, T. R., & Tollefson, N. (2003). Intellectual assessment of language minority students: What do school psychologists believe are acceptable practices? *Psychology in the Schools, 40*(6), 899-603.

Borghese, P., & Gronau, R. C. (2005). Convergent and discriminant validity of the Universal Nonverbal Intelligence Test with limited English proficient Mexican-American elementary students. *Journal of Psychoeducational Assessment, 23*, 128-139.

Bracken, B., & McCallum, R. S. (1998). *The Universal Nonverbal Intelligence Test.* Chicago, IL: Riverside.

Bradley-Johnson, S. (2001). Cognitive assessment for the youngest children: A critical review of tests. *Journal of Psychoeducational Assessment, 19*, 19-44.

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

Carter, A. S., Briggs-Gowan, M. J., Ornstein Davis, N. (2004). Assessment of young children's social-emotional development and psychopathology: Recent advances and recommendations for practice. *Journal of Child Psychology and Psychiatry, 45*(1), 109-134.

Christenson, S. L. (2004). The family-school partnership: An opportunity to promote learnign and competence of all students. *School Psychology Review, 33*(1), 83-104.

Clifford, D. et al. (2005). *Screening and assessment of young English-language learners*. Washington, DC: National Association for the Education of Young Children. Available online at http://www.naeyc.org/about/positions/ELL_Supplement.asp

Committee on the Foundations of Assessment (2004). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Coutinho, M. J., & Oswald, D. P. (2000). Disproportionate representation in special education : A synthesis and recommendations. *Journal of Child and Family Studies, 9*, 135-156.

De Avila, E. & Duncan, S. (1990). *Language assessment scales—oral.* Monterrey, CA: CTB McGraw-Hill.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232.

Duncan, S., & De Avila, E. (1988). *Language assessment scales—reading and writing*. Monterrey, CA: CTB McGraw-Hill.

Duncan, S. E., & DeAvila, E. (1998). *Pre-language assessment scale 2000*. Monterey, CA: CTB/McGraw-Hill.

Dunn. L. M., & Dunn, L. M. (1981). *Peabody picture vocabulary test.* Circle Pines, MN: American Guidance Service.

Dunn, L. M., Padilla, E. R., Lugo, D. E., & Dunn L. M. (1986). *Test de vocabulario en imágenes Peabody.* Circle Pines, MN: American Guidance Service.

Figueroa, R. A., & Hernandez, S. (2000). *Testing Hispanic students in the United States: Technical and policy issues*. Washington, DC: President's Advisory Commission on Educational Excellence for Hispanic Americans

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188-192.

Fugate, M. H. (1993). Review of the Bayley Scales of Infant Development, Second Edition. *Mental Measurements Yearbook, 13*.

Gándara, P. & Rumberger, R. (2009). Immigration, Language, andEducation: How Does Language Policy Structure Opportunity? *Teachers College Record,* 111, 6 - 27

García, E. E. (2005). *Teaching and learning in two languages: Bilingualism and schooling in the United States*. New York: Teachers College Press

Garcia, G. E., McKoon, G., & August, D. (2006). Synthesis: Language and literacy assessment. In D. August & T. Shanahan (Eds.), *Developing literacy in second language learners.* Mahwah, NJ: Lawrence Erlbaum Associates.

Genesee, F., Geva, E., Dressler, C., & Kamil, M. (2006). Synthesis: Cross-linguistic relationships. In D. August & T. Shanahan (Eds.), *Report of the national literacy panel on language minority youth and children.* Mahwah, NJ: Lawrence Erlbaum Associates.

Goldenberg, C., Rueda, R., & August, D. (2006). Synthesis: Sociocultural contexts and literacy development. In D. August & T. Shanahan (Eds.), *Report of the national literacy panel on language minority youth and children.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hakuta, K. & Beatty, A. (2000). *Testing English language learners in US schools*. Washington, DC: National Academy Press.

Harry, B. & Klingler, J. (2006). *Why are so many minority students in special education? Understanding race and disability in schools.* New York: Teachers College Press.

Mahoney, K., Haladyna, T., & MacSwan, J. (2007). A validity study of the Stanford English Language Proficiency Test (SELP) as used for classifying English Language Learners. Annual meeting of the University of California Linguistic Minority Research Institute (UCLMRI). Arizona State University, Tempe, May 2-5.

Hernandez, D. (2006). *Young Hispanic children in the US: A demographics portrait based on Census 2000*. Report to the National Task Force on Early Childhood Education for Hispanics. Tempe, AZ: Arizona State University.

McCardle, P., Mele-McCarthy, J., & Leos, K. (2005). English language learners and learning disabilities: Research agenda and implications for practice. *Learning Disabilities Research & Practice, 20*(1), 69-78.

National Association of School Psychologists. (2000). *Professional conduct manual.* Bethesda, MD: Author.

National Clearinghouse for English Language Acquisition (2006). *The growing numbers of limited English proficient students: 1993-94-2003/04*. Office of English Language Acquisition (OELA): US Department of Education.

National Research Council (2008). *Early childhood assessment: Why, what, and how?* Washington, DC: National Academies Press.

Ochoa, S. H., Galarza, S. & Amado, A. (1996). An investigation of school psychologists' assessment practices of language proficiency with bilingual and limited-English-proficient students. *Diagnostique, 21*(4), 17-36.

Ochoa, S. H., Gonzalez, D., Galarza, A., & Guillemard, L. (1996). The training and use of interpreters in bilingual psychoeducational assessment: An alternative in need of study. *Diagnostique, 21*(3), 19-40.

Paredes Scribner, A. (2002). Best assessment and intervention practices with second language learners. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology IV*. Bethesda, MD: National Association of School Psychologists.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Behavior Assessment System for Children, Second Edition.* Minneapolis, MN: Pearson.

Rhodes, R., Ochoa, S. H., & Ortiz, S. (2005). *Assessing culturally and linguistically diverse students: A practical guide*. New York: Guilford.

Rueda, R., MacGillivray, L., Monzó, L., & Arzubiaga, A. (2001). Engaged reading: A multi-level approach to considering sociocultural features with diverse learners. In D. McInerny & S. Van Etten (Eds.), *Research on sociocultural influences on motivation and learning* (pp. 233-264). Grennwuch, CT: Information Age.

Rueda, R. (2007). *Motivation, learning, and assessment of English learners*. Presented at the School of Education, California State University Northridge, Northridge, CA, April.

Rueda, R., & Yaden, D. (2006). The literacy education of linguistically and culturally diverse young children: An overview of outcomes, assessment, and large-scale interventions. In B. Spodek & O.N. Saracho (Eds.), *Handbook of Research on the Education of Young Children, 2nd Ed.*. (pp. 167-186).Mahwah, NF: Lawrence Erlbaum Assoc., Pub.

Santos, R.M., S. Lee, R. Valdivia, & C. Zhang. 2001. Translating translations: Selecting and using translated early childhood materials. *Teaching Exceptional Children* 34 (2): 26–31.

Shepard, L., Kagan, S. L. & Wurtz, L (Eds.) (1998). *Principles and recommendations for early childhood assessments.* Goal 1 Early Childhood Assessments Resource Group. Washington, DC: National Education Goals Panel. Retrieved online at http://eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/24/51/e6.pdf

Wechsler, D. (2003). *The Wechsler Intelligence Scale for Children,* 4th ed. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2004). *The Wechsler Intelligence Scale for Children—Spanish,* 4th ed. San Antonio, TX: Psychological Corporation.

Yzquierdo, Z., Blalock, G., & Torres-Velasquez, D. (2004). Language-appropriate assessments for determining eligibility of English language learners for special education services. *Assessment for Effective Intervention, 29*(2), 17-30.