

Copyright

by

Brent Alexander Schackmann

2015

**The Report Committee for Brent Alexander Schackmann  
Certifies that this is the approved version of the following report**

**PHP/HTML Design and Build of a Computer  
Adaptive Test to Assess English Fluency Among  
Native Spanish Speakers**

**APPROVED BY  
SUPERVISING COMMITTEE:**

**Supervisor:**

---

Paul von Hippel

---

Rajagopal Raghunathan

**PHP/HTML Design and Build of a Computer  
Adaptive Test to Assess English Fluency Among  
Native Spanish Speakers**

**by**

**Brent Alexander Schackmann, B.A.**

**Report**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Master of Public Affairs and Master of Business Administration**

**The University of Texas at Austin**

**May 2015**

## **Abstract**

# **PHP/HTML Design and Build of a Computer Adaptive Test to Assess English Fluency Among Native Spanish Speakers**

Brent Alexander Schackmann, MPAff, M.B.A.

The University of Texas at Austin, 2015

Supervisor: Paul von Hippel

Abstract: The following is a review of key findings from the implementation of a PHP/HTML web-based application to assess English fluency among native Spanish speakers. The scope of this professional report includes mainly the design, build, and implementation of a web based system accessible through [www.babelous.com](http://www.babelous.com). This written portion is intended to briefly summarize initial results from the implementation of the successfully built application, provide information on how to replicate the application, and detail areas of focus for future development.

## Table of Contents

Chapter 1: Introduction .....	1
Lexical Frequency Profile .....	4
Chapter 2: Methodology for the Test.....	5
The Binary Search Algorithm .....	9
Chapter 3: Implementation .....	14
Chapter 4: Data .....	16
Chapter 5: Results .....	18
Chapter 4: Conclusion.....	24
Appendix A.....	26
Appendix B .....	27
Bibliography .....	30

## Chapter 1: Introduction

K-12 teachers of English Language Learners (ELLs) face a challenge in designing academically appropriate content for students. In this environment teachers often have imprecise and outdated information about individual student fluency levels, making the task of teaching content standards difficult.

Many states use English Language Development (ELD) standards to guide the assessment of ELLs.<sup>12</sup> Schools often administer ELD tests at the beginning of each academic year, where tests are designed to cover reading, writing, and listening skills, and generate a score that places students into one of a few categories. California, for example, scores students as ELD1 through ELD4—categories defined by the State of California Department of Education. In California the categorization of ELD4 demonstrates the highest level of English fluency.<sup>3</sup> Similar to California, the consortium of 36 states using ELD standards defined by WIDA (based at the University of Wisconsin) assign students into proficiency category A, B, or C, breaking out some additional information related to reading, writing, and listening skill.<sup>4</sup> Teachers under either the California or WIDA ELD system receive similar information along with their classroom roster at the beginning of the year, and are expected to generate standards-

---

<sup>1</sup> "Consortium Members." WIDA: Member States. Accessed April 26, 2015. <https://www.wida.us/membership/states/>.

<sup>2</sup> "English Language Development Standards." Resources (CA Dept of Education). March 15, 2015. Accessed April 26, 2015. <http://www.cde.ca.gov/sp/el/er/eldstandards.asp>.

<sup>3</sup> Torlakson, Tom. "California Department of Education." *Overview of the California English Language Development Standards and Proficiency Level Descriptors*, 2012. Accessed March 11, 2015. <http://www.cde.ca.gov/sp/el/er/documents/sbeoverviewpld.pdf>.

<sup>4</sup> "WIDA's 2012 Amplification of the English Language Development Standards, Kindergarten–Grade 12." 2013. Accessed March 10, 2015. <https://www.wida.us/standards/eld.aspx>.

based content that can be understood by all students.<sup>5</sup> An example of the type of information provided to teachers using WIDA ELD standards is included as Appendix A.

While ELD testing provides useful feedback, the current process can be aided—and teaching quality improved—by adding an assessment to address three specific weaknesses in the current process. First, the results of ELD evaluations are coarse. The range of language fluency within each ELD category is broad, meaning a teacher has imprecise information about the true ability of a student given his or her classification. Designing an assessment that gives more precise fluency estimates to teachers could enhance curriculum design and ensure the broadest range of understanding among the classroom population.

The second major weakness in the current process is that ELD information is updated only once per year. Given the scope ELD testing (reading, writing, listening), and that tests are still primarily paper-based, it would be impractical to administer ELD tests more frequently. The annual overhead for administering these exams would double with biannual testing. While it is impractical to test students more frequently using traditional ELD assessments, it is true that a student's English fluency may improve dramatically over a single year. Teachers, therefore, may make curriculum decisions in May that are designed for fluency levels from September, which may fail to appropriately challenge students as they develop language skill. Ideally, to move students closer to English fluency, teachers would design academic content to push each student. A quick fluency assessment that is free and easy to use would deliver updated information to teachers and improve curriculum decisions over the course of the entire year.

---

<sup>5</sup> English Language Development Standards, California.

The third weakness of the ELD system in many states is that the scale is difficult to interpret. It is not impossible to figure out the exact meaning of each ELD category (ELD1-ELD4, or Category A, B, C), but it requires documentation, charts, and dozens of qualifications. The states using WIDA's ELD standards use a 138-page document that breaks down how to interpret and understand the information.<sup>6</sup> California's decoding document is 28-pages.<sup>7</sup> It is entirely possible then, that a teacher may lose all or part of the meaning for each classification because of the complexity. Using a more intuitive scale would benefit teachers, especially those dealing with numerous ELLs at many different ELD levels.

These three areas of concern highlight the need for improvement in the current system. ELD information is better than no information at all, but could be significantly aided if a new system—one that was fast, freely available and easy to interpret and access—was designed to provide precise estimates of fluency. An efficient system that delivers accurate fluency results would allow students to be tested more frequently, meaning teachers could better monitor progress and ensure students are appropriately challenged.

The aim of this report is to investigate and develop software to assess student fluency. Specifically, the software system, called Babelous, attempts to aid or improve the current ELD classification paradigm through the use of computer adaptive testing technology. Babelous is a system designed to be easy to interpret—generating a fluency score out of 100%; because, for example, a 75% fluency score is immediately understandable without the need for lengthy documentation. As Babelous is software

---

<sup>6</sup> "WIDA's 2012 Amplification of the English Language Development Standards, Kindergarten–Grade 12."

<sup>7</sup> Torlakson, Tom, *Overview of the California English Language Development Standards and Proficiency Level Descriptors*.



based, the test would also be easy to administer at any point in the year, is inherently scalable across schools because it is accessible by anyone with an internet connection, and is fundamentally low-cost as compared to traditional paper-based methods. The key to successfully building Babelous is accurately estimating student fluency level efficiently and reliably. To accomplish the estimation this report draws heavily on the research of Laufer and Nation's *lexical frequency profile*<sup>8</sup>, while the efficiency and reliability component is handled through research into computer adaptive testing best practices.

### **LEXICAL FREQUENCY PROFILE**

80% of written English uses only the 2,000 most frequent English words, and 95% of written English uses only the 5,000 most frequent words.<sup>9</sup> Laufer and Nation's original research suggests that estimating the lexical frequency profile for an individual correlates strongly with that individual's ability to understand both written and spoken English.<sup>10</sup> Therefore, an assessment designed to determine a student's lexical frequency profile can be used to estimate fluency level as well. Developing the lexical frequency assessment around best practices in computer adaptive testing allows for an efficient, interpretable, and scalable test, which can deliver fluency information directly to teachers as frequently as necessary. If successful, this model could greatly aid the ELD process and improve educational outcomes for ELLs in classrooms using the implementation.

---

<sup>8</sup> Laufer, B., and P. Nation. "Vocabulary Size And Use: Lexical Richness In L2 Written Production." *Applied Linguistics*, 1994, 307-22.

<sup>9</sup> Laufer, B., and P. Nation. "A Vocabulary-size Test of Controlled Productive Ability." *Language Testing*, 1999, 36-55.

<sup>10</sup> Laufer, 1995.

## Chapter 2: Methodology for the Test

In order to build a successful computer adaptive test—and in accordance with Laufer and Nation’s research regarding how many words are required to reach certain levels of fluency—it is necessary to start with a database of the 5,000 most frequently occurring English words. To determine which English words occur most frequently, scholars from Brigham Young University (BYU) built the Corpus of Contemporary American English and developed a machine-learning algorithm to count word instances.<sup>11</sup> The Corpus consists of 450 million English words from five document types, chosen to represent words across different contexts. The five document types include: spoken language (transcribed), fiction literature, magazines, newspapers, and academic content.<sup>12</sup> After processing the algorithm, a list of the 5,000 most frequent English words was generated by Mark Davies, professor of Linguistics at BYU.<sup>13</sup> Below is a sample of the information contained in the wordlist:

rank	Lemma/word	PoS	freq	dispersion
7	to	t	6332195	0.98
14	you	p	3085642	0.92
21	they	p	1865844	0.96
28	not	x	1638883	0.98
35	go	v	1151045	0.93
42	her	a	969591	0.91
49	as	i	829018	0.95
56	think	v	772787	0.91

Table 1: Sample English Frequency Wordlist.

<sup>11</sup> "Corpus of Contemporary American English (COCA)." Corpus of Contemporary American English (COCA). January 1, 2012. Accessed April 26, 2015. <http://corpus.byu.edu/coca/>.

<sup>12</sup> Corpus of Contemporary American English (COCA).

<sup>13</sup> Davies, Mark. "Word Frequency Data." Word Frequency: Based on 450 Million Word COCA Corpus. Accessed April 26, 2015. <http://www.wordfrequency.info/intro.asp>.

The rank represents the position of each word, sorted by its frequency value. Frequency is the raw number of times the words appears across the 450 million-word Corpus. Dispersion is a scaled value indicating how evenly the word appears across the Corpus and across the five document types—a value of 1 indicates the word appears in all five-document types and multiple times across the entire Corpus. Finally, PoS stands for part of speech, which can be decoded according to the CLAWS7 tagset.<sup>14</sup>

With a reliable list of the 5,000 most commonly occurring English words, which allows for estimation of fluency according to the student's lexical frequency profile, the computer adaptive test can be built to efficiently handle fluency estimation. The initial computer adaptive test is designed for native Spanish speaking ELLs. For the scope of this project only one ELL population (Spanish speakers) could be tested, as each additional language requires translation of the common English words database.

In this case, the English words database was translated into Spanish. To verify accuracy, 50 Spanish-translated words were randomly sampled and presented to volunteers fluent in English and Spanish. The randomly sampled English words and Spanish translations were 100% accurate, lending confidence that the translation procedure produced a relatively accurate list of English words and Spanish equivalents. Below is the final structure of the wordlist database, which is a merged list of English word frequencies and Spanish translations:

---

<sup>14</sup> Davies, Mark. "Word Frequency Data."

word_id	rank	english_word	spanish_word	pos	frequency	log_freq
849	850	discuss	discutir	verb	46852	10.75474898
850	851	indeed	en efecto	adverb	46184	10.7403887
851	852	force	forzar	verb	44931	10.71288326
852	853	truth	verdad	noun	45155	10.71785629
853	854	song	canción	noun	45352	10.72220956
854	855	example	ejemplo	noun	47134	10.76074989

Table 2: Final wordlist structure. Future adaptations for additional languages can be made easily by translating the English word set into any other language in the world.

The final database contains rank, PoS, frequency, and English\_word from the original wordlist (explained above) along with columns for word\_id, Spanish\_word and log\_freq. Word\_id represents a unique identifying value for each row. This value differs from rank because in some instances rank is a repeating value (for words that have the same frequency score). Log\_freq is the natural log of the frequency value (frequency, defined above, is the raw number of times a word appears across the Corpus).

This database of the 5,000 most frequent English words and Spanish equivalents, along with the frequency value for each word, provides the information structure to estimate a Spanish-speaking ELLs level of English fluency. With the data in place, a PHP/HTML computer adaptive test can fetch words to test user understanding, allowing for each individual's lexical frequency profile to be determined. From this their true fluency rate can be estimated. Laufer and Nation's original research<sup>15</sup> along with additional research from Laufer and Nation<sup>16</sup> and Lembier<sup>17</sup> point to the usefulness of a multiple choice test for this process. A multiple-choice test presents an English word and

---

<sup>15</sup> Laufer, 1995.

<sup>16</sup> Laufer, 1999.

<sup>17</sup> Lemhöfer, Kristin, and Mirjam Broersma. "Introducing LexTALE: A Quick and Valid Lexical Test for Advanced Learners of English." *Behavior Research Methods*, 2012, 325-43.

part of speech, expecting students will match these to the Spanish word with the closest meaning if they know the English word, while they will select an incorrect Spanish translation if they do not know the English word. Below is an example of the general testing structure:

Please identify which Spanish word most closely matches the meaning of the **noun: viewer**

- desde
- eso
- estos
- espectador
- este
- a través de
- vida
- nuestro
- entonces

[Next Question](#)

Illustration 1: Example of the testing structure.

Notice two things: first the number of possible selections. The number of possible choices helps reduce random correct guesses to a probability of 1 in 9. Second, the answer choices are intentionally selected from the database to be words that satisfy one of two criteria: each answer choice must contain similar letters to the correct Spanish word, or to the English word in question. The logic to handle this selection is based in SQL string matching. The database is queried for words that contain the first two letters of the correct Spanish word, or the first two letters of the English word. In the above example

the correct translation for the English word *viewer* is the Spanish word *espectador*. Each of the answer choices contains a letter combination of ‘es’ or ‘vi’. This helps eliminate obvious throwaway choices for random guessers, and also presents false cognates to users (e.g. the Spanish word *la arena* means *sand* and not *arena*). Both factors intend to lessen the impact of random guessing and provide more accurate results with fewer questions.

The assumption in this model, and in much of the research, is that a student who knows a word ranked as the 2500th most frequent is statistically likely to know other words with a similar frequency rank. Therefore, presenting a student with incrementally harder words is an inefficient way to estimate the lexical frequency profile. In accordance with generally accepted standards in computer adaptive testing, the goal is to present each student with a word they are about 50 percent likely to know. Prior to a sufficient sample of students taking a basic version of the test it is impossible to estimate which words follow this 50/50 rule. In the initial version of the test, then, the decision on which word to present next is made using a variant binary search algorithm.

#### **THE BINARY SEARCH ALGORITHM**

A critical component of any computer adaptive testing model is programming the decision making rule. Simply, how will the computer decide which word to give the user next? Given the general 50/50 rule, it is critical to create a decision algorithm, which presents a user with harder words after a correct answer and easier words after an incorrect answer. The model is optimal if it presents a word that, given the students previous answers, is estimated to have a 50 percent probability of eliciting a correct answer.

Decision rules in state-of-the-art adaptive testing systems are based on item response theory (IRT).<sup>18</sup> While incorporation of an IRT model is a desired feature for a future version of Babelous, the initial prototype uses a much simpler binary search algorithm based on the log frequency. More specifically, Babelous presents the user with a word whose log frequency is midway between the log frequency of the last word they translated correctly (assumed to be the first word in the database until a word is answered correctly), and the last word they did not know (assumed to be the last word in the database until a word is answered incorrectly).

The motivation for the log frequency rule is the Hick-Hyman<sup>19</sup> law, which claims that, in a stimulus-response task, users' response time is linearly related not to the frequency but to the log frequency with which they have been exposed to the stimulus.<sup>20</sup> We assume that the log frequency is also related to the probability of a correct response. Note that the Hick-Hyman law was originally developed using data from a small number of experimental subjects, and may be only approximately correct. In addition, the psychological law governing response time may be different from that governing the probability of a correct answer. Data collected from Babelous may be used to test the Hick-Hyman law and develop alternatives.

The first word presented to the user should also follow the 50/50 rule. The initial decision rule for first word choice is based on two factors. First, the word should be approximately in the middle of the list (somewhat close to word 2500) to best facilitate the binary splitting decision for the remaining words in the list. And second, the word

---

<sup>18</sup> Muñiz, José, Wim J. Van Der Linden, and Ronald K. Hambleton. "Handbook of Modern Item Response Theory." *European Journal of Psychological Assessment*, 1997.

<sup>19</sup> Hyman, Ray. "Stimulus Information As A Determinant Of Reaction Time." *Journal of Experimental Psychology*, 1953, 188-96.

<sup>20</sup> Seow, Steven. "Information Theoretic Models Of HCI: A Comparison Of The Hick-Hyman Law And Fitts' Law." *Human-Computer Interaction*, 2005, 315-52.

should approximately follow the Hick-Hyman principle guiding the other word decisions in the test—namely the first word should be close to the average log-frequency value for the entire wordlist. Given the two decision points, a set of words 100 words best satisfying both was identified. From the list of 100, the final start words were narrowed to 34 words, where obvious cognates were eliminated.

The following example walks through the Babelous process; notice the range of remaining words (difference between the floor and ceiling words) shrinks considerably after each question due to the binary decision rule:

- 1. First word:** selected randomly from the list of 34 words. The English word is *retirement*, with the Spanish equivalent *jubilación*. Word rank for *retirement* is 2464 and log-frequency is 9.544. **The user gets this word correct.**
- 2. Second word:** Because word one was correct, Babelous selects the next word by taking the new floor word rank value of 2464 (equal to the word rank of the previous correct answer) and the ceiling word rank value (5000 because no word has been answered incorrectly yet) and calculates the log frequency average for all words in the database between words 2464 and 5000. The next word that is presented to the user is the word whose log frequency value is closest to the log-frequency average for all words between the floor word (2464) and the ceiling word (5000). Babelous calculates the log frequency value as 8.9847 and selects the English word *trait* with the Spanish equivalent *rasgo*. Word rank for *trait* is 3778 and the log-frequency is 8.98469. **The user gets this word incorrect.**
- 3. Third word:** A new ceiling word rank is established at the previous incorrect word and Babelous now calculates the log frequency average for all words between the established word floor (2464) and the new word ceiling (3778). Babelous calculates the log frequency value as 9.25302 and selects the English



word *pipe* with the Spanish equivalent *pipa*. Word rank for *pipe* is 3005 and log-frequency is 9.2526.

- 4. Fourth through final word:** The above process repeats with a new floor established with each correct answer, and a new ceiling established for each incorrect answer.

After 6-8 questions, the difference between the floor word and ceiling word is small, only a few words, allowing the computer to estimate the lexical frequency profile within a small range on which to generate the estimate for English fluency.

The binary search algorithm is extremely efficient in estimating a student's lexical frequency profile. However, the current decision rule is vulnerable to uncharacteristic answers, especially early in the test set, which can lead to poor results. For example, consider a student who knows little English but recently had a grandparent retire. This student may have tacitly learned the English word *retirement* as one of the few English words in his/her English vocabulary. When Babelous presents this student with *retirement* (1/34 probability this happens) as a first word, and the student gets the word correct, Babelous assumes the student also knows words 1 through 2464—the word rank for *retirement*—in the database. Even if the student gets every remaining word on the test wrong, Babelous will assess their English fluency at approximately 50%, though their true fluency rate may be considerably lower.

Under the current binary decision rule, it is highly recommended for students to take the test multiple times. Consider that if the average test is 7 questions in length, two passes through the exam only requires 14 questions, and should greatly improve the likelihood that fluency estimates approach a true value. Averaging two passes helps down-weight the significance of uncharacteristic right or wrong answers under the binary decision model. Ideally the test would be structured such that the second (or third) pass

happened automatically, without revealing a true fluency score until all 14 (or 21) questions had been answered and the fluency estimates for each pass calculated. Babelous could then present the average fluency estimate of all passes as the fluency estimate. More optimally, additional passes would not restart students in the middle of the set—randomly in the list of 34 words—but near where the previous pass ended. Because each pass amounts to a restart, the floor and ceiling would be reset each time, allowing Babelous to evaluate the accuracy of its previous estimate(s). If the previous pass estimates fluency at 65%, presenting the user with a word around the 65% level in the database as the first word in the next pass, without an established floor or ceiling, allows Babelous to determine how close to 65% the user actually is on the second pass.

Babelous' current binary decision rule is adequate for the gathering of initial data, but will need to be replaced by an IRT model in future versions. Although there is room for improvement regarding the binary search rule, it is possible the current version could produce useful feedback to students and teachers. This could be especially true if the teacher were to record two or more rounds of data on a single student. Taking the average of multiple assessments might establish a reasonable estimate to teachers and students about the students' current lexical frequency profile, and therefore the current level of English fluency. A test of the current version is examined in the results section.

## Chapter 3: Implementation

To effectively implement Babelous as a computer adaptive test, the web-based application was constructed using PHP, HTML, and MySQL. The wordlist database (described above), along with the folder structure for babelous.com reside on server space leased through Dreamhost. The database was built using phpMyAdmin, used to import the merged English wordlist and Spanish translations. Additionally, a second data table was designed and setup to capture results from each submitted test. After correctly structuring the database, the front-end PHP/HTML logic was built to connect and render data appropriately.

The PHP/HTML design for Babelous is intentionally modular. What this means is each key component of logic is broken out as a unique function. Thus, the current test can be modified to incorporate a smarter decision rule relatively easily. Over time it will be critical to continue building the database of input data. After enough representative data is collected, a statistical model can be estimated and programmed into the current testing environment. The end result should be a computer adaptive test that address the three concerns noted in regard to the current ELD classifications; namely that the current classifications are too broad, they are difficult to interpret, and the test is only administered once per year.

Below is a representation of how the Babelous system is built and how information moves through the environment:

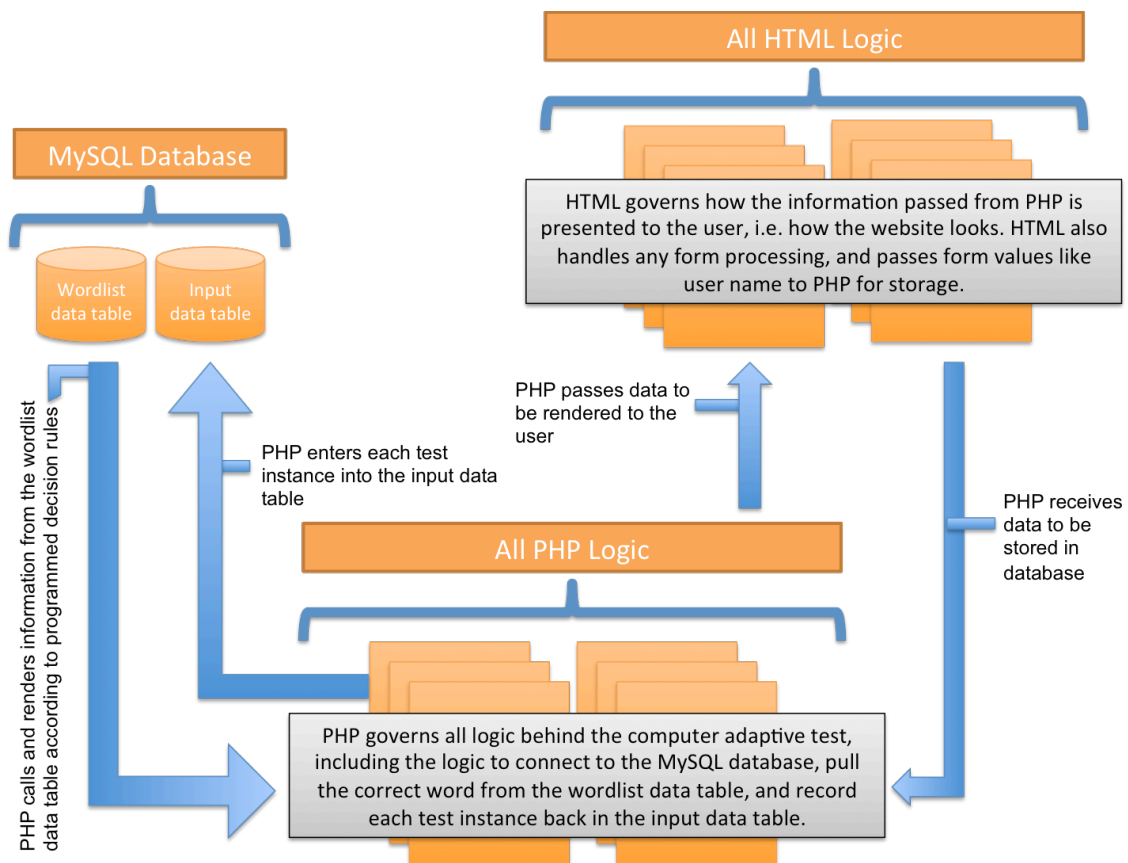


Diagram 1: Example of the testing structure.

The system is publically available via [www.babelous.com](http://www.babelous.com) and the test is free and easy to take. Additionally, the source code with all relevant files has been shared with a supervising professor.

## Chapter 4: Data

In order to review results, and evaluate the current version of the test, it is important to understand the profile of the subjects. Initial results from Babelous are from 20 ELL student volunteers who each took the assessment twice. All of the students are high school aged and in Burbank, California. Given California's annual ELD testing, each result from Babelous is compared against the ELD classification (in California ELD students are classified as ELD1 through EL4) for each student.

The population of the 20 participating ELL students are ELD classified as follows:

Students 1-2	ELD1
Students 3-6	ELD2
Students 7-12 <sup>21</sup>	ELD3
Students 13-20	ELD4

Table 3: ELD categories by student participants. ELD4 indicates the highest level of English fluency.

First, note that these students do not demonstrate a particularly diverse or representative sample of ELL students. They are all high school aged and there are far more ELD4 students than any other category. The average number of years this ELL population has been living in the United States is 6.2. Younger students, or students in lower grade levels, may better fill the ELD1 and ELD2 categories. The dataset does provide initially useful feedback for analysis, but would need to be validated and further tested with a broader, and more representative sample of students in all ELD levels.

---

<sup>21</sup> There is no student 11 in the sample, as a response with this ID was not submitted.

Additionally, broader and more representative sampling would lend increasing strength to the statistically based IRT decision rule for future test iterations.

Because the test subjects are from California, a breakdown of California's ELD classifications is included as Appendix B. Reviewing these standards reveals the differences between each classification, and what students generally need to demonstrate to be categorized in each of the four groups. Understanding what is meant by each of the ELD categories can help frame the results section.

## Chapter 5: Results

Initial test results—which estimate English fluency percentage based on the lexical frequency profile research—correlate well with ELD categorization. However, among the ELD2 and ELD3 population there is wide variance in the fluency estimated by Babelous, which is not unexpected given the current binary decision rule. Under the current binary decision rule the model loses sensitivity at the margins. Every time a new floor or ceiling is set the current decision rule is making an approximation for the 50/50 rule—when a new floor or new ceiling is never set (or is not set until several questions into the test) because students keep getting questions right or wrong, the breadth of the approximation for the 50/50 rule shrinks considerably. Therefore it is not unexpected for the model to perform best at identifying ELD1 and ELD4 students, and to demonstrate wider variance among ELD2 and ELD 3 students.

The first set of results suggests promise in current methodology, while clearly highlighting the need for a better decision rule than the binary search algorithm. The goal, of course, is to use this type of data to inform an IRT model that will ultimately replace the binary decision rule. Results are shown below for the first pass from each student, then the second pass from each student, and finally from the average of both passes for each student. Each tick mark represents an individual student score assessed out of 100% by Babelous, and the average score by ELD classification is shown as the trend line.

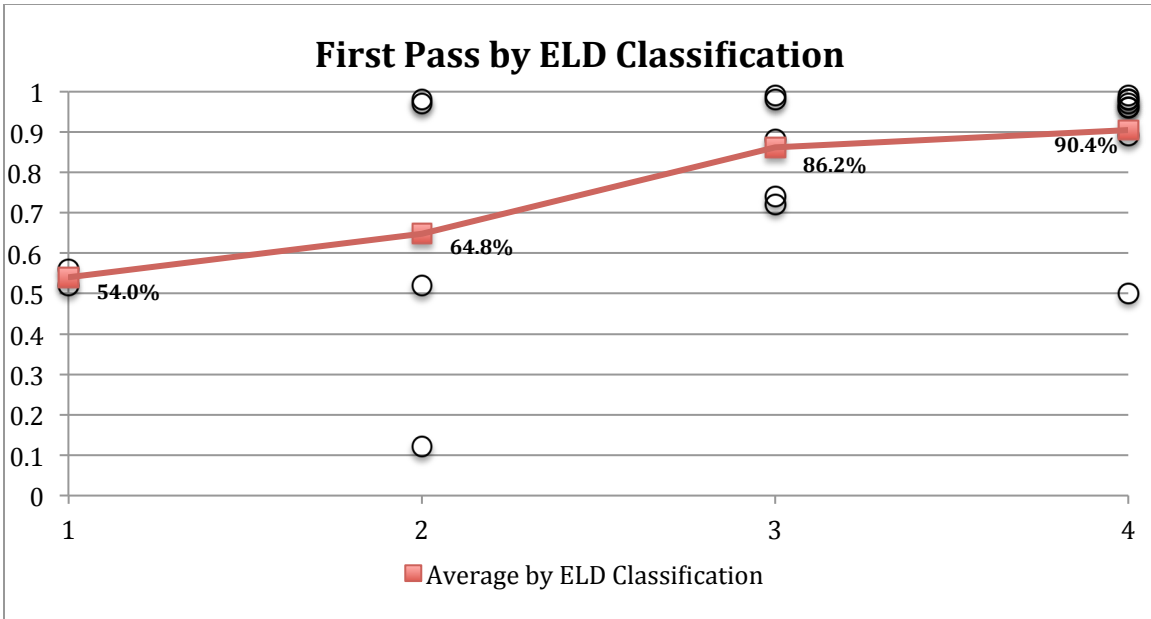


Figure 1: Babelous test scores against ELD classification. The horizontal axis represents the ELD classification and the vertical axis represents the fluency percentage score as determined by Babelous.

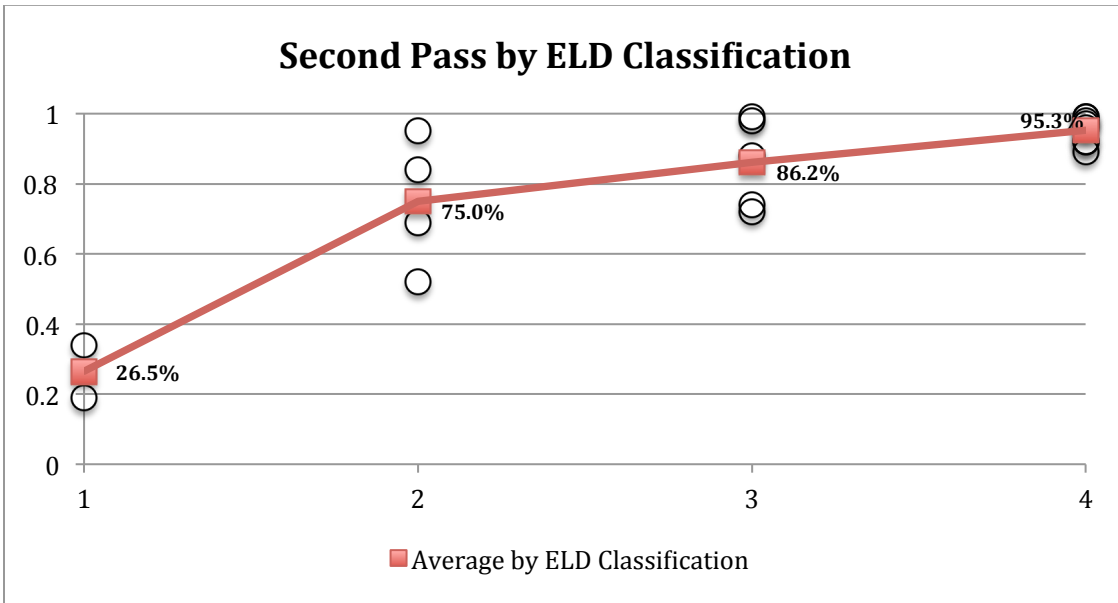


Figure 2: Babelous test scores against ELD classification. The horizontal axis represents the ELD classification and the vertical axis represents the fluency percentage score as determined by Babelous.



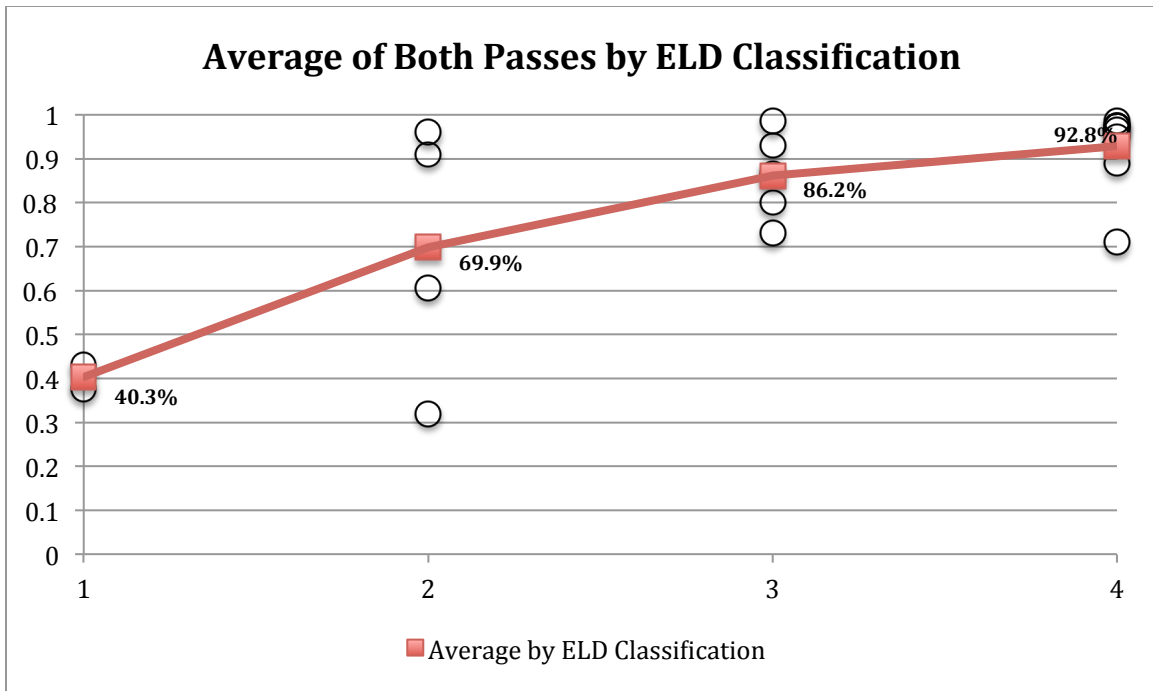


Figure 3: Babelous test scores against ELD classification. The horizontal axis represents the ELD classification and the vertical axis represents the fluency percentage score as determined by Babelous.

In general higher assessments by Babelous correlate with higher ELD classifications. Again, the range of fluency scores among ELD2s and ELD3s is widest, and indicative that the current model is not optimized. The major flaw related to the binary decision rule is highlighted well by the ELD4 student assessed at the 50% fluency level in the first pass. Reviewing the data demonstrates that in this case the ELD4 student got the first question wrong, which establishes a new ceiling score at the 50% level. The student then answered every other question correctly, but was capped at 50% according to the current binary decision rule. This same student scored a 92% in the next attempt. The case of this ELD4 student underscores the discussion in the methodology section regarding uncharacteristic right or wrong answers leading to poor fluency assessments. It also shows the usefulness of averaging multiple passes, as this student's fluency

assessment is averaged as 71% after a second pass, likely to be much closer to the student’s actual fluency level than 50%. Taking the test a third time may again add valuable information to the overall profile for this student.

Another important measure of the viability of the computer adaptive testing model is reliability. Reliability is the correlation between two test scores for the same student. This value determines how appropriately the test identifies a student’s score. A high reliability value indicates that students taking the test multiple times should expect to produce similar results each time—which generally means the model is estimating fluency level well. Below is the data used in the reliability analysis:

<b>Student Id</b>	<b>ELD Level</b>	<b>First Score</b>	<b>Second Score</b>
<b>Student 1</b>	1	0.56	0.19
<b>Student 2</b>	1	0.52	0.34
<b>Student 3</b>	2	0.52	0.69
<b>Student 4</b>	2	0.98	0.84
<b>Student 5</b>	2	0.12	0.52
<b>Student 6</b>	2	0.97	0.95
<b>Student 7</b>	3	0.33	0.51
<b>Student 8</b>	3	0.98	0.88
<b>Student 9</b>	3	0.72	0.74
<b>Student 10</b>	3	0.99	0.98
<b>Student 12</b>	3	0.54	0.79
<b>Student 13</b>	4	0.50	0.92
<b>Student 14</b>	4	0.98	0.99
<b>Student 15</b>	4	0.89	0.89
<b>Student 16</b>	4	0.97	0.98
<b>Student 17</b>	4	0.96	0.99
<b>Student 18</b>	4	0.98	0.92
<b>Student 19</b>	4	0.99	0.96
<b>Student 20</b>	4	0.96	0.97

Table 4: Reliability data.

The reliability estimate for a single administration of the test (correlation between first score and second score) is  $r=0.69$ . The reliability for two administrations averaged together is  $1-(1-r)/2=.85$ , which is comparable to the reliability of many professionally developed tests. For example, the TAKS tests that were until recently required of Texas students in grades 3-10 typically had reliabilities between .8 and .9. However, the TAKS took hours to administer on paper and results were not returned for weeks or months. Two passes of Babelous can be taken in a few minutes and results are provided immediately.

The reliability may be improved by starting the second pass where the first pass ended, or by replacing the binary search algorithm with a more sophisticated approach, for example one based on an IRT model. To accurately evaluate the reliability of a later version, it should be administered to a larger and more diverse set of users.

Another important test is the assumption of the Hick-Hyman principal, which guided the log frequency choice in the binary decision rule. Below is a chart of log frequency bands against the percent of students answering questions in the log frequency band correctly. Ideally this chart would identify a linear relationship between the log frequency value and the percent of students answering correctly:

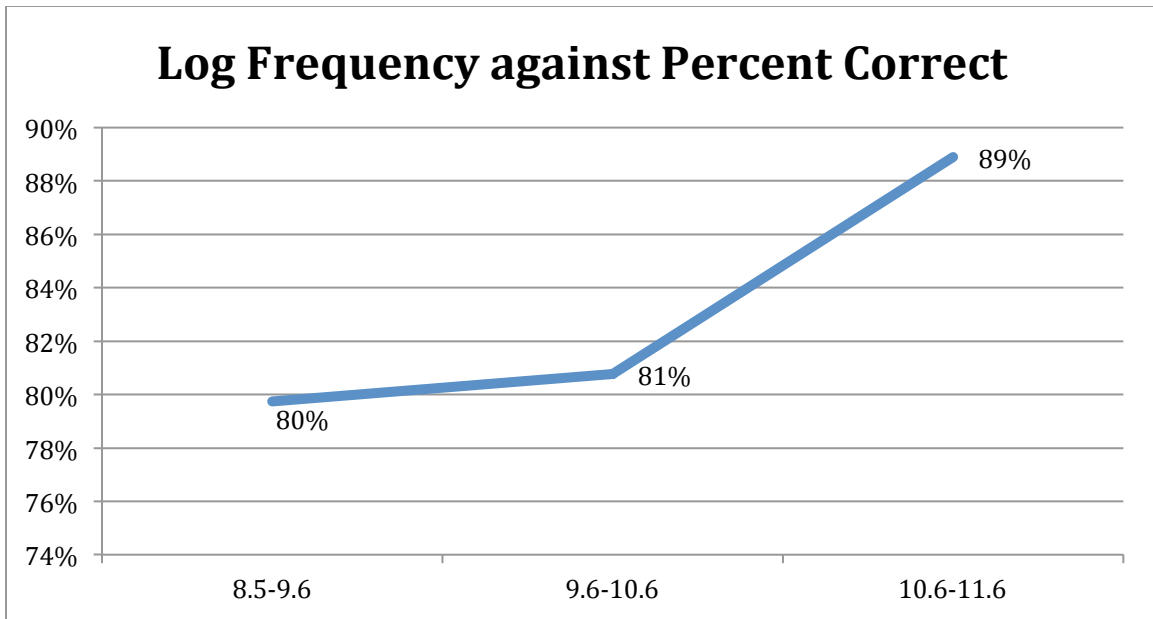


Figure 4: The above data shows the percent of users who correctly answered a word, grouped by similar log frequency values.

The graph above approaches linearity, a trend that may continue as the population of students in the sample increases to include students at all ELD levels.

Those that have used Babelous so far report it being easy to understand, functional, and highly interpretable. That is a good start because if the accuracy of the system can be improved over time, it will likely be a solution that addresses the three major weaknesses of the current ELD system.

## Chapter 4: Conclusion

The Babelous system can help teachers who currently struggle to understand the fluency makeup of their classroom, especially as more advanced decision rules replace the binary search algorithm. More precise and interpretable information, delivered efficiently at any point in the school year, can have a major impact on student success, and on teacher effectiveness. Babelous takes advantage of the growing influence computers have on improving educational outcomes. It utilizes basic adaptive methodology, with a goal to incorporate sophisticated IRT decision modeling in future iterations. Computer adaptive instruction is increasingly influential, as seen by software such as DreamBox,<sup>22</sup> designed to improve math outcomes among K-8 students by individualizing instruction in a way teachers cannot. DreamBox uses advanced adaptive technology to assess students individually by analyzing responses and response times, and presents questions that challenge students appropriately. This type of technology can understand and respond to students immediately, presenting stimulus targeted at each student's level of understanding. The future of education will include a significant software component as adaptive instruction technology advances to more and more arenas.


For this reasons the future of the Babelous software seems encouraging. Teachers and students increasingly have ways to access technology in the classroom, and are increasingly familiar with the benefits of an adaptive system. For Babelous, the next steps involve adding additional pass automation—that is that each user is given a second or third pass through the test automatically. Once this version is designed, more accurate data regarding reliability and fluency assessment can be captured and used to take the

---

<sup>22</sup> "DreamBox Learning." DreamBox Learning. Accessed April 20, 2015. <http://www.dreambox.com/>.

next step—to build an IRT model. Based on the positive results from the current binary decision model, which has obvious and significant flaws, we believe an IRT version of Babelous could revolutionize the way fluency is assessed by teachers of ELLs. It seems highly plausible that an optimized version of Babelous could significantly aid the traditional ELD system, by addressing the three major weakness, and not entirely implausible that a system similar to Babelous could eventually replace paper-based ELD tests altogether. In either case, with a few modifications, Babelous system seems likely to help teachers of English Language Learners.

## Appendix A

		<b>ACCESS for ELLs<sup>®</sup></b> English Language Proficiency Test										District: Sample District School: Memorial Middle School Grade: 6						
		<b>STUDENT ROSTER REPORT – 2011</b>																
STUDENT NAME STUDENT ID	Tier	Cluster	Listening		Speaking		Reading		Writing		Oral Language <sup>a</sup>		Literacy <sup>b</sup>		Comprehension <sup>c</sup>		Overall Score <sup>d</sup>	
			Scale Score	Prof Level	Scale Score	Prof Level	Scale Score	Prof Level	Scale Score	Prof Level	Scale Score	Prof Level	Scale Score	Prof Level	Scale Score	Prof Level	Scale Score	Prof Level
Lastname, Firstname T 123456789	B	6-8	380	5.0	359	4.3	366	5.0	373	4.4	370	4.6	370	4.5	370	5.0	370	4.6
Lastname, Firstname U 234567891	C	6-8	406	5.9	425	6.0	361	4.2	373	4.4	416	6.0	367	4.4	375	5.3	382	5.2
Lastname, Firstname V 345678912	B	6-8	380	5.0	340	3.2	354	3.7	337	3.3	360	4.2	346	3.4	362	4.2	350	3.7
Lastname, Firstname W 456789123	B	6-8	380	5.0	340	3.2	354	3.7	345	3.5	360	4.2	350	3.6	362	4.2	353	3.8
Lastname, Firstname X 567891234	A	6-8	328	3.0	239	1.5	353	3.7	325	2.9	284	1.9	339	3.2	346	3.4	332	2.6
Lastname, Firstname Y 678912345	B	6-8	380	5.0	293	1.9	354	3.7	328	2.9	337	3.2	341	3.2	362	4.2	340	3.3
Lastname, Firstname Z 789123456	B	6-8	380	5.0	381	5.2	343	3.2	356	3.9	381	5.1	350	3.6	354	3.8	359	4.0
Lastname, Firstname A 891234567	A	6-8	359	4.0	308	1.9	360	4.0	361	4.0	334	3.0	361	4.0	360	4.0	352	3.8

A - Oral Language = 50% Listening + 50% Speaking  
 B - Literacy = 50% Reading + 50% Writing  
 NA - Not Attempted - Student Booklet is marked with a Non-Scoring Code of Absent, Invalidated, Declined or Special Education/504 Exemption  
 C - Comprehension = 70% Reading + 30% Listening  
 D - Overall Score = 35% Reading + 35% Writing + 15% Listening + 15% Speaking  
 Overall Scores are computed when all 4 domains have been completed

Tier is the overall fluency assessment and is given as Tier A, Tier B, or Tier C.

## Appendix B

ELD 1 falls into the Emerging category

Mode of Communication	English Language Development →-----Emerging-----→	
	At the <i>early stages</i> of the Emerging level, students are able to:	At <i>exit</i> from the Emerging level, students are able to:
Collaborative	<ul style="list-style-type: none"> <li>• express basic personal and safety needs, ideas, and respond to questions on social and academic topics with gestures and words or short phrases;</li> <li>• use basic social conventions to participate in conversations;</li> </ul>	<ul style="list-style-type: none"> <li>• express basic personal and safety needs, ideas, and respond to questions on social and academic topics with phrases and short sentences;</li> <li>• participate in simple, face-to-face conversations with peers and others;</li> </ul>
Interpretive	<ul style="list-style-type: none"> <li>• comprehend frequently occurring words and basic phrases in immediate physical surroundings;</li> <li>• read very brief grade-appropriate text with simple sentences and familiar vocabulary, supported by graphics or pictures;</li> <li>• comprehend familiar words, phrases, and questions drawn from content areas;</li> </ul>	<ul style="list-style-type: none"> <li>• comprehend a sequence of information on familiar topics as presented through stories and face-to-face conversations;</li> <li>• read brief grade-appropriate text with simple sentences and mostly familiar vocabulary, supported by graphics or pictures;</li> <li>• demonstrate understanding of words and phrases from previously learned content material;</li> </ul>
Productive	<ul style="list-style-type: none"> <li>• produce learned words and phrases and use gestures to communicate basic information;</li> <li>• express ideas using visuals such as drawings or charts, or graphic organizers; and</li> <li>• write or use familiar words and phrases related to everyday and academic topics.</li> </ul>	<ul style="list-style-type: none"> <li>• produce basic statements and ask questions in direct informational exchanges on familiar and routine subjects;</li> <li>• express ideas using information and short responses within structured contexts; and</li> <li>• write or use learned vocabulary drawn from academic content areas.</li> </ul>



ELD 2s and 3s fall into the Expanding category

<b>nt: Proficiency Level Continuum</b> -----Expanding----->	
At the <b>early stages</b> of the Expanding level, students are able to:	At <b>exit</b> from the Expanding level, students are able to:
<ul style="list-style-type: none"> <li>• express a variety of personal needs, ideas, and opinions and respond to questions using short sentences;</li> <li>• initiate simple conversations on social and academic topics;</li> </ul>	<ul style="list-style-type: none"> <li>• express more complex feelings, needs, ideas, and opinions using extended oral and written production; respond to questions using extended discourse</li> <li>• participate actively in collaborative conversations in all content areas with moderate to light support as appropriate;</li> </ul>
<ul style="list-style-type: none"> <li>• comprehend information on familiar topics and on some unfamiliar topics in contextualized settings;</li> <li>• independently read a variety of grade-appropriate text with simple sentences ;</li> <li>• read more complex text supported by graphics or pictures;</li> <li>• comprehend basic concepts in content areas;</li> </ul>	<ul style="list-style-type: none"> <li>• comprehend detailed information with fewer contextual clues on unfamiliar topics;</li> <li>• read increasingly complex grade-level text while relying on context and prior knowledge to obtain meaning from print;</li> <li>• read technical text on familiar topics supported by pictures or graphics;</li> </ul>
<ul style="list-style-type: none"> <li>• produce sustained informational exchanges with others on an expanding variety of topics;</li> <li>• express ideas in highly structured and scaffolded academic interactions; and</li> <li>• write or use expanded vocabulary to provide information and extended responses in contextualized settings.</li> </ul>	<ul style="list-style-type: none"> <li>• produce, initiate, and sustain spontaneous interactions on a variety of topics; and</li> <li>• write and express ideas to meet most social and academic needs through the recombination of learned vocabulary and structures with support.</li> </ul>

ELD 4 falls into the **Bridging category**

Mode of Communication	English Language Development: Proficiency Level Continuum →-----Bridging-----→	
	At the <i>early stages</i> of the Bridging level, students are able to:	At <i>exit</i> from the Bridging level, students are able to:
Collaborative	<ul style="list-style-type: none"> <li>express increasingly complex feelings, needs, ideas, and opinions in a variety of settings; respond to questions using extended, more elaborated discourse</li> <li>initiate and sustain dialogue on a variety of grade-level academic and social topics;</li> </ul>	<ul style="list-style-type: none"> <li>participate fully in all collaborative conversations in all content areas at grade level with occasional support as necessary;</li> <li>participate fully in both academic and non-academic settings requiring English;</li> </ul>
Interpretive	<ul style="list-style-type: none"> <li>comprehend concrete and many abstract topics and begin to recognize language subtleties in a variety of communicative settings;</li> <li>read increasingly complex text at grade level;</li> <li>read technical text supported by pictures or graphics;</li> </ul>	<ul style="list-style-type: none"> <li>comprehend concrete and abstract topics and recognize language subtleties in a variety of communicative settings;</li> <li>read, with limited comprehension difficulty, a variety of grade-level and technical texts, in all content areas;</li> </ul>
Productive	<ul style="list-style-type: none"> <li>produce, initiate, and sustain interactions with increasing awareness of tailoring language to specific purposes and audiences; and</li> <li>write and express ideas to meet increasingly complex academic demands for specific purposes and audiences.</li> </ul>	<ul style="list-style-type: none"> <li>produce, initiate, and sustain extended interactions tailored to specific purposes and audiences; and</li> <li>write and express ideas to meet a variety of social needs and academic demands for specific purposes and audiences.</li> </ul>

## Bibliography

1. "Consortium Members." WIDA: Member States. Accessed April 26, 2015.  
<https://www.wida.us/membership/states/>.
2. "Corpus of Contemporary American English (COCA)." Corpus of Contemporary American English (COCA). January 1, 2012. Accessed April 26, 2015.  
<http://corpus.byu.edu/coca/>.
3. Davies, Mark. "Word Frequency Data." Word Frequency: Based on 450 Million Word COCA Corpus. Accessed April 26, 2015.  
<http://www.wordfrequency.info/intro.asp>.
4. "DreamBox Learning." DreamBox Learning. Accessed April 20, 2015.  
<http://www.dreambox.com/>.
5. "English Language Development Standards." Resources (CA Dept of Education). March 1, 2015. Accessed April 26, 2015.  
<http://www.cde.ca.gov/sp/el/er/eldstandards.asp>.
6. Hyman, Ray. "Stimulus Information As A Determinant Of Reaction Time." *Journal of Experimental Psychology*, 1953, 188-96.
7. Laufer, B., and P. Nation. "A Vocabulary-size Test of Controlled Productive Ability." *Language Testing*, 1999, 36-55.
8. Laufer, B., and P. Nation. "Vocabulary Size And Use: Lexical Richness In L2 Written Production." *Applied Linguistics*, 1994, 307-22.
9. Lemhöfer, Kristin, and Mirjam Broersma. "Introducing LexTALE: A Quick and Valid Lexical Test for Advanced Learners of English." *Behavior Research Methods*, 2012, 325-43.
10. Muñoz, José, Wim J. Van Der Linden, and Ronald K. Hambleton. "Handbook of Modern Item Response Theory." *European Journal of Psychological Assessment*, 1997.
11. Seow, Steven. "Information Theoretic Models Of HCI: A Comparison Of The Hick-Hyman Law And Fitts' Law." *Human-Computer Interaction*, 2005, 315-52.
12. Torlakson, Tom. "California Department of Education." *Overview of the California English Language Development Standards and Proficiency Level Descriptors*, 2012. Accessed March 11, 2015.  
<http://www.cde.ca.gov/sp/el/er/documents/sbeoverviewpld.pdf>.
13. "WIDA's 2012 Amplification of the English Language Development Standards, Kindergarten–Grade 12." 2013. Accessed March 10, 2015.  
<https://www.wida.us/standards/eld.aspx>.