PEARSON

technical report

Assessing English Language Proficiency: Using Valid Results to Optimize Instruction

Agnes Stephenson, Ph.D., Psychometrician

Diane F. Johnson, Senior Assessment Specialist

Margaret A. Jorgensen, Ph.D., Senior Vice President of Product Research and Innovation

Michael J. Young, Ph.D., Director of Psychometrics and Research Services

Paper presented at the 81st Annual Meeting of the
California Educational Research Association (CERA)

November 14, 2003 *(Revision 2, January 2004)*

Assessing English Language Proficiency: Using Valid Results to Optimize Instruction

Introduction

The *No Child Left Behind Act* of 2001 (NCLB) has focused increased attention on the appropriate assessment of English language learners (ELL students) in U.S. public schools. NCLB specifically requires that English proficiency be assessed and that ELL students participate in a standards-based English language testing program from day one. For the more than 4 million ELL students in K–12 public education across the United States, the federal expectation is that they will be able to function in regular classrooms within three years as proficient speakers, readers, and writers of English.

There is much diversity among ELL students in U.S. public schools. Language, culture, and economic differences are evident, but also influencing students' acquisition of English is their native language literacy and education experience, motivation, and opportunity to learn, as well as the ability of teachers to meet the individual learning needs of these students. English language proficiency measures must have a meaningful relationship with requirements of the classroom culture.

From the perspective of testing, the challenges are to:

- understand the complexities of acquiring English language proficiency;
- determine how assessments can support effective teaching;
- build reliable and valid assessments that are most likely to elicit critical evidence of English language acquisition;
- document the psychometric properties of these items and tests; and
- validate "proficiency" so that there is a rational link between the English proficiency of ELL students and native English speakers.

The purpose of this report is to describe an assessment strategy that fulfills the requirements of NCLB and does so by supporting effective instruction for the complex population of ELL students in U.S. K–12 public schools.

Understand the Complexities of Acquiring and Assessing English Proficiency

For teachers of ELL students in U.S. public schools, the goal of helping their students attain English language proficiency is inherently complex. Proficiency is often referred to as if it were a uniform state attainable by most students in a specifically defined time frame. This notion of a one-dimensional, global language proficiency is, however, only a theoretical construct. Attaining language proficiency is not a neat process—language skills (listening, speaking, reading, and writing) can be acquired at very different rates with spikes and plateaus during the learning.

So, when is a student considered proficient in English? An ELL student does not need to be as fluent as a native speaker to be considered proficient. Rather, an ELL student needs to be proficient enough to participate in regular classes conducted in English without requiring significant English language support. Furthermore, the proficient ELL student should have a good chance of success in those classes.

During the language acquisition process, immigrant children often achieve conversational fluency within one to two years, but their ability to reach grade-appropriate academic proficiency can take up to five years or longer. For these children, language can generally be divided into social language and academic language. Jim Cummins (1979) identified this linguistic phenomenon as basic interpersonal communicative skills (BICS) and cognitive academic language proficiency (CALP).

The complementary relationship between academic and social English language skills supports the structure of an English Language Proficiency test called for in NCLB. In general, the content areas of reading, writing conventions, and writing represent academic skills, and those of listening and speaking represent social skills. Accurately assessing these two aspects of English language skills, academic and social, provides a clear picture of a student's overall English proficiency. A student's level of language proficiency is directly related to his or her success in regular classrooms.

Determine How Assessments Can Support Effective Teaching

English language proficiency tests best serve the purposes of testing when they reflect both research and excellent teaching practices. Good testing and good teaching are grounded in research from the field of second language acquisition. Decisions about test design—test constructs and item constructs—based on quantifiable evidence from current research yield stronger assessment instruments. At the same time, this research also provides a link between assessment and the thinking about what constitutes current best practices in instruction.

When experienced teachers of ELL students who are knowledgeable about the students and curriculum to be tested, provide the basic content for ELP tests, those tests are much more likely to be age-appropriate and in line with curriculum. So, teachers who have helped author English language development standards are in the best position to know which standards are the most important ones to be tested, and these teachers are also the best people to furnish content for testing the standards.

What Standards Represent English Language Acquisition?

Most states are preparing or have completed content standards in English language development. However, content standards published by Teachers of English to Speakers of Other Languages (TESOL) are commonly referred to as the "national model" for English language acquisition. The TESOL standards represent a widely accepted framework upon which many states are building their specific standards.

The *Stanford English Language Proficiency Test* (Stanford ELP), first published by Pearson in 2003, assesses students' general acquisition of English by measuring:

- **Listening, Writing Conventions,** and **Reading** using multiple-choice items;
- **Writing,** using an open-ended direct writing assessment; and
- **Speaking,** using a performance test.

At the same time, state content standards are addressed by Stanford ELP. The development of Stanford ELP began with comprehensive reviews and careful analyses of current state and district English language curricula and educational objectives. Pearson also reviewed current second language acquisition research and considered the trends and directions established by national professional organizations.

In order to meet the expectations of today's professionals and school officials responsible for teaching ELL students, Pearson took all the research into account when developing Stanford ELP. Test blueprints based on Pearson's analyses address the language skills to be assessed at each grade level. The blueprint for each language skill outlines the topics to be covered, the instructional standards associated with each topic, and the proportion of test content to be devoted to each topic.

After development of the Stanford ELP test blueprint, national experts reviewed it. These experts commented on all aspects of the blueprints: (1) the instructional objectives included at each test level, (2) the test level at which the objectives were introduced, and (3) the proportion of test content developed for each objective. The blueprints were then revised, and the final blueprints became the framework upon which the test forms were constructed.

The final consideration in developing Stanford ELP was what specific language should be tested. Language, especially spoken language, is fluid and ever-changing. It is essential that ELP tests focus on fresh, vibrant language, the language that is actually used in classrooms and the community. The goal must be to measure what matters to students—to have teachers say, "What I really like about this test is its ability to test language that students really need."

As shown in Table 1, the topics and vocabulary used in each of the four levels of the Stanford ELP are age-appropriate for the corresponding grade range.

Table 1. Stanford ELP Test Levels and Corresponding Grade Ranges

Stanford ELP Test Level	Grade Range
Primary	K–2
Elementary	3–5
Middle Grades	6–8
High School	9–12

For each test level, particular care must be taken to ensure that both the language proficiency and the chronological or developmental status of students are taken into consideration. It is most important that all students are fully engaged in the content. To accomplish this, students' interests must be addressed.

Build Reliable and Valid Assessments that Measure English Acquisition

Following the review of all the appropriate informational materials (publications of national professional organizations, district and state standards, and input from teachers and educators), Pearson developed a set of test specifications (blueprints) for Stanford ELP. The blueprints include the number of test levels necessary for complete coverage across all grades, the content areas to be assessed, and the instructional standards to be included in each content area. The blueprints specify the actual subtests to be included at each test level and the instructional standards to be assessed across the grade levels. The number of items that should assess each of the instructional standards to ensure breadth of content and reliability of assessment were also included in the blueprint. At every test level, the specifications required that all forms of the test be parallel in terms of content, standards, and difficulty and that each form be unique. All items for Stanford ELP were newly written, and each item appears on only one form.

Test specifications, or blueprints, are the cornerstone of quality form construction. In addition, they are the content specialists' plan for test construction. They include any of the following:

- Instructional standards to be tested,
- Number of items per test form
- Acceptable range of *p*-values
- Number and/or percentage of items within desired *p*-value ranges
- Number and/or percentage of items within a specified level of knowledge
- Other specified psychometric properties

Alignment

To align Stanford ELP with instructional standards and curricula taught at the respective grade levels, Pearson utilized the criteria identified by Norman L. Webb (2002) as a model during the planning stages of development. The alignment criteria identified by Webb are:

1. *Categorical Concurrence.* This criterion is met when the same or consistent categories of content appear in both instructional standards and assessments.

2. *Depth-of-Knowledge Consistency.* This criterion is met when test questions presented to the students on the assessment are as cognitively demanding as what students are expected to know and do as stated in the instructional standards.

3. *Range-of-Knowledge Correspondence.* This criterion is met when the comparable span of knowledge expected of students by an instructional standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities.

4. *Balance of Representation.* This criterion is met when the degree to which the emphasis given to one instructional standard on the assessment is comparable to the emphasis given to the other instructional standards.

5. *Source-of-Challenge*. This criterion is met when the primary difficulty of the assessment items is highly related to students' knowledge and skill with the content area as represented in the instructional standards.

Once content and test construction experts reviewed and approved the completed test specifications, the development of item specifications for Stanford ELP began. The item specifications included the following information:

- Item format
- Content restrictions
- Option requirements
- Sample items

The final item specifications became the framework that drove the item development process.

Content specialists and Pearson-trained item writers created pools of items in accordance with the item specifications in their areas of expertise. These writers included practicing teachers who had a solid base of knowledge and teaching experience in ESL. These teachers were able to ensure appropriateness of topic, vocabulary, and language structure for items at each grade level.

Item writers were trained in the principles of item development and item review procedures. They received detailed specifications for the types of items they were to write, as well as lists of objectives and examples of both good and bad items. As item writers wrote and submitted items, the items also went through an internal process that included reviews by content experts, psychometricians, and editorial specialists.

The items developed present engaging, accessible content. As the following examples demonstrate, it is possible to measure academic and social language skills in engaging ways. The examples shown below demonstrate how functional academic language has been incorporated into Stanford ELP items. The examples are similar to items found in Stanford ELP Listening, Speaking, Writing, and Writing Conventions subtests.

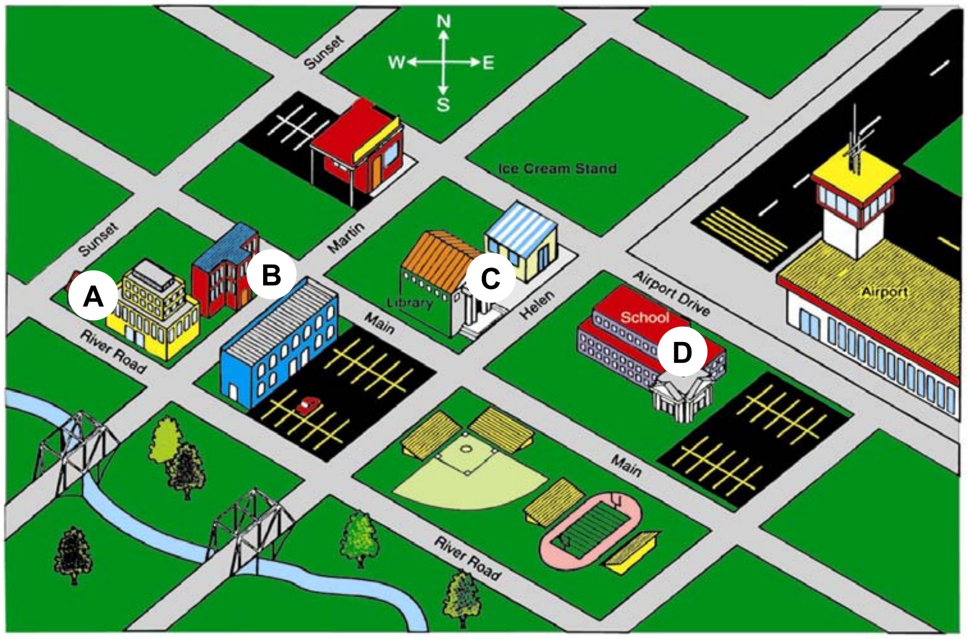
EXAMPLE 1—Listening, Middle Grades Level

Question

Where will you work on your group science project tomorrow?

Listening script (Dictated only)

Listen to the phone message from your classmate from school. Hi, this is Julie. I hope you got the science books from the library. Let's meet at 2:00 o'clock tomorrow at my house and then walk over to Sam's—his house is at the corner of Sunset and River Road. We can finish our project on recycling there. Don't forget—we've got to turn in all our work to Mr. Thomas at school next Thursday.



Answer options

- ***A*** *
- ***B***
- ***C***
- ***D***

After listening to and reading the question, the student looks at the graphic above while listening to the script. The student then decides which option, labeled A, B, C, or D in the graphic, is correct and marks it on the answer document. The context of Example No. 1 is a group of students working together on a science project. Thus, the item requires the test taker to comprehend and synthesize functional academic language that is needed when students work together cooperatively (i.e., *get science books from the library … let's meet at 2:00 o'clock … we can finish our project on recycling there … we've got to turn in all our work to Mr. Thomas at school next Thursday*).

EXAMPLE 2—Speaking (Social Interaction), High School Level

Items grounded in academic contexts are found throughout the Speaking subtest. This example, typical of a conversation that might take place at school, addresses sociolinguistic competence. The stimulus is the first part of a conversation, a single sentence, to which students respond with an appropriate rejoinder.

Prompt

I don't remember how many pages the teacher wanted us to read.

Possible student responses

I know.

Neither do I.

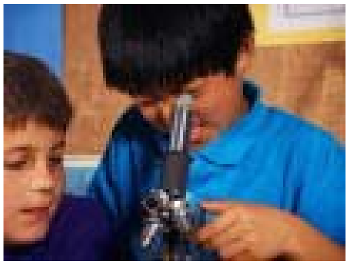
These are the pages.

EXAMPLE 3—Writing, Primary Level

Academic situations are portrayed in the graphics of writing prompts, as shown in this example. The picture of two boys with a microscope can elicit from students a wide range of language that will demonstrate their understanding of schools and classroom procedures, teaching and learning, and students and their behavior.

Prompt

Directions: Look at the picture. Write about what you see in the picture. Tell a story about this picture.



EXAMPLE 4—Writing Conventions, Elementary Level

Some Writing Conventions items exemplify tasks that are commonly used for studying language. One such task is shown in this example, which requires students to read a dictionary definition and then apply their understanding.

> **DICTIONARY**
>
> **Cu•ri•ous (kyo͝or' ē-əs)** *adj.* **1.** Very interested in getting information or knowledge.

Based on what the dictionary says about this word, which sentence is correct?

A *Carol is a curious student so she asks a lot of questions in class.* *

B *Carol is a curious student so she always finishes her homework on time.*

C *Carol is a curious student so she is going to be in the fifth grade next year.*

D *Carol is a curious student so she likes doing math problems better than reading.*

Bias/Sensitivity Review

The Advisory Board of ESL experts reviewed all multiple-choice items after they were assembled into field test forms. There were two aspects to the process: 1) a content review and 2) a bias/sensitivity review. For purposes of an ESL test, evaluating for bias/sensitivity is seeing that no language group is advantaged over another and that there is no material that would be offensive to any cultural or religious group. The panel suggested changes to items, directions, and format. Following the Advisory Board's review for bias, further revisions were made to the Stanford ELP items.

Document the Psychometric Properties of These Items and Tests

Field Testing of Items

Following their development and review, Stanford ELP multiple-choice test items were assembled into multiple parallel field test forms at each of the four test levels. The results for all field test items were analyzed using traditional item-analysis methods, which resulted in the types of information presented in Table 2 and described below.

Table 2. Data Obtained from the Spring 2002 Field Testing of Items—Form A

Stanford ELP Test Levels, Forms, & Subtests		Cronbach's Correlation Coefficient Alpha[1]	Standard Error of Measurement (SEM)[2]	Mean *p*-Value[3]	Median Point-Biserial Correlation Coefficient[4]
Primary					
Form A	Overall	0.88	2.48	0.68	0.382
	Reading	0.66	1.75	0.57	0.402
	Listening	0.52	1.30	0.84	0.332
	Writing Conventions	0.39	1.33	0.62	0.420
Elementary					
Form A	Overall	0.91	2.91	0.55	0.420
	Reading	0.77	1.48	0.64	0.449
	Listening	0.61	1.86	0.47	0.341
	Writing Conventions	0.76	1.79	0.55	0.497
Middle Grades					
Form A	Overall	0.93	2.68	0.55	0.422
	Reading	0.84	1.49	0.51	0.505
	Listening	0.68	1.77	0.60	0.403
	Writing Conventions	0.70	1.48	0.52	0.405
High School					
Form A	Overall	0.93	2.66	0.55	0.439
	Reading	0.83	1.35	0.55	0.517
	Listening	0.48	1.72	0.58	0.362
	Writing Conventions	0.83	1.63	0.51	0.529

[1] **Cronbach's Correlation Coefficient Alpha.** This value reveals the internal consistency (inter-item correlation) of the test.

[2] **Standard Error of Measurement (SEM).** This value is a way of describing the variability of test scores. The standard deviation of the distribution of a group's test scores is the standard error of measurement. A high/low SEM means that a student's observed score is less/more likely to represent his or her "true" score.

[3] **Mean *p*-Value.** The average of *p*-values. *p*-Value is the percentage of respondents answering the item correctly and represents the difficulty of the item. Mean *p*-value represents the average difficulty level of items on a particular subtest.

[4] **Median Point-Biserial Correlation Coefficient (Item Discrimination Index).** The point-biserial correlation coefficient represents the difference between the performance of the upper group on a particular item and the performance of the lower group on the same item. This value reveals how closely students' performance on the item relates to their performance on the entire test. The median is the "middle" value.

Non-native English speaking students participating in the Fall 2002 tryout of forms represented the fifteen language groups shown in Table 3.

Table 3. World Languages Included in the Fall 2002 Tryout of Forms Sample

Arabic	Hindi	Polish
Armenian	Japanese	Portuguese
Farsi	Khmer	Russian
Filipino	Korean	Spanish
Haitian (Creole)	Mandarin	Vietnamese

Pearson conducted validation research by administering Stanford ELP to two groups of students: Non-native English speakers (ELL students) and native English speakers. The summary statistics resulting from this study—N, mean, and standard deviation—are presented in Table 4 for each of the three multiple-choice subtests, Reading, Listening, and Writing Conventions.

The most fundamental function of validation research of tests of English language proficiency is to demonstrate that the language being assessed is the language actually used by native English speakers. The data presented in Table 4 provide evidence that native speakers performed much better on Stanford ELP than non-native speakers, demonstrating that the language tested is that of native English speakers.

Analysis of variance (ANOVA) calculations were conducted for each of the three subtests—Reading, Listening, and Writing Conventions—at each of the four test levels. The ANOVA results showed that the native English speakers scored significantly higher than the non-native speakers on all subtests and at all test levels.

Table 4. Performance Differences between Non-Native and Native English Speaking Students Taking the Stanford ELP

	Non-Native Speakers			Native Speakers			F Value	Pr>F
	N[5]	Mean[6]	SD[7]	N	Mean	SD		
Listening								
Primary	1151	14.81	3.32	1511	17.00	2.22	414.69	0.0001
Elemen.	885	11.81	4.13	1100	15.59	3.18	531.34	0.0001
Mid. Grds.	430	11.01	3.79	900	17.16	2.11	1437.72	0.0001
High Sch.	187	11.09	3.33	300	16.72	2.19	505.44	0.0001
Writing Conventions								
Primary	1151	9.54	4.24	1511	11.79	4.88	155.70	0.0001
Elemen.	885	12.52	4.57	1100	16.91	3.09	646.81	0.0001
Mid. Grds.	430	11.40	4.03	900	20.54	3.00	2144.10	0.0001
High Sch.	187	12.17	4.12	300	19.98	3.00	585.19	0.0001
Reading								
Primary	1151	9.78	3.96	1511	12.33	4.59	227.15	0.0001
Elemen.	885	11.91	4.22	1100	16.21	3.64	593.15	0.0001
Mid. Grds.	430	11.71	4.35	900	20.71	3.56	1606.52	0.0001
High Sch.	187	12.88	4.50	300	21.43	3.44	557.73	0.0001

The Chronbach's Alpha coefficients of reliability for the four Stanford ELP test levels ranged from 0.92 to 0.94, which are very high values. These values estimate the reliability of a composite score (see Table 5).

Table 5. Reliabilities by Test Level

Level	Reliability
Primary	.92
Elementary	.94
Middle Grades	.94
High School	.92

Combining the raw scores of certain Stanford ELP subtests as shown in Table 6 yields information about a student's skills in four important aspects of English language proficiency: Academic Language, Social Language, Language Comprehension (receptive language), and Productive Language. For example,

[5] **N.** This value is the number of examinees who took each subtest.

[6] **Mean.** This value is the average of the raw scores for each subtest. The mean is determined by adding all students' scores on the subtest and dividing by the total number of students.

[7] **Standard Deviation (SD).** This value is the square root of the average of the sum of squared deviation around the mean.

Table 6 shows that combining scores earned on the Listening and Speaking subtests yields a Social Language "subscore," which is a measure of a student's social language skills. As another example, Writing and Speaking combined yield a Productive Language subscore.

Table 6. Stanford ELP Subtests Comprising each English Language Subscore

▼Subscore►	Academic Language	Social Language
Language Comprehension	Reading	Listening
Productive Language	Writing Writing Conventions*	Speaking

*The Writing Conventions subtest score is included in only the Academic Language subscore.

It is important to examine the correlations between different pairs of subscores for each of the four Stanford ELP test levels (see Tables 7a–7d). An examination of the data shows that these correlations range from moderate to high.

Construct validity is evidenced by the high correlations between the Academic Language and Language Comprehension subscores as contrasted to the moderate correlations between the Academic Language and Social Language subscores, a pattern that holds across each of the Stanford ELP test levels (see Tables 7a–7d). The higher correlations reflect the similarity in meaning of the subscores; Academic Language and Language Comprehension both include Reading subtest scores. However, each of the four subscores is associated with a moderate correlation coefficient, demonstrating the power of each subscore to measure unique language skills.

Table 7a. Correlation Coefficients for Stanford ELP Subscore Combinations, Primary Test Level

Subscore	Academic Language	Language Comprehension	Productive Language	Social Language
Academic Language	1.00	.92	.62	.52
Language Comprehension		1.00	.59	.58
Productive Language			1.00	.96
Social Language				1.00

Table 7b. Correlation Coefficients for Stanford ELP Subscore Combinations, Elementary Test Level

Subscore	Academic Language	Language Comprehension	Productive Language	Social Language
Academic Language	1.00	.91	.76	.61
Language Comprehension		1.00	.67	.63
Productive Language			1.00	.93
Social Language				1.00

Table 7c. Correlation Coefficients for Stanford ELP Subscore Combinations, Middle Grades Test Level

Subscore	Academic Language	Language Comprehension	Productive Language	Social Language
Academic Language	1.00	.92	.76	.66
Language Comprehension		1.00	.69	.68
Productive Language			1.00	.95
Social Language				1.00

Table 7d. Correlation Coefficients for Stanford ELP Subscore Combinations, High School Test Level

Subscore	Academic Language	Language Comprehension	Productive Language	Social Language
Academic Language	1.00	.92	.74	.63
Language Comprehension		1.00	.69	.68
Productive Language			1.00	.95
Social Language				1.00

Validate "Proficiency" to Link ELL Student and Native Speaker Performance

Important external validity evidence is whether the classification of "proficient" for an ELL student who took the ELP test has meaning relative to the same student's performance on a test intended for native English speakers. For Stanford ELP, the data is clear and compelling. In order to examine this relationship, Pearson administered the *Stanford Achievement Test* Series, Ninth Edition (Stanford 9) Reading Comprehension subtest to a broad range of ELL students who also took Stanford ELP.

Table 8 presents the scores earned on the Stanford 9 Reading Comprehension subtest by second, third, and fourth grade ELL students classified as "non-proficient" and "proficient" by the Stanford ELP. The data show how language proficiency as measured by one instrument is confirmed by results obtained from a separate, equally valid, instrument.

Table 8. Scores Received on Stanford 9 Reading Comprehension Subtest by ELL Students Classified as Non-proficient and Proficient on Stanford ELP

	Non-Proficient			**Proficient**		
Grade	Mean Raw Score	Mean Scaled Score	Mean Percentile Rank	Mean Raw Score	Mean Scaled Score	Mean Percentile Rank
2	17.3	507	20	25.6	572	61
3	11.7	562	20	22.3	618	64
4	15.4	595	22	23.7	657	74

Standard Setting

The Modified Angoff Procedure (Angoff, 1984) was used to produce the recommended cut scores for the Stanford ELP. This standard setting method has a long and respected history in similar applications for both educational and professional certification assessments. The use of the procedure by Pearson provided a systematic technique for eliciting judgments from panels of experts (i.e., standard setting committees), producing consensus among these experts, and quantifying the results of the judgments. The Modified Angoff Procedure is widely recognized as the simplest method to use for setting performance level cut scores (Norcini, et al., 1988; Shepard, 1980). Moreover, research has shown that it produces ratings with better reliability and smaller variability among the ratings of judges than other standard setting procedures (Andrew and Hecht, 1976; Brennan and Lockwood, 1980; Cross, et al., 1984; Poggio, et al., 1981; Skakun and Kling, 1980). The Modified Angoff Procedure incorporates an appropriate balance between statistical rigor and informed opinion.

The process began with the establishment of standard setting committees. The five standard setting committees assembled by Pearson included outstanding ESL teachers of kindergarten through grade 12 from across the United States.

The training and experience of the members of the standard setting committees are essential in establishing the validity of their ratings (American Educational Research Association, et al., 1999). These committees were thoroughly trained in the procedures and methodology.

Five separate standard setting committees were established to identify cut scores for grades K–12. Each committee was responsible for establishing cut scores for each grade within an assigned group of grades as shown in Table 10. Each committee member was asked to set standards for the Stanford ELP.

Table 10. Standard Setting Committees, Grade Levels Assigned, and Stanford ELP Level(s) Reviewed

Committee	Grades Assigned	Stanford ELP Level(s) Reviewed
A	K	Primary
	1	
B	2	
	3	Elementary
C	4	
	5	
D	6	Middle Grades
	7	
	8	
E	9	High School
	10	
	11	
	12	

The judges' primary task was to "rate" each item on the Stanford ELP in terms of how a Basic, Intermediate, or Proficient student *should* perform on it, rather than how they do perform or will perform.

For a given multiple-choice item (Listening, Writing Conventions, and Reading subtests), the rating process involved answering the following questions:

1. What percentage of "Basic" performing students in the grade should correctly answer the item?
2. What percentage of "Intermediate" performing students in the grade should correctly answer the item?
3. What percentage of "Proficient" performing students in the grade should correctly answer the item?

Each committee member used Pearson's online standard-setting tool to record his/her judgment.

For a given open-ended or performance-based item (Writing and Speaking subtests, respectively), the rating process involved answering the following questions:

1. What average number of rubric points should a "Basic" performing student earn on this item?
2. What average number of rubric points should an "Intermediate" performing student earn on this item?
3. What average number of rubric points should a "Proficient" performing student earn on this item?

Judges were reminded that they were to rate an item based on only the group of students within a performance level stratum, rather than the total group. In other words, for a Basic rating, judges were to identify the percentage that should correctly answer the item based on only the group of students whose *best performance* is Basic; not the percentage based on the total group's performance.

Comparison of Cut Scores Across the Grades

In order to verify that cut scores ascended with the grade level, the raw cut scores were first converted to a proportion of the total possible points. The total possible points for grades K through 5 was 102, while that for grades 6 through 12 was 110.

Pearson identified four cut score points for the overall test at each grade (see Figure 1). An impact analysis was conducted after the standard setting procedure to confirm these cut scores. The final cut scores for this test were based on both teacher judgment and psychometric analysis. Ideally, states using the Stanford ELP will set their own proficiency levels and cut scores to reflect their needs and policies.

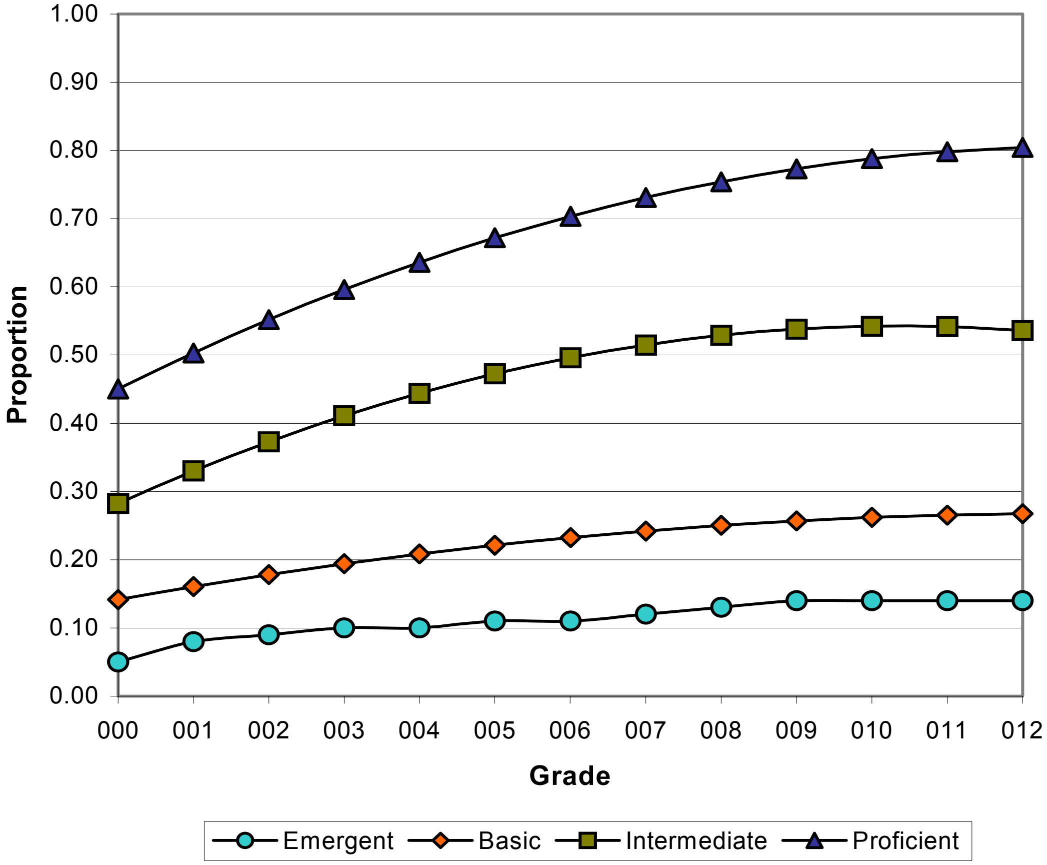


Figure 1. Recommended Proportion Cut Scores Across Grades for Four Performance Levels—Smoothed

Conclusion

The data presented in this report serve as evidence of validity for the *Stanford English Language Proficiency Test* (Stanford ELP). In addition to the four basic language domains (Listening, Speaking, Reading, and Writing), it reliably assesses and reports separate "subscores" for, four distinct aspects of language understood by native speakers: Academic Language, Social Language, Language Comprehension, and Productive Language. Results reported in this way are a new and powerful tool for tailoring instruction to English language learners. In light of requirements set forth in the *No Child Left Behind Act* of 2001 (NCLB)—that ELL students must be proficient in English after three consecutive years in public schools—the targeted information contained in Stanford ELP scores and subscores are effective tolls for school systems to use to achieve this goal.

References

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

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service. (Reprint of chapter in R. L. Thorndike (Ed.). (1971). *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement, 36*, 45-50.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4*, 219-240.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examination. *Journal of Educational Measurement, 21*, 113-129.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism, No. 19*, 121–129.

Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. *Review of Educational Research, 59*, 315-328.

Ganopole, S. J. (1980). "Using performance and preference data in setting standards for minimum competency assessment programs," in R. M. Jaeger and C. K. Tittle (Eds.), *Minimum competency achievement testing*. Berkley, CA: McCutchan, 406-418.

Hambleton, R. K., Powell, S., & Eignor, D. R. (1979). "Issues and methods for standard setting" in *A practitioner's guide to criterion-referenced test development, validation, and test score use: Laboratory of Psychometric and Evaluative Research Report No. 70* (2nd ed.). Amherst, MA: School of Education, University of Massachusetts.

Jaeger, R. M. (1982). An iterative structures judgment process for establishing standards on competency tests: Theory and applications. *Educational Evaluation and Policy Analysis, 4*, 461-475.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practices, 10*, 3-10.

Meskauskas, J. A. (1983). *An assessment of the state of the art of standard setting methodology: Characteristics of improved approaches and two new methods*. Paper presented at the Annual Meeting of the American Educational Research Association.

Millman, J. (1973). Passing scores and test lengths for domain-referenced measures. Review of *Educational Research, 43*, 205-216.

Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher, 18*, 5-9.

Mills, C. N., & Melican, G. J. (1988). Estimating and adjusting cutoff scores: Features of selected methods. *Applied Measurement in Education, 1*, 261-275.

Norcini, J. J., Shea, J. A., & Kanya, D. J. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement, 25* (1), 57-65.

Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1981). *An empirical investigation of the Angoff, Ebel, and Nedelsky standard setting methods*. Paper presented at the annual meeting of the American Educational Research Association.

Shepard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4*, 447-467.

Skakun, E. N., & Kling, S. (1980). Comparability of methods for setting standards. *Journal of Educational Measurement, 17*, 229-235.

Webb, N. L. (1997a). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 8). Washington DC: Council of Chief State School Officers.

Webb, N. L. (2003). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Madison, WI: University of Wisconsin-Madison.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Washington DC: Council of Chief State School Officers.

Zeiky, M., & Fremer, J. (1979). *Guidelines for developing and using minimum competency tests*. Paper presented at the Annual Meeting of the American Educational Research Association.

Additional copies of this and related documents are available from:
Pearson Inc.
19500 Bulverde Rd.
San Antonio, TX 78259
1-800-211-8378
1-877-576-1816 (fax)
http://www.hemweb.com/library/researchreports/index.htm