USE AND IMPACT OF ENGLISH-LANAGUAGE LEARNER ASSESSMENT IN ARIZONA

Stephen B. Lawton

Arizona State University

Paper presented at the annual conference of the Arizona Educational Research Organization, November 6-7, 2008, Phoenix, AZ. Revised May 20, 2009.

ABSTRACT

USE AND IMPACT OF ENGLISH-LANAGUAGE LEARNER ASSESSMENT IN ARIZONA

The Arizona English-Language Learner Assessment (AZELLA) is the backbone of Arizona's new English-language learner (ELL) policy in that it is used to assess students' English-language proficiency in order to place them into groups for English-language instruction and to determine when they have become proficient in English. This paper evaluates a central claim for this assessment – namely, that it is unidimensional – and the implications of the dimensionality of an assessment for practice. Unidimensionality is critical since it is required for Rasch scaling and item response theory (IRT) which are used to array individuals along a single scale that measures English-language proficiency. If AZELLA is not unidimensional, then this ranking process and the programmatic consequences that flow from it would not be valid. It is concluded that AZELLA is not sufficiently unidimensional for its present use, but that it could be used for tailoring English-language programs related to individuals' distinctive skills in listening, speaking, reading, and writing English. (Three tables).

USE AND IMPACT OF ENGLISH-LANAGUAGE LEARNER ASSESSMENT IN ARIZONA[1]

A key element of Arizona's new English-language learner (ELL) policy is the testing of ELLs to determine their proficiency in English so that they can be placed in appropriate groups for English-language instruction. The Arizona English-Language Learner Assessment (AZELLA) was developed for this purpose and it is the only criterion used to assess English language proficiency, so it is a "high stakes" assessment (Wolf, et al., 2008). This paper evaluates a central claim for this assessment – namely, that AZELLA is unidimensional – and considers the consequences that flow from the question as to whether or not the measurement instrument has one or more than one dimension. This issue is critical since unidimensionality is required for the use of a psychometric techniques such as Rasch scaling and item response theory (IRT) that claim to array individuals along a single scale indicating the extent to which they possess a given characteristic such as English-language proficiency. If an assessment is not unidimensional, then these claims may not be valid. The programmatic consequences that flow from the assumption – or violation of the assumption – of unidimensionality include the choice between a single, fixed program versus interventions targeted at the individual strengths and weaknesses of students.

[1] The author thanks Gene Glass, Jeff MacSwan, Jon Vreeken, and Amy Heineke for their suggestions. The author is solely responsible for the content and views expressed here.

Background

The effectiveness with which ELL students learn English in Arizona's publicly funded schools has been an issue in the state of Arizona for almost two decades. In *Flores v. Arizona* (2000), the federal court determined that the state was not funding or providing an effective instructional program for ELLs.[2] Students who were non-native speakers of English were not developing sufficient English-language proficiency for them to enjoy the full benefits of a public education. Subsequently, the passage of *Proposition 203* (2000) mandated the delivery of a structure English-immersion program (SEI) to all ELL students. As well, Title III of the *No Child Left Behind Act* of 2001 set the goal of English proficiency for those students who do not have English as their first language. In response, the Arizona legislature passed *Arizona House Bill 2064* (2006) that mandated the creation of the Arizona English-Language Learners' Task Force which was given a three-fold task: 1) design of a year-long SEI program lasting for four hours per day, 2) determination of the incremental cost of the ELL program, and 3) development of an assessment process to measure the English-language skills of students for placing them in appropriate groups or classes.

AZELLA

In accord with the Task Force's mandate, test experts with Harcourt Assessment Business (now part of the British company Pearson PLC) were employed to develop and administer an English-language proficiency assessment for Arizona's estimated 135,000

[2] In January 2009 the U.S. Supreme Court agreed to hear the state's appeal of the *Flores* case.

ELLs, as they have been in six other states (Wolf, et al., 2008). The resulting Arizona English-language learner assessment program has five levels of assessment: *Preliteracy* – K*; Primary* - 1st and 2nd grades; *Elementary* – 3rd, 4th, and 5th grades; *Middle* – 6th, 7th, and 9th grades; and *Secondary* – 9th, 10th, 11th, and 12th grades[3]. Students identified by school officials as having primary home language other than English (PHLOTE) are administered an assessment based upon their age/grade level.

AZELLA has several different phases in its administration. For preliteracy, students are assessed on four characteristics: Listening, Speaking, Prereading, and Prewriting. Older students, primary through high school, are assessed on five characteristics: Listening, Speaking, Reading, Writing, and Writing Conventions. Based on their scores, students are placed into one of five categories: 1) Pre-emergent, 2) Emergent, 3) Basic, 4) Intermediate, and 5) Proficient (Mesa Public Schools, 2009). The overall reliability for AZELLA scores at the different levels are quite high, ranging from .93 to .97 depending upon assessment level (ADE, 2007).

To verify the unidimensionality of AZELLA, developers used principal components analysis (PCA), which is a technique for simplifying data by discovering underlying traits or components. While there are as always many components as there are original items, items that measure similar information are clustered into a few components which account for most of the information contained in the instrument. For each component, there is an eigenvalue, a term arising in matrix algebra, which indicates the magnitude of each component's explanatory power. The test developers used an adaption of Lord's (1980) criteria that had been suggested by Divgi (1980) and reviewed

[3] Solano-Flores (2008) criticizes the assumption that linguistic ability in either a person's first language or English-as-a-second language is linked to grade level (p. 190). Mahoney, Haladyna, and MacSwann (2009) describe the need for multiple measures to assess language proficiency.

by Hattie (1985) to assess the unidimensionality of AZELLA. This criterion entails comparing the ratio of the difference between the first and second eigenvalues to the difference between the second and third eigenvalues. If the ratio of these two differences is greater than three (3), the authors state, then the instrument is unidimensional.[4]

Table 1 reports the ratios calculated and the values of the first three eigenvalues for each level of AZELLA.[5] It is evident that AZELLA met the Lord/Hattie criterion at all levels and the developers conclude "There is one dominant factor in AZELLA" and the "unidimensionality assumption…. is valid" (Arizona Department of Education, 2007, p. 29).

Insert Table 1 about here

Table 1 also reports the number of items used at each level of testing and the percentage of total variance in scores explained by the first component. The proportion of variance explained by a component is calculated by dividing its eigenvalue by the total number of test items. Kaiser (Hattie, 1985) advised dismissing all components with eigenvalues less than one since each of these components contains less information (i.e., explains less variance) than any single item alone. Dismissing factors with eigenvalues greater than one, however, is problematic. One solution, that adopted by AZELLA's

[4] Hattie (1985) in fact rejects this ratio, noting that "it is not difficult to … to construct cases when this index will fail. For example, given four common factors, if the second and third eigenvalues are nearly equal, then the index could be high. But in a three-factor case, if the difference between the second and third eigenvalues is large, then the index would be low. Consequently, the index would identify the four-factor case as unidimensional, but not the three–factor case!" (p. 146).

[5] Ratios and eigenvalues in Table 1 are taken from ADE (2007, p. 30). Rounding errors account for minor differences between reported numbers and ratios calculated from data. E.g., while the first ratio of the difference between the first two eigenvalues divided by the difference between the second two is [11.33-4.10]/[4.10-2.91] = 7.23/1.19=6.08, the number reported is 6.06 from the original.

developers, is to use Divgi's decision rule related to the ratio of differences being greater than 3. While this criterion was met, it also is evident from the last column of Table 1 that the first components of AZELLA explain between 21 percent and 29 percent of the total variance, depending on level. That is, other components explain from 70 percent to 79 percent of total variability, implying the magnitude or strength of the first factor is moderate, at best, at all levels.

Another important aspect of AZELLA is what is termed its internal structure; that is, the strength of relationships among its building blocks, the subtests for Listening, Speaking, Reading, Writing,[6] and Writing Conventions. For example, at the Kindergarten level, the correlation between Listening and Reading is 0.48 and that between Speaking and Reading is 0.39. Since the square of the correlation coefficient indicates the percentage of variance that one variable explains of another, these correlations indicate that Listening skills explain 23% of the variance in Reading skills, and that Speaking skills explain 15% of the variance in Reading skills. Put another way, ELL children's listening and speaking skills are rather poor predictor's of their reading skills at the kindergarten level.

The relationship between Speaking and Reading is even weaker at the 1st grade level, where the correlation is 0.34, implying Speaking skills explain 10% of the variance in Reading skills. The relationships strengthen at higher grade levels, with correlations at the 6th and 12th grades 0.60 and 0.65, implying explained variance of 36% and 42% respectively (Arizona Department of Education, 2007, pp. 26-28).

[6] For reporting purposes, the Writing and Writing Convention scores were combined into a single score (Arizona Department of Education, 2007), yielding four sub-values: listening, speaking, reading, and writing. Solano-Flores (2008) also notes that "each ELL has a unique set of strengths and weaknesses in each language mode (i.e., listening, reading, speaking, and writing)" (p. 190) in both their first and second languages.

Critique

There is little doubt that AZELLA is a much better instrument than some others that are available, such as the Stanford English Language Proficiency (SELP) test on which AZELLA is in part based and which was previously used for the task of assessing the English-language proficiency of ELL students in Arizona (Mahoney, Haladyna, & MacSwann, 2009; Wolf, et al., 2008). Stephenson, Jiao, and Wall (2004) compared classification decisions for ELL and non-ELL students using SELP and found that students would be misclassified over half the time. Since AZELLA has many more items and higher reliability than SELP, one would expect lower rates of misclassification, although available documentation does not include a comparison of AZELLA scores for native and non-native speakers of English.

Being a strong instrument, however, does not necessarily mean that the fundamental conclusion of AZELLA's developers – that AZELLA is unidimensional – is correct or that decisions made based on the demonstrated degree of unidimensionality are valid.

One immediate issue is the sample of students on which AZELLA was tested. According to the Rasch model and other forms of item response theory, the sample on which an assessment is validated is not relevant since the measure being developed provides a single scale on which each individual is placed. However, if the instrument is multidimensional, then the lack of attention to the sample could be critical. For example, we do not know if performance is influenced by gender, yet educators in the field note

that girls tend to "test out" of ELL classes before boys. Also, we do not know if the first language spoken by a child – Spanish, Navajo, Cantonese, etc. – affects their performance on AZELLA, as might be predicted given the vastly different syntax and tonal traits of these languages and as reported by Solano-Flores (2008, p. 190). The latter also notes that tests like AZELLA, by focusing on the student's second language, "fail to provide important information about an ELL's language development." That is, some individuals may have limited proficiency in several domains of their first language, while others may be advanced in all domains. Persons at higher development levels in one language have advantages in learning a second. Finally, we do not know how native-English speakers would perform in comparison to non-native speakers; educators in the field express skepticism about the degree of success many of their native-English speakers would have if they were administered AZELLA.

Some of the issues involved can be illustrated by considering a principal component analysis of data on the height, weight, and waist size of a non-random sample of graduate students in education. This example is used since the variables are easily observed and understood, unlike inferred mental traits such as reading and writing skills. Tables 2 and 3 present the summary statistics and the PCA of the data. Notable from this example are 1) that the first component has an eigenvalue of 2.5, 2) that all other eigenvalues are less than one, and 3) that the ratio as defined by Divgi is 12. The first factor explains 83% of the total variance and could well substitute for the formula commonly used for calculating an individual's body mass index (BMI) to determine if a

person is at risk of health problems related to obesity.[7] Yet, we would not conclude from this example that individuals' size or stature is strictly one dimensional; height, weight, and waist size are all conceptually distinct albeit correlated traits. Depending on one's purposes – assessing an individual's health risks or tailoring a suit – a unidimensional index may or may not be useful.

Insert Table 2 about here

Insert Table 3 about here

With AZELLA, a similar but more powerful argument for concluding the existence of distinctive traits applies. The relative weakness of the first component, which explains on average just 26% of total score variance, and the weak to moderate correlations among speaking, listening, reading, and writing, mean that these four traits must be considered in their own right . That is, even though an AZELLA score is a useful overall indicator of English competence, it does not imply that individual students are equally proficient in all four language domains. A person can understand a language without great facility in speaking it and no ability to read or write it. Conversely, an individual can be very proficient in reading a language, but have no idea how it sounds and can neither comprehend the spoken language nor be able to speak it.

[7] BMI, using metric measures, is a person's weight divided by the square of the person's height. Using an online height calculator, the BMI for the average student in the sample is 28.4, which implies they are, on average, overweight. Combined with waist size, the risk of health problems for the average student is increased but not considered high. See http://www.nhlbi.nih.gov/health/public/heart/obesity/lose_wt/bmi_dis.htm

In sum, the claim that AZELLA is unidimensional is misleading, at best, and possibly false altogether. It is misleading in the sense that it implies that there is a single score that sums up a person's language ability, just as the BMI sums up the person's physical stature. And if the claim is false, then the entire conceptual structure that underlies AZELLA, including the use of the Rasch/IRT model, is invalid. This, along with the lack of test data from a diverse sample of students in diverse situations, means that using AZELLA scores to assign students to a uniform academic program is inappropriate. And, although this conclusion is demonstrated here only for AZELLA, it is quite likely that it applies other English proficiency measures developed from the same assumptions; i.e, the assessment instruments in the 30 or so other states that depend upon a single measure to assess English language learners (Wolf, et al., 2008).

Of course, the central issue may not be AZELLA per se, but the notion that all ELL students can and should be assessed on a single scale of English language proficiency, as called for in Title III of NCLB, and assigned to a uniform four-hour per day SEI program, as is legislatively mandated in Arizona. Such prescriptive approaches call for a prescriptive measurement instrument. It is quite possible that AZELLA is sound instrument for assigning students to an inappropriate program. That is, the problem may be with AZELLA's intended rather than with the instrument itself, a conclusion consistent with research on the complexity of context and first-language skills on second language learning (Solano-Flores, 2008).

Implications

If we reject the notion that AZELLA (and other similar assessment instruments) is unidimensional, then the logical alternative is that it is a composite index incorporating information on students' listening, reading, speaking, and writing skills. As a composite index, the overall AZELLA score may be a useful for approximate classifications yet be useless for tailoring programs to meet students' specific needs – needs that are in fact captured at least in part by AZELLA's subscales. That is, the appropriate program for a child who scores high on listening and reading is not the same as for a student who scores high on listening and speaking. The former would have high levels of comprehension and would need a program tailored to English expression, while the latter might lack even the rudiments of literacy and require a program focused on reading and writing.

Historically, schools are noted for their efforts in classifying and grouping students – by age, by gender, by race, by neighborhood, by academic ability (e.g., special education, general, and advanced programs). In Arizona, AZELLA along with the SEI program have, perhaps inadvertently, served to achieve many of these groupings at once. Classroom teachers and district administrators report that in schools with many ELL students, one or two classes at a grade level have been created for the ELL students who then no longer mix with native-English speakers. In schools with few ELL students, multi-aged withdrawal classes are common; students in these classes miss regular instruction on academic topics since it is not feasible to introduce grade-appropriate academic content for each student. Secondary students in four-hour SEI programs miss

out on academic courses needed for graduation and college (Lewin, 2009). As well, in some schools, the regular classes are predominately white or Chicano (American born of Hispanic descent) while the ELL classes are composed of recent Hispanic immigrants, who are stigmatized in the process. And with girls testing out of ELL classes before boys, in upper elementary grades classes may be separated by gender was well as ethnicity or immigration status.[8]

The alternative to this situation would probably involve individualized programs based on each child's language skills[9]; such an approach is not easy. Nevertheless, it would appear to be more sensible to design programs that build on students' strengths and attend to areas in which they are developing. AZELLA probably would be suitable for assessing students along its four dimensions – listening, speaking, reading, and writing – so that students might be provided instruction that matches their profile of strengths and weakness. Indeed, individual classifications and interpretive remarks are made for each mode on a report provided by the test publisher (Mesa Public Schools, 2009). Taking such a multidimensional approach to AZELLA and to instructional interventions is the logical outcome of this analysis. Indeed, in practice, some ELL teachers do adapt instruction beyond a student's numerical classification as being a "1," "2," or "3," or "4," although this flexibility may contravene the mandated SEI program.

To test this conclusion about the need to adapt the SEI program requires, first, strong evidence on the progress of students in Arizona's new four-hour SEI program. Such evidence should include data on initial student performance on each of AZELLA's

[8] Illustrations based on discussions and e-mail exchanges with a non-random sample of Arizona teachers and administrators during fall 2008 and spring 2009.

[9] Both first and second language skills are important; non-English speakers with excellent skills in all four modes of first languages other than English will respond very differently to English-language instruction than will students with limited reading and writing skills in their first language (Solano-Flores, 2008).

subscales, the actual classroom practices of teachers, and the subsequent performance of students reclassified as "Proficient" on both state and national assessments. If students test out of the ELL program because they possess strong speaking and listening skills, but fail high stakes tests due to inadequate reading or writing skills, then their failure will be due as much to the invalid assumption of AZELLA's unidimensionality as to the quality of the SEI program.[10] As well, if the program succeeds because of teachers' adaptation of the SEI program to student needs while disregarding mandated requirements, then this too needs to be revealed.

Second, systematic experimentation is needed on alternative interventions to determine if instruction tailored to students' individual language needs on each of the four dimensions is more effective than a uniform SEI program, either as originally planned or as adapted by teachers. As part of such studies, the impact of alternative approaches on social and academic variables and first language skills should be considered. Even if alternative programs are not any more effective in terms of developing English-language skills than the mandate four-hour SEI program, they would likely be less prone to inadvertently separate students into groups that limit students' social development, stigmatize them, fail to capitalize on their levels of linguistic development, or withhold them from academic courses needed to graduate and advance toward college.

[10] Mahone, Haladyna, and MacSwann (2009) report that students deemed Proficient on SELP, AZELLA's predecessor, had higher failure rate on state assessments than did English-first-language students.

References

Abedi, J. (Ed.) (2007). English language proficiency assessment in the nation: Current status and future practice. Davis, CA: University of California, Davis. Retrieved May 20, 2009, from http://education.ucdavis.edu/research/ELP_Report.pdf

Arizona Department of Education & Harcourt Assessment, Inc. (2007). *Arizona English Language Learner Assessment: Technical manual*. Phoenix: Author. Retrieved November 24, 2008, from http://www.ade.state.az.us/oelas/AZELLA/AZELLAAZ-1TechnicalManual.pdf

Arizona House Bill 2064 (47th Legislature, 2nd Regular Session, 2006). Retrieved November 2, 2008, from http://www.azleg.gov/legtext/47leg/2r/bills/hb2064c.pdf

Arizona Proposition 203(2000). Retrieved February 7, 2008, from http://www.azsos.gov/election/2000/Info/pubpamphlet/english/prop203.pdf

Divgi, D. R. (1980, April). Dimensionality of binary items: Use of a mixed model. Paper presented to the annual meeting of the National Council on Measurement in Education, Boston, MA.

Flores v. State of Arizona, 48 F.Supp. 2d 937 (D. Ariz. 1999); 172 Supp. 2d 1225(D. Ariz. 2000); 160 F.Supp. 2d 1043 (D. Ariz. 2000); U.S. Dist. 23178 (D. Ariz. 2002); 480 F.Supp. 2d 1157 (D. Ariz. 2007).

Hattie, J. (1985). Methodology review: Assessing unidimensionality. *Applied Psychological Measurement, 9*(2), 139-164.

Lewin, T. (2009, May 19). End is near in a fight on the teaching of English. New York Times. Retrieved May 21, 2009, from http://www.nytimes.com/2009/05/20/education/20flores.html

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* New York: Erlbaum Associates.

Mahoney, K., Haldyna, T., & MacSwan, J. (2009). The need for multiple measures in reclassification decisions: A validity study of the Stanford English Language Proficiency Test. Paper presented at the Annual Conference of the American Educational Research Association, San Diego, CA, April 13-17, 2009.

Mesa Public Schools, English Language Acquisition Department (2009). The AZELLA: Test components and scoring. Retrieved May 8, 2009, from http://www2.mpsaz.org/elad/servesource/infoppt/

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English Language Learners. *Educational Researcher, 37*(4), 189-199.

Stephenson, A., Jiao, H., & Wall, N. (2004). *A Performance Comparison of Native and Non-native Speakers of English on an English Language Proficiency Test*. Retrieved May 26, 2008, from http://harcourtassessment.com/NR/rdonlyres/EB7D1CF5-97D5-4851-A50C-4B2DF207D72E/0/NativeNonNative.pdf

Wolf, M. K., Kao, J., Griffin, N., Herman, J. L., Bachman, P. L., Chang, S. M., & Farnsworth, T. (2008). *Issues in assessing English language learners: Proficiency measures and accommodation uses. Practice Review (Part 2 of 3).* Los Angeles, CA: National Center for Research and Evaluation, Standards, and Student Testing, UCLA. Retrieved May 20, 2009, from http://www.cse.ucla.edu/products/summary.asp?report=732

Table 1: Ratios and eigenvalues for AZELLA by level

Level	Ratio	Eigenvalue 1	Eigenvalue 2	Eigenvalue 3	# Items	% Explained by First Component
Preliteracy	6.06	11.33	4.10	2.91	53	21.4
Primary	4.24	16.54	5.48	2.88	76	21.8
Elementary	7.19	21.73	4.41	2.01	76	28.6
Middle	7.48	24.69	4.66	1.98	84	29.4
High School	8.79	23.08	4.01	1.84	84	27.5

Table 2: Descriptive statistics for a sample of graduate students (n = 22)

Variable	Height	Weight	Waist
Height (inches)	67.7 (5.52)	0.62	0.69
Weight (lbs.)		187.0 (63.9)	0.79
Waist (inches)			34.0 (5.76)

* Means and standard deviations on diagonal and correlations off diagonal.

Table 3: Principal component analysis for height, weight, and waist data

Ratio	Eigenvalue 1	Eigenvalue 2	Eigenvalue 3	# Items	% Explained by First Component
11.62	2.504	0.341	0.155	3	83.5