The Design and Analysis of Everett Alvarez High School's Placement Test for Students of English as a Second Language

Jeff Mattison, Designer
B1 Revised Project
June 20, 2005

This paper will present all of the relevant information on the second version of the placement test for Everett Alvarez High School's English as a Second Language department. Hereafter, the test will be known as the EAHSESL. The paper will be organized in the following way

1. Theoretical Foundations
2. Test Specifications
3. Statistical Analysis
4. Item Analysis
5. Suggestions for Further Research
6. Conclusion

The reader should keep in mind that this is a report of a test that is still in the development stage. The reliability and validity of the test have been investigated, but not established. Therefore, the concern of the reader should not be, "Are the test scores reliable and valid?" Instead, the concern of the reader should be, "Has the designer proven to me that he knows how to measure and explain the test scores' reliability and validity so that I know he will be able to handle the task when the scores are ultimately reliable and valid?" Having established a framework for the reader's evaluation of this report, the designer will explain the theories that support the test specifications.

Theoretical Foundations

There are several theories that informed the item construction and materials selection for this test. Two of them are from Swain (1984), start from somewhere and concentrate on content. Another involves schema theory according to Carrell and

Eisterhold (1983). The final theory is dynamic assessment, which is surveyed by Poehner and Lantolf (2003) and modeled by Guterman (2002). Before the test specifications are introduced, the designer will briefly describe the theories that informed them. During the explanation of the test specifications, the features of each theory will be highlighted.

Start from Somewhere

When Swain (1984) encourages that test design should, "start from somewhere," she means that each test designer should begin with a clear idea of what his/her test will measure. For the public school teacher, this means that a test should measure the skills and knowledge set forth by the state department of education. Although the test designer was an outsider to the EAHS ESL department and California State Curricula, he made an effort to link this test to the existing goals and objectives of each. Because the EAHS ESL department already administers the California English Language Development Test (CELDT) orally for its placement procedure (Appendix A), direct reading and indirect writing sections were designed to compliment the speaking and listening that would be assessed by the oral CELDT. The EAHSESL test contains stimulus material that is a part of the social science curriculum of the state (California Department of Education, 2002). The chosen stimulus material also has beneficial properties for the next criteria of Swain's (1984) criteria for test design, concentrate on content.

Concentrate on content

One of the motivating factors for the revision was the content. Swain (1984) advises that the content should be motivating, substantitive, integrated, and interactive. Content includes the material for the activities as well as the tasks for communicative

language acts. For motivating material, a newspaper article about EAHS students was chosen (Appendix B). Students love to see their own names in print. The subject of the article, a student-designed play about Salinas Valley farm workers, is relevant to the students. Most of the students who come to EAHS's ESL department are in farm working families or are a generation removed from one. Therefore, the students are very familiar with the content of the stimulus material.

Some material is new to the students like Maricela's narrative. Personal narratives are especially interesting; therefore they can motivate the test-taker's interest. Because Maricela's story is set up as a dictocomp listening activity, not all of the information is available upon first hearing it. This simulates an information gap activity and therefore a communicative language exercise. Appendix C illustrates how the content of the newspaper article and Maricela's story are integrated: They both deal with the theme of farm workers' personal narratives.

Finally, the content is interactive. The schema activating activity before the test gives students the opportunity to write their opinion about farm workers and theatre directors. The test could have included an interactive oral section using the newspaper article and/or personal narrative, but it was not feasible to coordinate an oral productive task with 29 students or a creative written task within the time constraints.

Schema Theory

A schema is an organizational pattern in the mind (Carrell & Eisterhold, 1983). Therefore a schema activator in testing is a device that alerts the test-taker to think of a subject in a certain way. A content schema is the information contained in the material

presented and a formal schema is the structure of the information presented (Bailey, 1998).

The EAHSESL test contains content schemata activating material. The pictures in the article illustrate actions and topics from the text. The questions that the test-taker is encouraged to answer before reading the article activate prior knowledge about theater directors and farm workers. It also contains formal schemata activating material. In the instructions read aloud and displayed on the TV screen from the DVD, the proctor announces that the test-taker is about to hear the story of a Salinas Valley farm worker. Students already familiar with the content schemata of such a story (i.e., the trip to USA, hardship, the better life sought, current struggle) can expect the details of the story to be presented in a certain order. Therefore, students may pay more attention to the exact details of Maricela's story. If the proctor just said, "listen twice, then write down as much as you can remember," test takers would have to figure out the formal schema on the first listening. This would leave only one more listening for the test-taker to focus on the specific details of the story.

In addition to formal and content schemata building, the test-taker must also utilize bottom-up (vocabulary, sentence recognition) and top-down (organizational familiarity, predicting structure) during the course of reading and listening comprehension activities such as those used in the EAHSESL test. Carrell and Eisterhold (1983) explain why

> Bottom-up processing ensures that the listeners/readers will be sensitive to information that is novel or that does not fit their ongoing hypotheses about the content or structure of the text; top-down processing helps the

> listeners/readers to resolve ambiguities or to select between alternative possible interpretations of the incoming data. (p. 557)

Therefore, giving the test taker the necessary formal and content schemata is only half of the process for enabling good performance. Short of being given interventionary assistance by the proctor, the test-taker needs to use top-down and bottom-up processing techniques for reading in order to use the schemata. Introducing schemata interactively or that encourages intramental speech approaches a dynamic method of measuring language ability. Before the test-taker has completed the test by him/her self, there is also a way to measure the amount of intervention that a proctor gives to each student to arrive at the correct answer? It is called dynamic assessment.

Dynamic Testing, Dynamic Assessment

The concept of dynamic assessment is a recent development in the United States, but has long been used in the former Soviet Union with children in special education due to the legacy of Vygotsky and his Zone of Proximal Development (ZPD). Because dynamic assessment is such a young field in the United States, it has taken many forms that Poehner and Lantolf (2003) have recently attempted to classify into measurements of dynamic testing (which lack personal intervention) and dynamic assessment (which include personal intervention).

In the EAHSESL test, the designer used a modified version of Guterman's (2002) metacognitive awareness guidance (MCAG) to add an element of sociocultural theory (SCT), mediation, which could assist the test taker to perform better. The MCAG encourages the test taker to use intramental speech to confirm the metacognitive strategies that the student has activated by completing the prompts that answer open-

ended questions about the subject of the upcoming reading passage. This procedure is an example of assistance from an "other" which Vygotsky identifies as cognitive activity taking place within ZPD. With the learning activity heightened in the test taker, Guterman hypothesizes that the student will perform better on the rest of the test.

The MCAG was not administered with any research design that would give the designer any authority to write about its effectiveness. The purpose for including the MCAG was to apply SCT to a language assessment device and make research-based improvements in the future. When she used an MCAG on a reading comprehension test for Israeli fourth graders, Guterman (2002) asserted that her MCAG functions within Vygotsky's ZPD and addresses the test-taker's potential development. Poehner and Lantolf (2003) would disagree with her on this point; they would categorize Guterman's MCAG as dynamic testing instead of dynamic assessment. The authors contrast the two constructs in this way

> Dynamic testing emphasizes the individual's or group's uptake of a predetermined repertoire of mediational means and as such attempts to discover the extent to which people will or will not change when offered pre-fabricated assistance. DA [dynamic assessment]… can neither limit nor prespecify the types of mediation required and must therefore allow the appropriate assistance to emerge in the dialogue between examiner and examinee as they jointly engage in concrete tasks. (p. 6)

The EAHSESL test is not a device of dynamic assessment according to Poehner and Lantolf (2003) because it was piloted to a large group under a short time frame of

design that didn't allow the proctor enough time to draft structured procedures for interacting one-on-one with 29 students.

Dynamic assessment must contend with the issues of practicality and reliability because it is usually administered one-on-one and the assistance given by the proctor is not consistent between each test taker. The difference between a test-taker's performance alone and with assistance measures his/her potential development. A student demonstrates improvement by requiring less outside assistance from an expert than s/he used in the previous test. This measurement of progress contrasts with the traditional construct of reliability that requires the same treatment to be applied to each student each time s/he takes the test. Future test developers will have to negotiate the best way to support the reliability of dynamic assessment devices. For the moment, the designer will forsake a discussion of alternate ways to measure the reliability of dynamic assessment devices and proceed with a description of the EAHSESL test specifications.

Test Specifications

Background

This test was designed as a placement device for migrant students entering the ESL program at Everett Alvarez High School (EAHS) in Salinas. It measures the students' reading, listening, and writing abilities that are required to participate in the language discourse of an ESL classroom at EAHS. In that domain, students must listen to classroom lectures given by the teacher, work in groups with other ESL students of a similar proficiency level, and read passages from a textbook that has semi-authentic material modified to an appropriate reading level.

The students who piloted the test were between 15 and18 years of age, from Mexico or another Central American country, and were in the ESL level 2 (beginner-intermediate) class taught by Michael Hedgpeth. A brief description of the ESL levels at EAHS can be found in Appendix A. The community in which the students live is a middle class neighborhood where many people are employed in the service or agricultural sectors. The content material of the test was chosen with these demographic features in mind.

The test has two norm-referenced sections. There is a direct test of reading comprehension and a direct test of listening and writing. The reading comprehension section contains a schema activating activity that is not scored and a reading passage with multiple-choice items that are objectively scored. The listening/writing section contains a writing task that is subjectively scored by two raters.

Prompt Attributes

Each test section contains directions that include a simple description of the task requested of the test-taker, as well as tips on good test-taking behaviors.

Reading Comprehension

In this section, students will read three questions that are designed to activate their prior knowledge about theatre, farming, and migrant workers. This portion of the test will not be scored, but it is included in the test to raise each student's metacognitive awareness about the subject of the reading passage in the next subsection. There, students will read a 274-word reading passage and answer 10 multiple-choice items about what they just read. This section will test the student's ability to

- Define unknown vocabulary using context clues and familiarity with idioms (a content schema)
- Summarize the main idea of the story (a formal schema)
- Choose appropriate registers of formality
- Rephrase idiomatic expressions
- Identifying the title of the story (a formal schema)
- Identify the referent of a third-person pronoun
- Rephrase an active voice sentence using the passive voice
- Construct subject-verb agreement

The text in this section should cover the experience of students who discover the history of the local area as it relates to agriculture and migrant workers. Preferably, the text should be authentic, contain less than 300 words, and of a 8th grade reading level on the Flesch-Kinkaid scale. If any of these criteria cannot be met, the text should be modified so that it fulfills all three requirements.

Listening/Writing

In this section, students will watch a DVD containing instructions that are projected on the TV screen as well as read by the test designer, a picture of a Salinas Valley farmworker, and two duplicate, 80-second recordings of a Salinas Valley farmworker telling her life story. There is a 27-second pause between the two recordings. After the DVD has been played, students are prompted to write down as many details of the story as they can remember. This section will measure the student's ability to

- Make inferences about data in the recording

- Compose sentences with proper syntax
- Recall information from short-term memory

The text for this section should contain information about the personal experience of a migrant worker in California. It may be authentic, but it may also be modified to meet length and language requirements. The text should take no more than 90 seconds to read. The language therein should be appropriate for a 8th grade reading level on the Flesch-Kinkaid scale. The text should be read in English by a person whose first language is a variety of Spanish spoken in the region from which most of the students come, in order to minimize comprehension difficulties with accent.

Response Attributes

Reading Comprehension

Students will first complete a three-word prompt with a one-sentence answer in response to three questions that are designed to activate prior knowledge about theatre, farming, and migrant workers. Students will answer in English. Then, students will whisper three affirmations that reinforce metacognitive strategies for finding information in the text and cognitive strategies for understanding the text. During the reading of the story, students are directed to underline important information. After the student has read the reading passage, he or she should write one sentence that summarizes the main idea of the story. Next, students will circle the letter next to the correct answer in the multiple-choice section.

Listening/Writing

Students are directed to refrain from taking notes when the first recording of the narrative is played. During the 27-second interval between recordings and the

reading of the second recording, the students may take notes. After the DVD stops, students must write, in English, as many details of the narrative as they can remember, using proper syntax and punctuation. There are 15 main points that students are expected to produce.

Sample Items

Reading Comprehension

Metacognitive Awareness Guide (MCAG)

Directions: The title of the story that you are about to read is called, "Students Write Their Own Play About Local Farm Workers".

In the story, the author describes:
- How the play was created
- What the play is about
- What the actors learned from the play

Before reading the story, please try to answer the following questions. They are not part of the test. They will help you to do better on the test.

1. How do you think a director creates a play? *In my opinion,* _____

_____

_____

Multiple-Choice Items

Directions:

1. Read the story.
2. Notice the numbers next to each line of the story. The line numbers will help you find a place in the story when you answer the questions after the story.
3. Use your pencil or pen to mark any sentences that tell information that you think is important.

4. Look at the pictures. Think about what they tell you in addition to the words in the story.
5. At the end of the story, write in one sentence what you would tell a friend if he or she asked you what the story was about.
6. Answer the 10 questions about the information and English in the story.
7. Circle the letter next to the correct answer.
8. You may go back to the story and your writing before the story to answer the questions.

Members of the Everett Alvarez High School drama program are making a final run through of "Whispers From the Fields." Drama teacher Michael Roddy's research on Salinas Valley field workers, their stories, and history inspired the play. The story is about Mexican immigrants who struggle to make a living in the California fields before the United Farm Workers era.

1) What is the title of the play?
   a) They carry a lot of wisdom
   b) La llorona
   c) Whispers from the field
   d) Farm worker history

Listening/Writing

Directions: You are about to hear the story of a Salinas Valley farm worker. You will hear the story two times. The first time, just listen carefully. Do not take notes. The

second time, you may take notes. After the story has been played twice, write down as much of the story as you can remember, using full sentences.

My name is Maricela. My family came to Salinas from San Andreas, Oaxaca because there was no work. Now we're living a life dominated by work, spending 10 hours a day in the fields, six days a week, which leaves only 3 or 4 hours of sleep before rising after midnight. We have lived in Salinas since 1993. We have four kids and one that is grown and married. She lives with her husband in Salinas. Our youngest was born here…

Example: Maricela is from …(San Andreas, Oaxaca) . Now, She lives in … (Salinas, California).

Specification Supplement

Objectively Scored Section

All items will be scored according to the scoring key, which is located in Appendix D. There is only one correct answer per item. Correct answers are worth one point; incorrect answers are worth zero points. Ten points are possible in this section. The scores from this section will be multiplied by a factor of 4.5 to weigh it equal to the subjectively scored section when the final grade is calculated.

Subjectively Scored Section

All items in this section will be holistically scored according to a 4-point rubric. Possible scores for each of the 15 equally weighted, main points can be 0, 1, 2, or 3; the maximum score is 45. The scoring rubric is as follows

0 – the response does not contain any element of the event

1 – the response contains an element of the event but is either incomplete or inaccurate with the rest of the event

2 – the response contains all elements of the event but is written in an incomplete sentence

3 – the response contains all elements of the event and is written in a complete sentence

Statistical Analysis

On March 14, the designer piloted the EAHSESL with 29 intermediate students in Michael Hedgpeth's ESL class. No extraordinary events occurred during the test. Students took the estimated time allotted, 40 minutes for the reading comprehension (Section One) and 20 minutes for the listening comprehension (Section Two), to complete the test. All of the audiovisual equipment (TV and DVD player) worked well and were visible to every student in the classroom.

The designer entered the data from the multiple-choice items of Section One by himself. Two students did not answer one item each; otherwise, every student attempted every item. With the help of a colleague from the first version of this test, Lindsey Babcock, the designer marked the Section Two responses, which required subjective scoring. The norming procedure consisted of the designer explaining the scoring rubric to Ms. Babcock, then the raters scored two tests that were pre-piloted by two English-only speaking friends of the test designer. Next, the raters negotiated for the viability of certain responses that weren't considered in the original rubric. The

rubric was revised to include these possibilities. Finally, the raters graded each test with a minimal amount of conversation until they finished.

Reliability

Estimates of reliability were calculated for equivalence in judgment of raters in the subjectively scored listening/writing section and for consistency within the objectively scored reading comprehension section. Both measures found that the reliability of the test scores is not sufficient for full-scale testing at this time.

Inter-rater Reliability

The listening/writing section was subjectively scored by two raters, the test designer and his partner from the first version of the test. Spearman's rho was used to measure the inter-rater reliability (IRR) because the number of test-takers was under 30 and the scores in the listening/writing section, as seen in Figure 1, were not normally distributed.

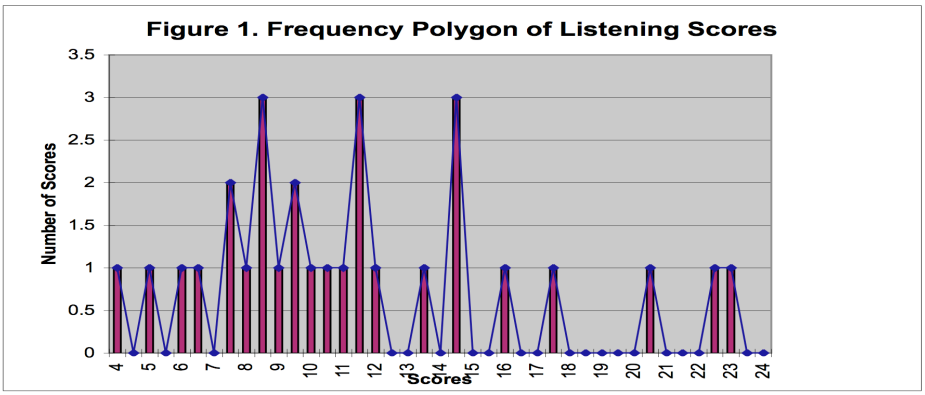
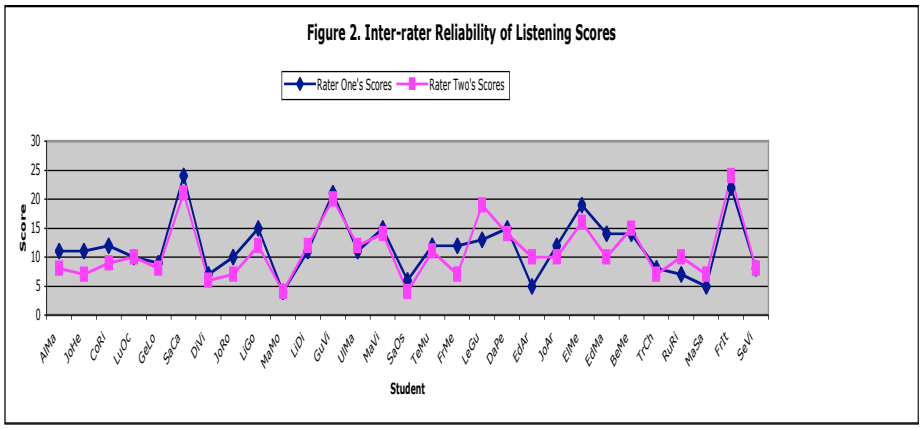


Table 1. Listening Section Measures of Central Tendency (45 pts.)	
Median	10.50

Mode	8.50
Mean	11.47
Range	20.00
Standard Deviation	4.87

Using the software program SPSS, the designer entered the data into the computer. The two sets of interval data were converted into ordinal data and ties of scores were counted. The correlation coefficient between the raters was 0.813, which is significant at the 0.01 level. Although 0.813 is a strong correlation (Brown, 1988; Bailey, 1998), it is weaker than the Pearson's r correlation of 0.97 for the previous version of the listening/writing section. This correlation reflects that, compared to each other, the raters scored the present version of the listening/writing section less consistently than the previous version. This means that the norming procedure should be revised.



There are three places that inter-rater reliability could be improved. First, the test designer did not provide any benchmark or anchor writing samples for the raters to

familiarize themselves with the scoring rubric. In addition, the raters only had one round of two writing samples to align their scores. Third, monolingual English speakers pre-piloted the test. To remedy these problems, the test designer could present the raters with examples of sentences that would earn 0, 1, 2, and 3 points on the scoring rubric. Then the raters would have a more clear understanding of what to look for when they score the writing samples. For the second problem, the norming session should include two or three rounds of grading five writing samples, it would produce more data to reveal where the raters need to align their scoring judgments. Finally, individuals whose second language is English should pre-pilot the test. All of these changes could easily be implemented in the next phase of development for the EAHSESL test.

Internal Consistency

The multiple-choice items did not measure one construct, therefore there should be little consistency between the 10 multiple-choice items. Using Spearman's rho because the scores for the reading comprehension section, as seen in Figure 3, were not normally distributed, the designer calculated the correlation coefficient between the odd and even items to be 0.276.

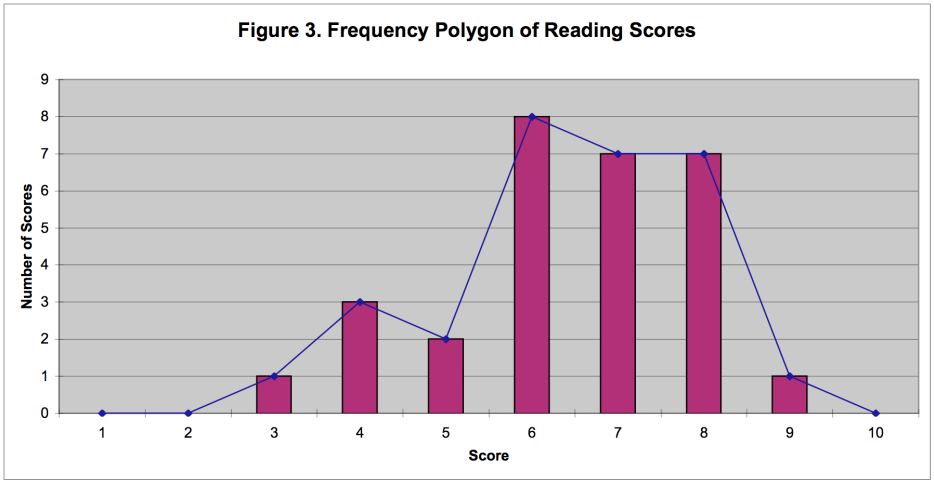


Table 2. Reading Section Measures of Central Tendency (10 pts.)	
Median	7.00
Mode	6.00
Mean	6.45
Range	7.00
Standard Deviation	1.48

This is a weak correlation, meaning that the items inconsistently measured the same construct. In other words, the items measured different constructs. This confirms what the test specifications call for in the reading comprehension section.

Validity

Hatch and Lazaraton (1991) mention five threats to validity that can often be avoided with careful planning by the test designers. These threats are

1. Invalid application of tests
2. Inappropriate content
3. Lack of cooperation from the examinees
4. Inappropriate norming population
5. Invalid constructs (pp. 541-542)

It is important to note that a test is not valid just because it avoids these five threats to validity. A test can begin to be measured as valid once steps have been taken to diminish these threats. The designer worked to avoid the above threats to validity by

1. Designing the test specifically for the type of student who piloted it
2. Choosing content that was about the local area and interesting to the students who took the test
3. Coordinating with the test-takers' teacher to ensure that the piloting of the test did not conflict with other school activities which would reduce their desire to cooperate with the test designer
4. Observing the discourse patterns of classroom language used by the norming population and using that data to construct appropriate test items.

Having established that the circumstances and materials of the test design do not reveal any gross violations of its validity, the designer will begin to discuss the statistical and qualitative data that measure some types of internal and external validity of the EAHSESL.

Content Validity

Although the two sections of the test measure a variety of language abilities, they still fall short of representing the entire range of English language abilities that the test

taker will need to perform academically in an ESL class at EAHS. If there were more people involved in this version of the test design, it may have been possible to create more items to measure more language abilities. The content validity of the test would be strengthened by the increased number of abilities measured. However, because the test designer had to work alone, under a short time frame, it wasn't possible to achieve a satisfactory degree of content validity in this iteration of the placement test.

Face Validity

Each test section uses a direct method of assessment to measure the ability or abilities being tested. The directions for each section did not request any procedure that was out of the ordinary for a reading, listening, or writing exercise. Therefore, *prima facie*, the EAHSESL was accepted by the students as a test that measured reading, listening, and writing. The only new method of assessment was the MCAG. If the response material on this section were graded, it may threaten the face validity of the test because the test takers would be unfamiliar with the stimulus material and therefore confused about what that section is measuring. Until test takers and designers become more familiar with the concept and appearance of dynamic assessment methods, its designers will have a hard time establishing the face validity of tests that include it.

Concurrent Validity

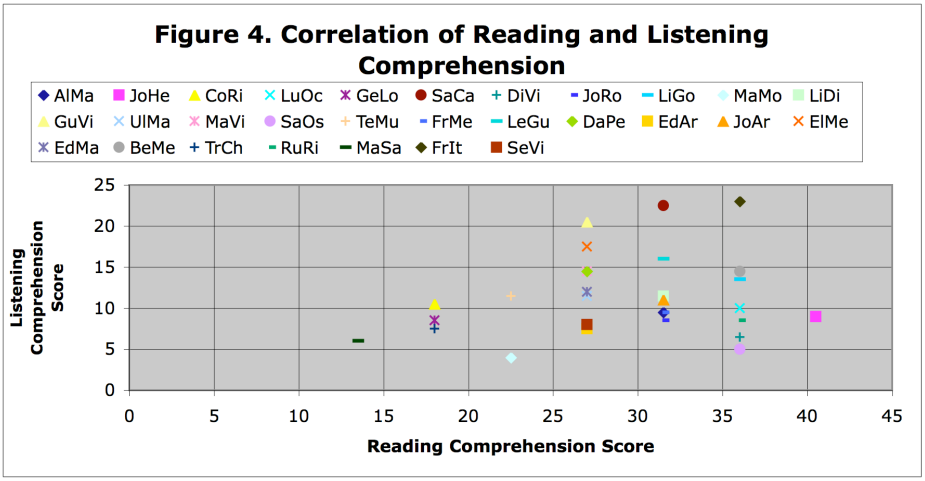
A good placement test should put a student in the appropriate level of classroom instruction that matches his or her current level of language proficiency. It could be argued that students who are placed in the appropriate classroom, will perform well academically. A student may suffer academically if he or she is placed in a class level too

low (becoming bored) or is placed in a class level too high (experiencing anxiety). Academic performance can be measured by the student's grade in the class or by an evaluation from the classroom teacher. Therefore, if the result of a placement test can correlate with a student's academic performance or teacher evaluation shortly after arriving in class but before the student has an opportunity to improve, it can be said that the test has sufficient concurrent validity. Alderson, Clapham, and Wall (1995) contend that this correlation could also be considered as evidence of predictive validity.

In the case of the EAHSESL, it does not currently have concurrent validity because no correlation procedure has been implemented at this time. It could be possible to correlate results of the EAHSESL with test takers' grades or evaluation by their ESL teacher if the principal, teacher, and parents give permission to release the information. This could be something to pursue in the next phase of development for the EAHSESL test.

Construct Validity

The small scale of this project and the short time frame that the sole test designer had to complete the project hampered his ability to generate data to support the test's construct validity. It was not feasible for him to assemble a panel of experts or design an experimental treatment to evaluate the degree to which the EAHSESL could operationalize the two theories, schema theory and dynamic testing, underlying the test design. The next phase of the project could include a quasi-experimental design to measure the effect that the MCAG has on the student's proficiency test performance.



However, the test designer was able to calculate a correlation between the scores of the two sections to measure the amount of overlapping variance. Spearman's rho was selected as the statistical test to measure correlation between the reading comprehension and listening/writing sections because the number of participants was below 30. In order to meet the assumptions for Pearson's r, the population should be at least 30 and the data should be normally distributed (Hatch & Lazaraton, 1991). The correlation coefficient between the two sections was 0.146. This is a weak correlation. It means that the two sections measure different things and that language skills required for one section don't enable the test taker to perform in the other section.

Consequential Validity

One societal impact that could result from students' interaction with the test is an increase in school and community pride. This is because the subject matter, EAHS students producing a play about the local history of migrant farm workers, is something to which most test takers will be able to relate. By using a DVD player to deliver the

content for the listening/writing section, the EAHSESL test utilizes the latest technology and conveys the message that classroom instruction will also use interactive media. The presence of the MCAG at the beginning of the test highlights the ESL teachers' desire for their students to use metacognitive strategies in the learning process. Overall, the consequential validity of the test is a strong feature that should be accentuated in future editions.

In conclusion, the scores of this test are yet valid. The test designer avoided some key threats to validity, which allowed further investigations of validity to proceed. At this time, these investigations have not yielded results that confirm the test scores are valid and reliable. There are still more constructs to measure, more correlations to calculate, and more testing methods with which to familiarize the participants, before the validity of the EAHSESL can be established. With a longer timeline, a larger staff, and a budget, it may be feasible to climb the necessary steps to sufficient validity for the EAHSESL.

Item Analysis

Most of the items on the reading comprehension section had an adequate degree of discriminibility. Two of the items did not discriminate and the designer has learned something from them.

Table 3. Item Discrimination for Reading Section											
Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Top Eight Scorers											
JoHe	1	1	1	1	1	1	0	1	1	1	9
LuOc	1	1	1	1	1	1	1	1	0	0	8
DiVi	1	1	1	1	0	1	1	1	0	1	8
LiGo	1	1	1	0	1	1	1	1	0	1	8
SaOs	1	1	1	1	1	0	1	1	0	1	8
BeMe	1	1	1	1	1	1	1	1	0	0	8
RuRi	1	1	1	1	1	1	0	1	0	1	8
FrIt	1	1	1	0	1	1	1	1	0	1	8
Bottom Eight Scorers											
MaSa	0	1	0	0	1	0	0	0	0	1	3
TrCh	0	1	1	0	1	0	0	1	0	0	4
GeLo	0	0	1	0	1	1	0	1	0	0	4
CoRi	0	1	1	0	0	1	1	0	0	0	4
TeMu	0	0	1	1	1	1	0	1	0	0	5
MaMo	0	1	1	0	0	1	1	1	0	0	5
SeVi	1	1	1	1	0	0	0	1	0	1	6
DaPe	0	1	1	1	0	1	1	0	0	1	6
ID (N=8)											
27.5%	**0.88**	**0.25**	**0.13**	**0.38**	**0.38**	**0.25**	**0.38**	**0.38**	**0.13**	**0.38**	

One of the items, number 3, asked the participant to translate a Spanish word into English. All but one of the participants got this item correct. The test designer constructed the item from the perspective of an English speaker having to translate a Spanish word back into English. As a result, he did not choose options with a high degree of semantic nuance between them. What the designer should have done was put

himself in the position of a Spanish speaker translating a word into English. Possible options in the next draft should include: paternal grandparents, maternal grandparents, and elders. These would be more suitable options that are plausible, yet incorrect (Celce-Murcia, Kooshian, and Gosak, 1974). The second item that did not discriminate well, item 9, was too difficult because it was the only cloze-passage type item on the test.

Table 4. Distracter Analysis for Reading Section

Option/Item	1	2	3	4	5	6	7	8	9	10	Total
Total "a"	0	26*	1	6	0	24*	2	0	14	9	82
Total "b"	0	0	28*	3	4	3	4	25*	1	17*	85
Total "c"	17*	0	0	13*	8	0	7	2	4*	2	53
Total "d"	12	3	0	5	17*	2	16*	2	10	1	68

Note: * identifies the correct option

It is evident in Table 4 that option "a" of item 9, "searches," could have worked grammatically but was not the best option. There was not enough context around the sentence to rule out that the woman may have been searching for something.

The facility, or difficulty level, of each item in the reading comprehension section, can be seen in Table 5. Item facility (IF) is measured by dividing the number of test-takers who got the

Table 5. Item Facility for Reading Section

Item No.	1	2	3	4	5	6	7	8	9	10	TOTAL
Total Correct (N=29)	17	26	28	13	17	24	16	25	4	17	187
Item Facility	0.59	0.90	0.97	0.45	0.59	0.83	0.55	0.86	0.14	0.59	

item correct by the total number of test-takers. Bailey (1998) and Oller (1979) agree that an IF of 0.15 to 0.85 is acceptable. According to this scale, it is evident that items 2, 3, and 8 are too easy; item 9 is too difficult. Using Table 4, "Distracter analysis for reading section" it is clear what options were good and poor distracters. Items 4, 7, 9, and 10 all had at least one student choose each option. Even though item 9 will be replaced because it is too difficult, at least the test designer did a good job of writing distracters that were plausible, yet incorrect.

Items 1 and 3 should have two of their options replaced because students selected only two of the four options. Item 1 was the strongest discriminator between the high and low scoring students, visible in Table 3. In the next phase of revisions on this test, the test designer will maintain effective discrimination between high and low scorers while improving the distraction between options. However, if an item discriminates well, the diversity of distracters may be inconsequential.

In summary, the item analyses reveal some strong points in the designer's ability to construct multiple choice items that measure a certain construct, discriminate between high and low scorers, and provide credible distractors. At the same time, the designer's inexperience with item construction allowed some options to be too easy or hard for the age group being tested. More experience designing items will give the test designer better judgment in creating plausible multiple-choice items.

Suggestions for Further Research

Six things can be done to improve the EAHSESL before it is piloted again. First, if the EAHSESL will be used to complement the oral delivery of the CELDT in placing

students, it will be necessary to conduct a correlation design study between the two tests. This statistical procedure will reveal to what degree the two tests measure the same construct. It would be desirable that the two tests do not measure the same things and therefore test all four language skills of each student.

Second, a quasi-experimental study should be conducted to determine the effectiveness of the MCAG on student performance on the EAHSESL as measured by their reading comprehension section score. The results from such a study could inform the test designer of any revisions that the MCAG would need with this population of students to perform its intended purpose of raising metacognitive awareness to increase performance on the reading comprehension section.

Third, the timing or population of the pilot session should be reconsidered. The test version in this paper was piloted in February, after the participants had had at least five months of English instruction. If the test is going to place students who could potentially have moved to an English-speaking community for the first time, the population of the group piloting the test should reflect that profile as well. This could be addressed in two ways: Either assemble a group of students who have just moved to the area or conduct the pilot session with an intact group of students at the beginning of the school year before they have shown progress from any kind of instruction.

Fourth, the prompt attributes of the reading comprehension section should be adjusted. Items in the multiple-choice section should be revised to fit within the acceptable range of discriminibility. The inclusion of a person whose first language is Spanish in the creation of test items would benefit them because s/he could relate to the

same mental process of translating vocabulary from first to second language that the student will experience when answering the test items.

Fifth, it has already been discussed how the norming procedure should be revised to produce a higher correlation coefficient between raters. If a third rater is added to an improved norming process for scoring the subjective listening/writing section, it could also benefit the inter-rater reliability by making each student's mean score less volatile when the rater's scores are averaged.

Finally, the test designer could change the EAHSESL from a norm-referenced test to a criterion-referenced test. If students must exit high school with a certain set of communicative competencies, then their placement into high school should be measured according to the same criteria. This change could be the most fundamental to the EAHSESL because it would require the test designer(s) to use a new set of statistical procedures to measure reliability and validity and to draft standards against which to measure student performance. However, the change would reduce anxiety and competition amongst the students because the test measures them according to ability instead of against each other.

Conclusion

The current version of the EAHSESL test has shown some improvements over its predecessor, but its scores are neither reliable nor valid. Improvements include the integration of schema and dynamic assessment theory into the reading comprehension section, which brings the latest assessment methods to bear on the placement needs of Everett Alvarez High School's ESL department. The qualitative and quantitative analyses

reveal that the EAHSESL is neither reliable nor valid in its present iteration. However, the test designer's ability to conduct statistical procedures and interpret their outcomes has improved. His raised awareness of the test's shortcomings will motivate the revisions in the next phase of design. Overall, the designer has learned that developing a language placement test requires multiple cycles of design, piloting, and analysis to create an effective assessment tool that minimizes any mismatch between a class level and a student's ability.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). Language test construction and evaluation. Cambridge: Cambridge University Press.

Bachman, L. (1990). *Fundamental considerations in language testing*. New York: Oxford University Press.

Bailey, K. M. (1998). Learning about language assessment: Dilemmas, decisions, and directions. Boston: Heinle & Heinle.

California Department of Education. (2002). *History-social science content standards for California public schools: Kindergarten through grade twelve*. Sacramento, California: Author.

Carrell, P. L., & Eisterhold, J. C. (1983). Schema theory and ESL reading pedagogy. *TESOL Quarterly*, *17*, 553-573.

Celce-Murcia, M., Kooshian, G. B., & Gosak, A. J. (1974). Goal: Good multiple-choice language test items. *English Language Teaching*, *28*, 257-262.

Guterman, E. (2002). Toward dynamic assessment of reading: Applying metacognitive awareness guidance to reading assessment tasks. *Journal of Research in Reading*, *25*, 283-298.

Hatch, E., & Lazaraton, A. (1991). The research manual: Design and statistics for applied linguistics. Boston: Heinle & Heinle.

Hughes, A. (1989). Testing for language teachers. Cambridge: Cambridge University Press.

Larsen-Freeman, D. (2003). Teaching language: From grammar to grammaring. Boston: Thompson-Heinle.

Oller, J. W., Jr. (1979). Language tests at school. London: Longman Group Ltd.

Poehner, M. E. & Lantolf, J. P. (2003). "Dynamic assessment of L2 development: Bringing the past into the future." *CALPER Working Papers Series, No. 1*. The Pennsylvania State University, Center for Advanced Language Proficiency, Education and Research.

Swain, M. (1984). Large-scale communicative language testing: A case study. In S. J. Savignon and M. Burns (Eds.), *Initiatives in communicative language teaching*. Reading, UK: Addison-Wesley.