Do the AZELLA Cut Scores Meet the Standards?

A Validation Review of the Arizona English Language Learner Assessment

Ida Rose Florez
Arizona State University

July 2010

The Civil Rights Project
Proyecto Derechos Civiles

Abstract

The Arizona English Language Learners Assessment (AZELLA) is used by the Arizona Department of Education to determine which children should receive English support services. AZELLA results are used to determine if children are either proficient in English or have English language skills in one of four pre-proficient categories (pre-emergent, emergent, basic, intermediate). Children who test at or above the proficient cut score in English are placed in mainstream classes without English language support. Children who obtain scores below the proficient cut scores receive English language support services in state-mandated Structured English Immersion classes. Whenever tests are used to make high-stakes decisions, especially about vulnerable populations (e.g., children), it is the test developers' responsibility to ensure the instrument yields fair and valid results. When cut scores are used as the primary interpretation of the test they are key to establishing the test's validity. This validation study found that cut scores for the AZELLA are of questionable validity. The procedure used to set the cut scores is criticized by national measurement experts as ineffective and obsolete. Further, the test developers do not adequately establish the expertise of the judges used to set the cut scores. Evidence from the cut-score-setting process indicates judges did not come to consensus at the kindergarten level. Analysis of empirical evidence suggests cut scores over-identify kindergarten children and under-identify older children. Finally, the test developers rejected 85% of the cut scores recommended by the standard-setting judges, setting cut scores higher than recommended for kindergarten and lower than recommended for older children, without describing their process or rationale.

Do the AZELLA Kindergarten Cut Scores meet the *Standards*[1]? A Validation Review of the Arizona English Language Learner Assessment

Between 2000 and 2010, court decisions, voter initiatives, and federal mandates dramatically changed English proficiency assessment policies for Arizona public school children whose first language is not English (see Mahoney, Haladyna, & Macswan, 2010, for a detailed history of English proficiency testing in Arizona from 2000-2010). In response to these policy mandates the Arizona Department of Education (ADE) partnered with Harcourt Assessment, Inc. to develop the *Arizona English Language Learner Assessment* (AZELLA; ADE & Harcourt, 2007). The test is given to students in grades K to 12 whose parents indicate on a Home Language Survey that the child's primary language is not English. AZELLA results are used to determine if children are either proficient in English or have English language skills in one of four non-proficient categories (pre-emergent, emergent, basic, intermediate). Placement in each of these categories (also called proficiency levels) is determined by comparing the child's AZELLA score with *cut scores*. A cut score is the minimum score a child must obtain to fall into one of the five proficiency levels. Children who test at or above the proficient cut score in English are placed in mainstream classes without English language support. Children who obtain scores below the proficient cut scores receive English language support services in state-mandated Structured English Immersion (SEI) classes. In 2006, 83,167 students were tested with the AZELLA. The percent of students at each grade level falling at or above the proficient cut score ranged from 1.87 in 1st grade to 33.99 in 5th grade.

Determining the English proficiency of English Language Learners (ELLs) substantially impacts children's educational opportunities. U.S. courts (e.g., *Lau v. Nichols*, 1974) and federal legislation (e.g., Bilingual Education Act of 1968; Equal Educational Opportunity Act of 1974) have recognized the right of ELLs to receive appropriate language support services. In *Lau v. Nichols* (1974), speaking of educating ELLs, the U.S. Supreme Court wrote:

> "…there is no equality of treatment merely by providing students with the same facilities, textbooks, teachers, and curriculum, for students who do not understand English are effectively foreclosed from any meaningful education. Basic English skills are at the very core of what these public schools teach. Imposition of a requirement that, before a child can effectively participate in the educational program, he must already have acquired those basic skills is to make a mockery of public education. We know that those who do not understand English are certain to find their classroom experiences wholly incomprehensible and in no way meaningful."

Whenever tests are used to make high-stakes decisions, especially about vulnerable populations (e.g., young children), it is the test developers' responsibility to ensure the instrument yields fair and valid results (American Educational Research Association, the

[1] This refers to the AERA/APA/NCME 1999 Joint Standards for testing that are the "standard" for the field of educational and psychological testing.

American Psychological Association, and the National Council on Measurement in Education, 1999). Safeguards for fairness and validity are delineated in the *Standards for Educational and Psychological Testing* (*Standards*; AERA, APA and NCME, 1999). The *Standards* were developed by a joint committee appointed by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). They are widely accepted by academic and practicing professionals in the field of educational and psychological testing as the criteria for judging the quality and fairness of assessment instruments. Indeed, the AZELLA *Technical Manual* (*Manual:* ADE & Harcourt, 2007) recognizes the importance of the *Standards* by indicating that the instrument "is in accordance with" them (*Manual*, p. 5).

Emphasizing that the "improper use of tests…can cause considerable harm to test takers and other parties affected by test-based decisions" (*Standards*, p. 1), the *Standards* are specifically designed "to provide criteria for the evaluation of tests" (*Standards*, p. 2). The National Association for the Education of Young Children (NAEYC) & National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE), speaking on behalf of young children, also argue that the improper use of tests can harm test-takers. The two organizations jointly issued a widely accepted position statement outlining the proper use of assessment with children birth through age 8 (NAEYC & NAECS/SDE, 2003). The position statement warns against basing high-stakes decisions for young children on standardized tests (i.e., tests that require young children to demonstrate skills in predetermined ways) and against the practice of making high-stakes decisions based on a single instrument. Using the AZELLA (or any other English language proficiency test) in isolation to determine which young children receive English language support calls into question the validity of the assessment mandating substantial evidence that the instrument is clearly valid for its intended purpose.

According to the *Standards*, validity is "the most fundamental consideration in developing and evaluating tests" (AERA, APA and NCME, 1999, p. 9). Validating a test involves evaluating how well the results, and the *interpretations* of results, accurately fulfill the test's purpose. The AZELLA was designed to meet the assessment mandates of the federal *No Child Left Behind* (NCLB) Act of 2001 (ADE & Harcourt, 2007), which requires states to assess the progress of ELLs towards English proficiency and their ability to "meet challenging State academic content standards" (p. 5). According to ADE & Harcourt (2007), "the purposes of the AZELLA are (1) to determine appropriate placement of students who have a Primary Home Language Other Than English (PHLOTE) and (2) to measure PHLOTE students' annual improvements in achieving English language proficiency. Thus, the test results provide the criteria for entry into SEI programs and the criteria for exiting SEI programs" (p. 5). Because assignment to an SEI classroom in Arizona has significant consequences for students with respect to the curriculum they will be provided (or not), and the exposure they will have to mainstream students in their school, this test can be considered "very high stakes" and therefore should meet the most stringent criteria for validity and reliability.

Evaluating a test's validity is a complex process that includes explicating evidence to support the rationale for how the test was constructed and how well the test actually accomplishes its purposes. When cut scores are used as the primary interpretation of the test (as they are in the AZELLA), they are key to establishing the test's validity. According to the

Standards, where cut scores "embody the rules according to which tests…are interpreted" test validity "may hinge on the cut scores" (AERA et al., 1999, p. 53).

Cut scores are not determined by an exact science. To set cut scores, test developers choose from a variety of methods referred to as *standard setting procedures*. Some standard setting procedures, such as the one used to determine the AZELLA cut scores, rely on the judgments of a panel of people (called *judges*) who are familiar with the test content and purpose. Thus, validating the AZELLA cut scores is critical to determining if the test fairly and accurately determines which children need English support services and which have English proficiency skills sufficient for them to learn adequately in mainstream English classes. This critical evaluation examines five validity issues related to AZELLA cut scores: (1) standard setting procedures; (2) consensus of judges' recommended cut scores for kindergarten; (3) selection and qualifications of judges; (4) the final selection of cut scores; and (4) the empirical evidence for AZELLA cut score validity. I argue that the AZELLA cut scores substantively fail to meet the *Standards* in these five areas and that the cumulative effect of these failures calls into question the validity of the AZELLA for identifying the English proficiency in of Arizona's ELLs.

Standard Setting Procedures

The *Standards* emphasize that when "results of the standard-setting process have highly significant consequences, and especially where large numbers of examinees are involved, those responsible for establishing cut scores should be concerned that the process by which cut scores are determined be clearly documented and defensible" (*Standards*, p. 54). The AZELLA *Manual* provides information about the standard setting process used to derive the five proficiency levels (pre-emergent, emergent, basic, intermediate, and proficient) and four cut-scores (emergent, basic, intermediate, and proficient). Although all proficiency levels are important, the proficient level cut score is most critical as it determines which children enter, remain in, and exit language support services.

The *Standards* require test developers to document "the rationale and procedures used for establishing cut scores" (*Standards*, p. 59). The AZELLA *Manual* indicates that the selected standard setting procedure, a modified-Angoff (1984) method, "is sometimes referred to as the ACT/NAGB [American College Test/National Assessment Governing Board] standard setting process" and "has a long and successful history in similar applications for both educational and professional certification assessments. … This method has been applied successfully and it is a widely recognized method" (ADE & Harcourt, Inc., 2007, p. 42). Contrary to these claims, standard setting procedures based on the Angoff method have been widely, and intensely, criticized (for a more thorough discussion of the issues surrounding standard setting for English proficiency assessments see Abedi, (2007). Angoff methods require panels of judges to review test items and estimate the performance levels of "minimally competent" test-taker for each level. Along with other criticisms, Angoff methods have been criticized for the subjectivity and complexity of the judgments standard-setting panelists are required to make. Indeed, under pressure from "significant members of the profession" (Brown, 2000, p. 17) the National Assessment Governing Board (NAGB), which oversees the National Assessment of Educational Progress (NAEP), commissioned a study of standard setting methods in 1992. According to

Brown (2000), "the study was decidedly critical of the Angoff method" (p. 22). The study, conducted by Lorrie Shephard, Robert Glaser and Robert Linn, three nationally recognized educational measurement experts, found that:

> The Angoff method requires panelists to make conceptual judgments that are complex: (1) they must conceptualize borderline performance specifically enough to make p-value estimates; (2) they must be able to understand relationships among the content framework, the achievement level descriptions, the NAEP assessment items, and the performance levels; and (3) there should be evidence that panelists are grounding their estimates to factors in the achievement level descriptors to the NAEP items. The study concluded that the achievement level descriptions evolved significantly during the process of setting achievement levels. The panelists reported that personal and experiential background influenced their judgments as well as the achievement level descriptions. As a result, the study concluded that "many participants were not making systematic judgments based in specific features of the descriptions" (Brown, 2000, p. 21).

Based on these recommendations, the NAEP no longer uses a modified-Angoff method, but now uses a book-marking system to determine cut scores (Brown, 2000). Bookmarking standard-setting methods use booklets of test items. The items are placed one to a page in order of difficulty. Judges place a bookmark between the pages of the items where they believe the cut score should fall. Bookmarking procedures also involve deliberate attempts to build consensus among judges. According to Reckase (2000), "Following placement of the bookmarks, the judges have a discussion session with the goal of reaching consensus on their placement. If consensus is not reached, the average bookmark placement is used for the cut-score" (p. 59). Most other statewide English proficiency assessment development projects followed NAEP's lead and have chosen book-marking methods over Angoff-based procedures. For example, the Mountain West Assessment Consortium (MWAC), in its development of a NCLB-compliant English proficiency assessment, recommended book-marking procedures for setting cut scores to the consortium's eleven states (Alaska, Colorado, Idaho, Michigan, Nevada, New Mexico, North Dakota, Oregon, Utah, and Wyoming). According to the report *English Language Proficiency Assessment in the Nation: Current Status and Future Practice* (Abedi, 2007), forty-three states determine cut-scores with methods other than modified-Angoff. Of the remaining seven states that use modified-Angoff methods for their NCLB-mandated English proficiency assessment, six use tests developed by Harcourt.

The AZELLA *Manual* fails to document any rationale for using modified-Angoff procedures. In light of the criticisms surrounding Angoff-based procedures and the growing body of evidence supporting other methods (such as book-marking) as more valid, the rationale for the selected standard-setting procedure is judged to be inadequate, rendering the cut scores, and thus the high-stakes decisions based on them, of questionable validity.

Selection Process and Qualification of Judges

According to the *Standards*, "When a validation rests in part on the opinions or decisions of expert judges, observers or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented" (*Standards*, p. 19). The *Standards* indicate test developers are responsible for documenting the selection process and qualifications of standard-setting judges. To ensure cut scores are fairly determined, developers of most large-scale assessments recruit panelists from multiple stakeholder groups (Hambleton & Pitoniak, 2006). Selection procedures often include a priori sampling schemes that indicate the proportion of judges that are to be selected from various categories of stakeholders. For example, federal law requires that standard-setting panelists for the NAEP broadly represent the nation on a variety of demographics including gender, ethnicity, region and community type (Bourque, 2009). In addition to selection procedures, test developers are expected to document the qualifications of standard-setting judges. Such documentation is especially important when test developers use standard-setting procedures that rely heavily on expert judgments, such as modified-Angoff methods. Given that Angoff methods require judges to make complex conceptual judgments as they determine cut scores, establishing the qualifications of the panelists is paramount to defending the credibility of the cut-scores. AZELLA test developers should report the minimum number of years of teaching (or other educational) experience required to be a judge, the number of years experience at the particular grade level, the number of years experience teaching dual-language learners, and the racial, ethnic, and gender make up of the expert panels. Panels should also include subject-area specialists. In developing a test of second-language English proficiency development, to meet the *Standards,* all standard-setting panels should include an expert in second language acquisition and development of English in non-native speakers.

ADE and Harcourt (2007) failed to provide adequate details about the qualifications of judges and the process by which they were selected. The description in the AZELLA *Manual* refers vaguely to judges needing to be "familiar" with teaching ESL, but the *Manual* does not specify how much grade-level or ELL teaching experience judges were required to have. No range or average number of years taught is provided. Specifically, the test developers fail to document the early childhood expertise of panelists setting kindergarten cut scores, they fail to document the presence of second language acquisition experts on any panel, and they fail to document the educational level or demographics of the panelists. There is also no indication of a sampling plan designed to ensure gender, racial and ethnic representation, and the statement "balanced regional representation" is undefined.

Cut Score Consensus

One of the primary goals of the standard-setting process is to build consensus among judges. In a modified Angoff procedure, judges are asked to estimate appropriate cut scores in multiple rounds, and are given feedback about the relationship of their cut scores to other judges' scores or to external criteria such as information about how students performed at their recommended score (Reckase, 2000). For valid standard setting, the differences between judges' recommended cut scores should decrease each round, indicating greater consensus around the

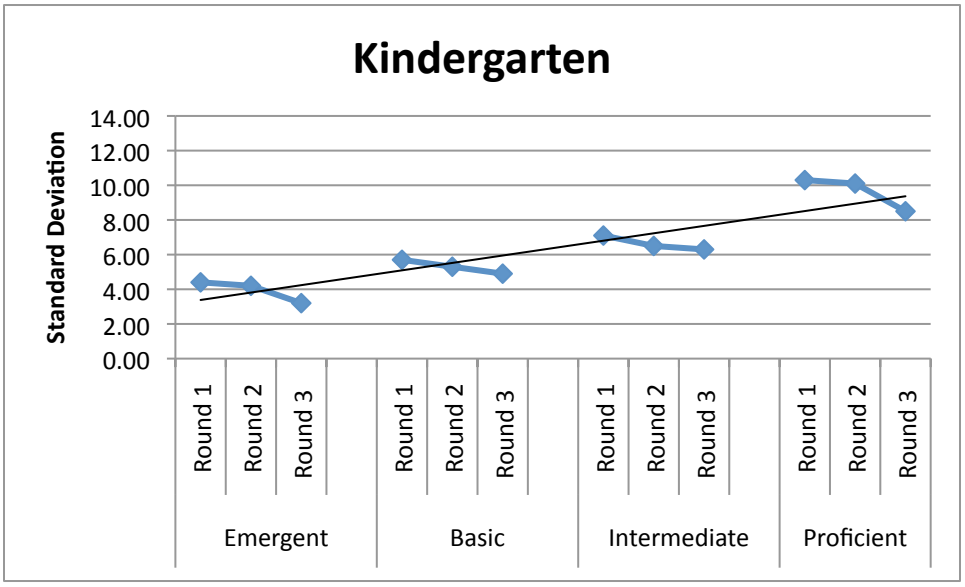
appropriate place for cut scores to fall. The greatest consensus should be found for the cut score associated with the highest stake decisions, in this case, the "proficient" cut score. Although all cut scores should be defensible, the proficient cut score results in the most critical decisions: to enter, retain, or exit a child from the English language support program. Finally, as differences in judges' recommended cut scores decrease across standard-setting rounds, the overlap of recommendations for cut scores at each proficiency level should also decrease. It is undesirable, for example, for some judges to recommend, in the final round, that some scores fall below the proficient cut score while others recommend that the same scores represent proficient English language skills. Thus, analysis of data from successful standard-setting should find decreases in variability between judges' recommendations across rounds at each grade level, the greatest consensus among judges' opinions for the proficient cut score, and little to no overlap in the range of recommended cut scores for adjacent proficiency levels by judges in the final round.

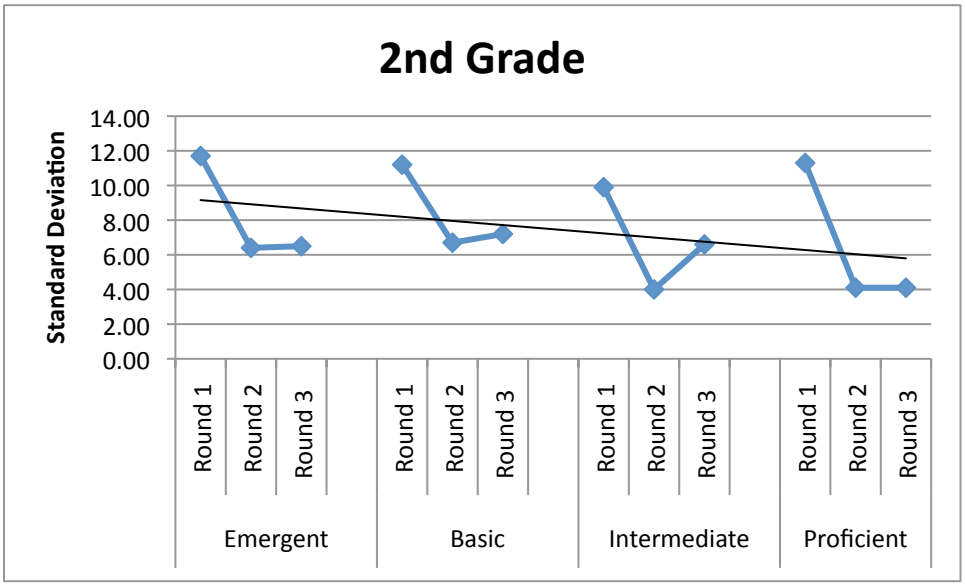
Figure 1 depicts variability (standard deviations taken from data provided in the AZELLA *Manual*) from the three standard-setting rounds. An analysis of trends in variability in judges' recommended cut scores indicates that the standard-setting processes yielded varying rates of consensus around cut scores with more agreement for older grades and less consensus at the kindergarten level.

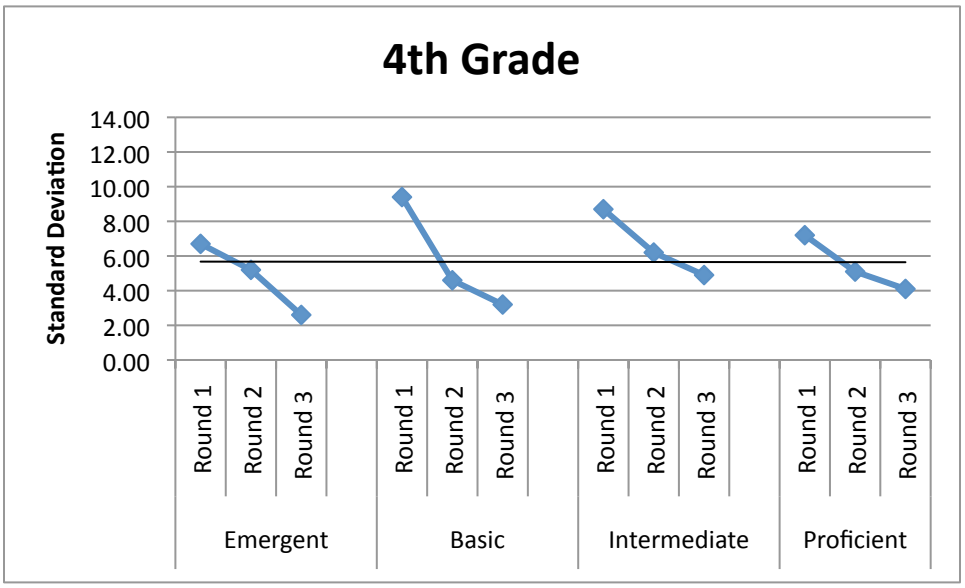
The solid line without data point indicators on Figure 1 represents the variability trends across proficiency levels. For 2nd, 7th, and 9th grades judges reached the greatest consensus for the proficient cut score. For 4th grade variability remained even across cut score levels. For kindergarten, disagreement increased from the emergent to the proficient level, with variability (i.e., disagreement) highest for the high-stakes proficient cut score. The data trend for kindergarten is exactly opposite the desired trend. These findings are of particular concern for several reasons. The majority of children tested with the AZELLA are kindergarten students. In 2006, nearly 28,500 kindergarten children[2] were tested with the AZELLA, compared to 8987 1st graders, 7013 2nd graders, and 3500-5000 students at each of the other grade levels. Evidence from longitudinal studies of children's development indicates English language proficiency at the end of kindergarten predicts key indicators of educational attainment and academic achievement such as 4th grade reading rates (Lesaux, Rupp, & Siegel, 2007). Thus, inaccurate high-stakes placement decisions based on faulty AZELLA results at kindergarten have the potential to adversely affect tens of thousands of young children every year; effects that may negatively impact their entire school career and quality of adult life. Young children who would be better served in classrooms where they have access to the same curriculum as their native English-speaking peers will be denied this opportunity and likely set them back academically; those who need services but are not properly identified face similar long-term consequences as they cannot be expected to succeed in classrooms where they do not understand the lessons. Further, inaccurate placement decisions at kindergarten make it difficult to determine the efficacy of English support services. If kindergarten children are over-identified then reclassification rates will be artificially elevated, which may be erroneously interpreted as evidence of program effectiveness.

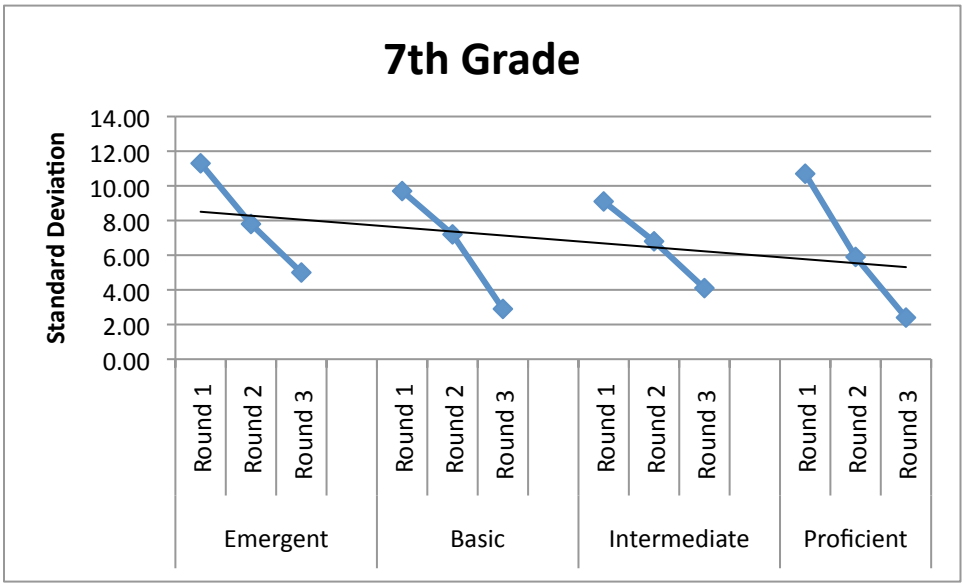
[2] The exceptionally high number of kindergarteners tested is due to the state having tested "all *new* PHLOTES" in 2006, and as such, the bulk of all *new* PHLOTES are found in kindergarten.

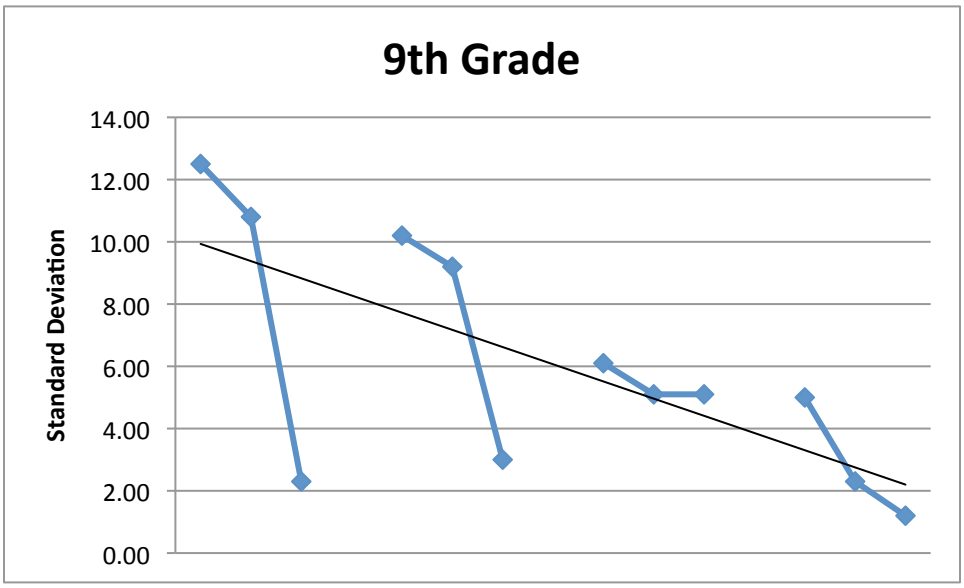
Figure 1. Variability (in standard deviations) in cut score recommendations across standard-setting rounds, proficiency levels and grades (data from Appendix c.3, AZELLA *Manual*).






4th Grade
Standard Deviation
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
Round 1
Round 2
Round 3
Emergent
Basic
Intermediate
Proficient


7th Grade
Standard Deviation
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
Round 1
Round 2
Round 3
Emergent
Basic
Intermediate
Proficient



To further evaluate judges' agreement around cut scores at the kindergarten level, a comparison was made of the range of judges' recommended cut scores for kindergarten and 9th grade. Figure 2 presents a rough depiction the judges' cut score recommendations from each of the three standard-setting rounds for all four proficiency levels for kindergarten and 9th grade.

The bands represent the range of recommended cut scores from lowest to highest (rounded to the nearest 5). Cut score consensus should result in little if any overlap, especially by Round 3. For example, at the 9th grade level the recommended ranges for the four cut scores in Round 3 do not overlap. There is consensus among the judges that the cut scores fall somewhere within the bands, and, importantly, do *not* fall outside the range. The picture is not so clear for kindergarten. Judges' recommendations for the Intermediate and Proficiency cut score in Round 3 overlapped by 14 points, or by 30% of the total range for the two proficiency levels (18 to 65). At the kindergarten level judges did not agree on the boundaries for any of the cut scores.

This analysis further confirms a lack of judges' consensus at the kindergarten level and leads one to wonder about the qualifications of judges at the pre-literacy level. As discussed earlier, assessment of language development is complicated. Assessment of any skill in early childhood is extremely difficult and frequently unreliable (Meisels, 1986; Nagle, 2007), and there is even more reason for concern with students who are not primarily English speakers and as a consequence may feel disoriented or even frightened by the testing procedure. Often, individuals with little expertise in early childhood development or assessment are selected to administer and make judgments about tests that will be used with young children (NAEYC & NAECS/SDE, 2003). It is impossible to judge the adequacy of the pre-literacy qualifications of the judges used for this panel as no information about their credentials, experience or skills has been documented.

Figure 2. Overlap of judges' recommended cut scores for each proficiency level, by standard-setting round, for Kindergarten and 9th grade.

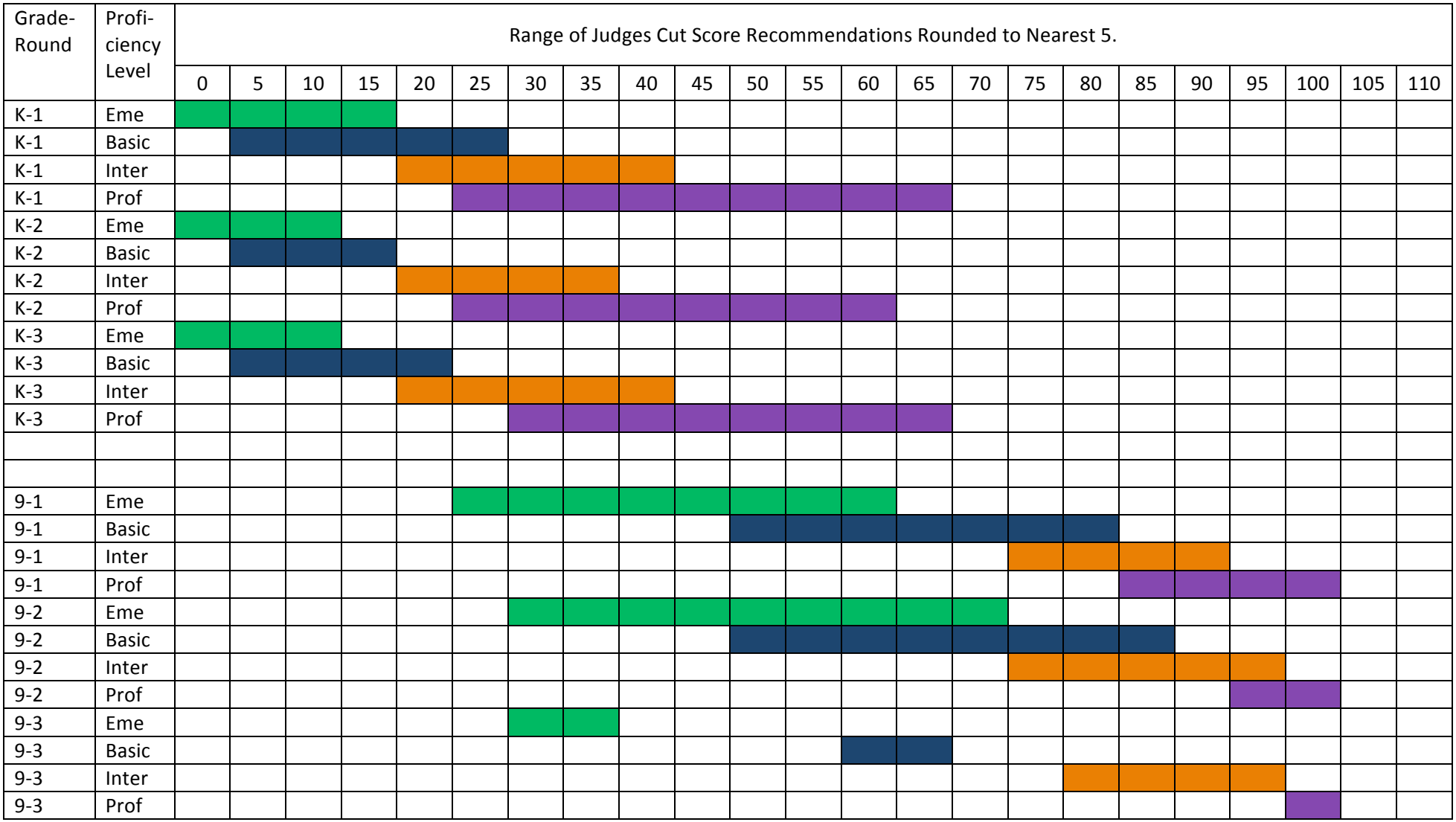


Table 1. Range of recommended cut scores, recommended cut scores and final cut scores for each proficiency level for grades kindergarten, 2nd, 4th, 7th, & 9th. Source: AZELLA *Manual* (ADE & Harcourt, 2007).

Total Test	Emergent			Basic			Intermediate			Proficient		
Grade	Range Of Rec'd Cut Scores	Rec'd Cut Score	Final Cut Score	Range Of Rec'd Cut Score	Rec'd Cut Score	Final Cut Score	Range Of Rec'd Cut Score	Rec'd Cut Score	Final Cut Score	Range Of Rec'd Cut Score	Rec'd Cut Score	Final Cut Score
K	1-11	5	11	5-22	14	18	18-41	27	40	29-65	43	58
2nd	4-25	15	20	19-50	40	30	50-75	60	60	81-95	85	85
4th	17-25	20	18	35-46	42	27	56-75	67	59	76-90	87	82
7th	30-45	30*	19	55-64	57	30	75-86	83	64	97-108	100	91
9th	30-40	38	17	60-70	61	28	80-97	85	61	100-105	102	88

*Error in the original data as reported in the AZELLA *Manual*

Final Cut Score Selection

The AZELLA *Manual* standard-setting procedures yielded three rounds of recommended cut-scores. Data for each of the five (i.e., kindergarten, 2nd, 4th, 7th and 9th grade) teams included the range of recommended scores, means, standard deviations and median (ADE & Harcourt, 2007, p. 120). According to the *Manual*, the median Round 3 score was submitted to ADE as the recommended cut score, however, ADE was not obligated to adopt the recommended score. Table 1 contains the range of recommended scores (from round 3), the recommended (median) score (from round 3) and the final cut score for all proficiency levels for each grade.

Of the 20 cut scores, only 3 (2nd grade Intermediate and Proficient; 4th grade Emergent) fall within two points of the recommended score. Ten of the twenty final cut scores are ten or more points discrepant from the recommended score. Notably, all final kindergarten cut scores fall *above* the judges' recommended score – including a 15 point discrepancy at the proficient cut score; and all 7th and 9th grade scores fall *below* the recommended score – by as much as 33 points (9th grade Basic). Thus, to be judged proficient, kindergarten children must score 15 points higher than the score recommended by the standard-setting procedure and high school students may score 14 points lower than the recommended score.

The AZELLA *Manual* gives no rationale for the substantial differences between the recommended scores and the final cut scores. It also fails to document how final cut scores were determined and the expertise of the person or persons who set the final scores. The substantial differences between the final cut scores and the recommended scores call into question the validity of the entire standard-setting process. If the test developers were confident in the standard-setting process, why were the recommended cut scores rejected? Further, kindergarten cut scores again follow a different pattern than the cut scores for older children. Raising the bar for kindergarteners above the judges' recommendations and lowering the bar for older students could result in over-identification of kindergarten students, under identification of older students, and inappropriately accelerated reclassification of students (reclassifying students as proficient before they are truly proficient). It could also produce artificially inflated reclassification rates for 1st graders as many may have actually met recommended criteria for proficiency at the kindergarten level if the cut score was set too high.

Empirical Evidence of Cut Score Validity

The *Standards* encourage the use of empirical data when cut scores define "categories with distinct substantive interpretations" (*Standards*, AERA et al., 1999, p. 60). Cut scores can be empirically set by analyzing how well they predict concurrent or future performance on a criterion task. "Ideally, cut scores delineating categories [of proficiency levels among school age students] would be based on research demonstrating empirically that pupils in successive score ranges did most often benefit more from the respective treatments to which they were assigned than from the alternatives available (*Standards*, AERA et al., 199, p. 53).

One way to empirically analyze the validity of cut scores with the limited AZELLA data available for independent review is to compare where cut scores fall on a normal curve for each grade level. Although using norm-referenced assessment to determine English language proficiency is ill advised (Mahoney et al., 2010), analyzing normative cut score data provides another point of information in the validation process. Table 2 provides estimated (rounded to the nearest tenth) percentile ranks for Basic, Intermediate and Proficient cut scores calculated from 2006 final form operational data provided in Table 4.1 of the AZELLA *Manual* (p. 17-18).

Table 2. Percentile ranks for Basic, Intermediate, and Proficient cut scores based on 2006 operational administration of the AZELLA.

Grade	Basic	Intermediate	Proficient
K	24.00	69.00	95.00
1	22.00	65.00	93.00
2	12.00	42.00	78.00
3	18.00	50.00	78.00
4	14.00	43.00	74.00
5	15.00	39.00	68.00
6	13.00	39.00	68.00
7	15.00	42.00	68.00
8	11.00	36.00	69.00
9	10.00	39.00	69.00
10	10.00	40.00	72.00
11	4.00	39.00	69.00
12	5.00	30.00	74.00

Percentile ranks indicate how many children out of 100 scored at or above the proficiency level for that grade. For example, the proficient cut score for kindergarten falls at the 95th percentile. This means that to score in the proficient range, a kindergartener must score better than 95 out of 100 other kindergarterns who took the AZELLA. The cut score falls at a similar percentile for 1st grade. In the 2nd through 12th grades, however, cut scores fall at a substantively lower normative level (68th to 78th percentile).

The presence of these normative differences across grade levels is problematic because they are unreported and unexplained. The 13 to 15-point difference between the kindergarten/1st grade normative levels and second grade level further supports the contention that cut score validity is suspect. In light of the subtantial variation in judges' cut score recommendations and the setting of final cut scores substantially different than the judges' recommendations, these normative differences in cut score placements demands an explanation. Requiring kindergarteners to have relatively stronger English language skills to pass the AZELLA may under-identify proficient kindergarteners resulting in many children who are ready for manistream classes not having exposure to mainstream classrooms (and daily discourse with

English-proficient peers). These differences have implications for interpreting the efficacy of the AZELLA's reclassification rates 1st through 3rd grades. If children are over-identified in kindergarten, they may also exit the program quickly (because their English language skills are already proficient) resulting in artificially high reclassification rates in the primary grades. Unfortunately, it is difficult to know to what extent this is true because ADE inexplicably does not report reclassifcation rates for kindergarten to first grade.[3] Artifically elevated reclassification rates would obfuscate efforts to accurately evaluate the efficacy of Arizona's SEI program.

Conclusion

Validly assessing the English proficiency of ELL children is critical for providing them with equal educational opportunity. Given the substantial numbers of children taking the AZELLA and the critical importance of establishing English proficiency early in children's school career, rigorously demonstrating the validity of AZELLA cut scores is imperative. The AZELLA test developers have failed to provide convincing evidence that they have met widely-established standards for establishing the cut scores used to determine which ELL children receive English language support and which are educated in mainstream classes. To date, there is no publically available empirical evidence that AZELLA cut scores accurately differentiate those children who need English language support and those who do not.

References

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

American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

Arizona Department of Education & Harcourt Assessments, Inc. (2007). *Arizona English Language Learner Assessment.* Phoenix, AZ: Arizona Department of Education [ADE] & Harcourt Assessments, Inc. [Harcourt].

Bourque, M. L. (2009). *A history of NAEP achivement levels: Issues, implementation, and impact 1989-2009.* Washington, D.C.: National Assessment Governing Board.

Brown, W. (2000). Reporting NAEP by achievement levels: An analysis of policy and external reviews. In M. L. Bourque, & S. Byrd, *Student performance standards on the National Assessment of Educational Progress: Affirmations and improvements.* Washington, D.C.: National Assessment Governing Board.

Hambleton, R.K., & Pitoniak, M.J. (2007). Setting performance standards. In R. L. Brennan *Educational Measurement, Fourth Edition*. pp.433-470: Praeger: Westport, CT.

Lesaux, N.K., Rupp, A.A., & Siegel, L.S. (2007). Growth in reading skills of children from diverse linguistic backgrounds: Findings from a 5-year longitudinal study. *Journal of Educational Psychology, 99*, 821-834.

Mahoney, K., Haladyna, T., & Macswan, J. (2010). The need for multiple measures in reclassification decisions: A validity study of the Standford English Language Proficiency Test. In P. Gandara, M. Hopkins, P. Gandara, & M. Hopkins (Eds.), *Forbidden Language: English learners and restrictive language policies* (pp. 240-262). New York, New York: Teachers College Press.

Meisels, S.J. (1986). Testing four- and five-year-olds: Response to Salzer and to Shepard and Smith. *Educational Leadership, 44*, 90-92.

National Association for the Education of Young Children (NAEYC) & National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE) (2003). *Early childhood curriculum, assessment and program evaluation: Building an effective, accountable system in programs for children birth through age 8 (position statement with expended resources)*. Retrieved from www.naeyc.org/about/positions/pdf/pscape.pdf Washington DC: National Association for the Education of Young Children.

Nagle, R.J. (2007). Issues in preschool assessment. In B.A. Bracken & R.J. Nagle (Eds.), *Psychoeducational assessment of preschool children: Fourth edition*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Reckase, M. D. (2000). A survey and evaluation of recently developed procedures for setting standards on educational tests. In M. L. Bourque, & S. Byrd (Eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmation and miprovmeents.* Washington, DC: National Assessment Governing Board.