RESEARCH REPORT

# Effectiveness of *On Our Way to English* as a Program for Development of Reading and Oral Proficiency by Elementary English Learners

## A Report of Randomized Experiments in a California and a Texas School District

Denis Newman
Empirical Education Inc.

Andrew Jaciw
Stanford University

May 20, 2005

Empirical Education Inc.
www.empiricaleducation.com
425 Sherman Avenue, Suite 210
Palo Alto, CA 94306
(650) 328-1734

## Acknowledgements

We are grateful to the people in the two school districts for their assistance and cooperation in this research and for providing access to their data under an agreement between Empirical Education Inc. and the districts. These experiments were sponsored by Harcourt Achieve and conducted under a subcontract to MarketingWorks, Inc. The reporting here is done as an independent research organization and the analyses reported were supported in part by a grant to Empirical Education (grant # R305E040031) from the US Department of Education, Institute for Education Sciences. The Department is not responsible for the content of this report.

### *About Empirical Education Inc.*

Empirical Education Inc. was founded to help K–12 school districts, publishers, and the educational R&D community assess new or proposed instructional programs through scientifically based pilot implementations. The company draws on the expertise of world-class researchers and methodologists assuring that the research is objective and takes advantage of current best practice in rigorous experimental design and statistical analysis. The company's findings let educators quantify the value of programs and help them partner with providers to implement those most effective for their students.

## Executive Summary

We were asked to find out whether *On Our Way to English*, a supplementary, text-based product to help elementary school students learn to read and speak English was more effective in a California and a Texas school district than the materials the districts already had in place. We conducted an experiment during the 2003-2004 school year.

**Intervention.** *On Our Way to English* (*OWE*) is a new product distinct from many existing products in its comprehensiveness and its simultaneous focus on English language skills and literacy skills. *OWE* consists of materials for kindergarten through fifth-grade classrooms and addresses oral language, reading, and writing at each level. Thematic units encompass about four weeks of instruction and contain large-format graphic organizers, chant and song charts with audio CDs, materials aimed at explicit instruction in phonics and skills, and classroom sets of leveled readers. Formal and informal assessments for teacher use are supplemented with standardized test practice. The teacher guide provides a comprehensive day-by-day sequence of activities integrating the array of resources. Teachers in our study received a one-day in-service session led by a consultant-trainer from the publisher.

**Settings.** In California, the research site was an urban center where 30% of students are designated as English learners, a percentage representative of the state population overall. California control group teachers used the existing materials, which consisted of the recently adopted reading textbook (Houghton Mifflin) supplemented by ELD materials such as Hampton Brown and teacher made materials. California's testing program for all English learners, the California English Language Development Test (CELDT), includes measures of reading, writing, and listening and is given in early fall of each year.

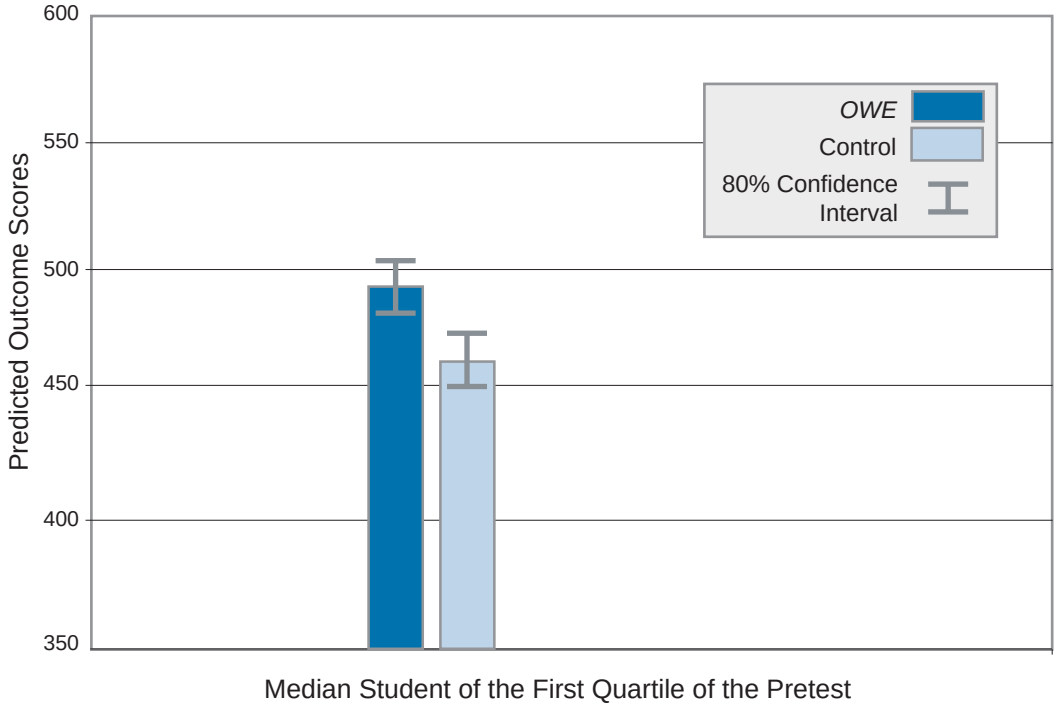
In the Texas site, also an urban center, approximately 9% of the students are designated as English learners. Most Texas control group teachers used materials they developed or collected themselves. Immersion teachers had a set of English-language leveled readers as well as the reading basal (Harcourt), which was used by some bilingual teachers, who also had a Spanish version of a similar text and a library of leveled readers in both English and Spanish. The district administers STAR Reading, a computer-adaptive English language reading test, to all students at the beginning and end of the school year. In addition, the individually administered Idea Proficiency Tests (IPT) Oral addresses vocabulary, comprehension, grammar/syntax, and verbal expression.

**Research design.** Our research design was a randomized experiment (or randomized controlled trial). This type of study is the best way to assure that the new program and not some characteristic of the teachers caused the differences observed between groups. Teachers who volunteered to participate were assigned by coin toss to the *OWE* group or to the control group. The process of assignment included a mechanism that assured approximately equal numbers of classes in the two grades involved in the study (second and fourth) and in bilingual vs. English immersion.

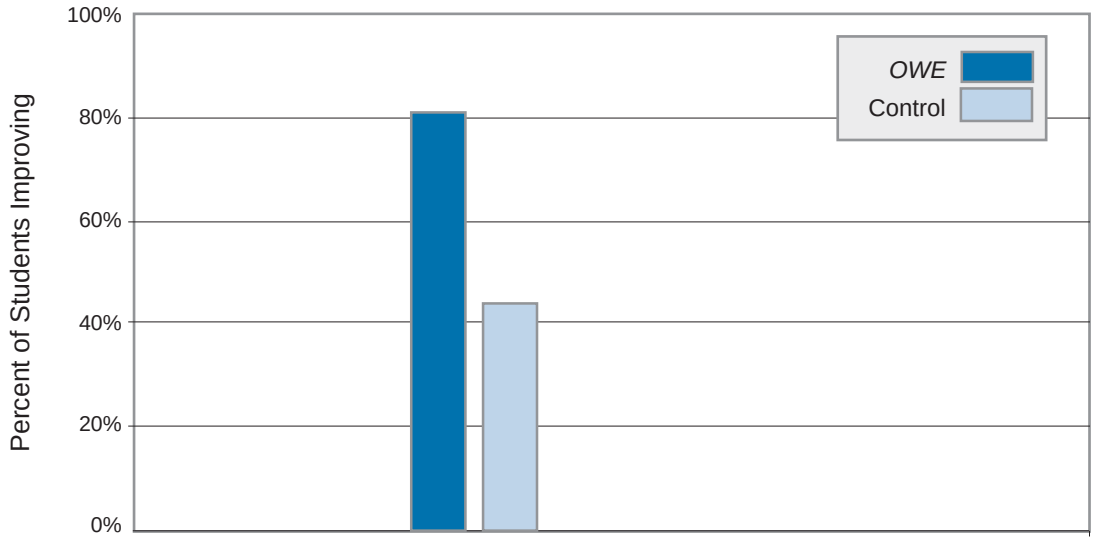
**Participants.** In the California study, a total of 384 English learners and 29 teachers participated. This included 6 bilingual classes and 21 immersion classes. In the Texas study, a total of 169 students and 20 teachers participated. This included 7 bilingual and 13 immersion classrooms.

**Statistical analysis.** We analyzed data from the two districts separately because of differences in the outcome measures. We used a variety of statistical methods but primarily a mixed model statistical analysis that involved two levels—students and classes. For both districts we performed separate analyses for reading and oral proficiency. Our statistical models distinguished between bilingual and immersion implementations and always made use of student pretest scores.

**Results.** The findings from the two settings were similar: *OWE* is generally as effective as the control programs for reading proficiency and is generally more effective than control in improving oral proficiency. In California the value of *OWE* in promoting oral proficiency was greater in the immersion setting. In that setting, the effect was greater for the students starting with low proficiency. The following graph shows the results for the students in the bottom quartile of the immersion classes.




However, in the Texas district, where immersion students were, for the most part, already proficient, the positive impact of *OWE* is observed for bilingual students. The following graph shows the comparison of *OWE* and control conditions for these students in terms of the percentage making gains during the year.



**Conclusion.** The primary value of *OWE* in these two districts was found in its effectiveness in improving oral proficiency. In interpreting these results it is important to consider them in the context of the different implementations—bilingual versus immersion—that were used. These implementations and the learning environment they provided differed also between the two districts. Variations in ways of serving English learners among districts suggest caution in generalizing these results to other locations. In assessing a new program such as *OWE*, it is also necessary to consider the programs already in place and the relative level of existing proficiency of the students. For the districts participating in this study, our recommendation is to focus the use of *On Our Way to English* in the area of oral proficiency.

# Effectiveness of *On Our Way to English* as a Program for Development of Reading and Oral Proficiency by Elementary English Learners

## A Report of Randomized Experiments in a California and a Texas School District

# Table of Contents

## Introduction

We report here on a research program aimed at producing scientifically based evidence of the effectiveness of *On Our Way to English* (*OWE*). *OWE* is a supplementary, text-based product designed to help in teaching literacy skills to elementary school students who are learning to speak English. Two school districts, one in California and the other in Texas, agreed to conduct structured pilots involving a small number of their teachers with the goal of determining whether the program was more effective than the programs already in place. We conducted year-long randomized experiments in the two districts. We measured the students' English learning by the local and state-mandated achievement tests. We were also interested in the effects of how the program was used and whether its impact was related to the proficiency level at which the student started.

The specific question addressed is whether elementary school English learners in classrooms using *OWE* would learn English faster than if they continued using the programs already in place in their districts. Secondarily to this question, we were interested in the context of instruction: in both districts, some English learners were in bilingual programs while others were in immersion programs, and the new program may be better suited to one implementation than the other. The other variable of interest was the students' initial level of English proficiency. We wanted to find out whether *OWE* would prove better suited for early or later English learners. Cross-cutting these questions were the measures of English proficiency, particularly reading and oral proficiency. It was important to know whether *OWE* was more effective for improving some aspects of English language development than others.

The primary purpose of this research is to advise the districts about wider deployment of the program beyond the small sample of students and teachers in the experiment. As an effectiveness study, the research placed *OWE* in a realistic implementation context, using each district's own standards for English proficiency and the districts' methods of assigning students to English learner programs. The design of the experiments reflects the requirements of the No Child Left Behind Act, which directs schools to consult reports of rigorous research in making adoption decisions about instructional programs. The US Department of Education (2003) has been explicit in interpreting this requirement in terms of randomized experimentation for determining effectiveness. In a randomized experiment, we reduce selection bias by tossing a coin to assign teachers to use the program or to continue using their current program. This design is considered the appropriate design for valid conclusions about effectiveness (Shadish, Cook & Campbell, 2003). Nevertheless, given the specifics of the implementations and the particular characteristics of the two districts studied here, we do not intend for the research by itself to apply generally to other districts.

## Method

### *Research Design*

Our study compared outcomes for students taught using the *OWE* program with outcomes for students taught using other materials. Second- and fourth-grade teachers volunteered for participation. From this pool of volunteers, we randomly assigned equal numbers of teachers to *OWE* (treatment) and control groups.

We conducted two independent experiments for this study. In the two districts, the basic question was the same: was *OWE* more effective than the programs currently in use? Because each district had different

programs already in place, the contrasts differed. Moreover, each district used a different testing program to measure English proficiency. Other important differences involved the students' home languages and the way students were assigned to English learner programs. Our approach was to work within each district, making use of their test results to answer their questions, without a primary concern with generalizable results. We anticipated, however, that if there were commonalities in the results in the two districts it would strengthen our conclusion about those aspects of the program.

## Materials

The *OWE* program was a new product that was distinct from many existing programs in its comprehensiveness and its simultaneous focus on English language skills and literacy skills. The product consists of materials for kindergarten through fifth-grade classrooms, at each level addressing oral language, reading, and writing. Thematic units that encompass about four weeks of instruction contain large-format graphic organizers, chant and song charts with audio CDs, materials aimed at explicit instruction in phonics and skills, and classroom sets of leveled readers. Formal and informal assessments for teacher use are supplemented with standardized test practice. The teacher guide provides comprehensive day-by-day sequence of activities integrating the array of resources. The teachers in the treatment group—those who were assigned to use *OWE* materials—received a one-day in-service session led by a Harcourt consultant-trainer.

## Site Descriptions

The research sites for this study were identified by the publisher's research and sales staff as interested in the product and willing to conduct a structured pilot with a subset of their classrooms. California and Texas were targeted because of their high concentrations of English learners.

### California District

The California site was an urban district with more than 80,000 students enrolled. The ethnic makeup of 54% Hispanic, 11% African American, 18% White, and 16% Asian includes a large Hmong population from Laos. The district has 30% English learners, and 76% are eligible for free or reduced-price lunch.

Classes in this district were identified as either bilingual or "structured English immersion," a formal definition that did not necessarily reflect a strong distinction in the concentration of non-English speakers. Several classes, for example, were composed of students who, as a group, had been a bilingual class the year before. (Because of the pressure in California to move students out of bilingual settings, the classes were re-designated as English immersion with the addition of a few native English speakers. One teacher explained that, for her, the major difference was that now she was expected to avoid speaking Spanish, although she did so when a student did not understand the instruction.) Another important factor in the California district was the large Hmong population, who were also English learners, but for whom bilingual classrooms were not an option.

The district had recently adopted the Houghton Mifflin textbook as its basal reading program and made it available in all classrooms. The control group teachers used a variety of materials for English instruction, including the English learner materials in the basal, teacher-made materials, and the Houghton Mifflin program.

### Texas District

The Texas site was an urban center of approximately 73,000 residents. The school district enrolls about 16,600 students, of whom 21% are Hispanic, 9.2% are African American, 67% are White, and 2.2% are Asian. About 9% of the students are English learners and 21% of the students are eligible for free or reduced-price lunch, a measure of their socioeconomic status.

Students are screened using an oral assessment when entering the district, and periodically thereafter, as the criterion for assignment as an English learner. Spanish-speaking English learners are, by default, assigned to a bilingual program. Non-Spanish-speaking English learners are placed in an immersion program, which is taught by a teacher with English language teaching qualifications but in a mainstream classroom. (Some Spanish-speaking parents request that their child be placed in an English immersion classroom, which is usually accommodated.) As a result, classes for "bilingual" and "immersion" students are distinct and their learning environments are quite different.

At the time of our study, this district had not yet adopted a set of materials for English language development. Most teachers depended on materials they developed or collected themselves. The elementary reading basal (Harcourt Collections, in English), adopted three years previously, was used in the English immersion classrooms and, to a lesser extent, in the bilingual classrooms. Bilingual classrooms used a Spanish version of a similar text as well as a library of leveled readers in both English and Spanish. English immersion classrooms also used a set of leveled readers (in English).

## *Random Assignment of Teachers*

We met with district staff members and principals to explain the details and procedures of the study. Principals identified eligible teachers, who were then invited to after-school meetings. The kick-off meeting for the research experiment in the California district occurred on October 14, 2003 with 20 teachers. The district was able to recruit nine more teachers, who met with us separately on November 6, 2003 and underwent the same process. A full day of training was conducted November 10 for all but three of the teachers (for whom a make-up session was arranged). A similar kick-off meeting in Texas, conducted on November 4, 2003, was attended by 20 teachers.

In both cases, the meeting began with a presentation of *OWE* by a Harcourt sales representative, followed by a discussion of the research procedures led by the research team. After a question-and-answer period, teachers who decided to participate engaged in a discussion of the important factors that they believed would affect the results. They were then separated into groups that went to different parts of the room. First, they were divided by the most important factor, which in both cases was grade level. Each grade group was further divided into bilingual and English immersion classroom teachers. The resulting four groups were instructed to form pairs of teachers who were most similar on the remaining important factors. These additional factors were decided by the teachers, who organized their own pairs of participants on the basis of similarity in their approaches to classroom teaching. Once the pairs were established, we tossed a coin to decide which member of the pair joined the *OWE* group and which one joined the control group. Where the group had an uneven number, we used a coin toss to decide the assignment of the unpaired member. This procedure allowed the groups to be both randomly assigned and equivalent in terms of the distribution of important factors.

### Data Collection

The data for these experiments were primarily those collected and provided to us by the school district staff. They consisted of classroom rosters, test scores, and demographic information on the students. In addition to conducting one-time teacher interviews, we also collected monthly web-based surveys from all participating teachers in each group.

### Demographics

The districts provided basic information of all students, including their age, sex, ethnicity, and home language. Student-level data for socioeconomic status, measured in terms of participation in the free/reduced-price lunch program, were available in Texas but not in California, due to privacy concerns.

### Classroom Assignments

The research team was diligent in identifying which students were in which classrooms during English language instruction. In the case of the California district, information from the district's student record database was supplemented by communication with each individual teacher. In some schools, students are sent to teachers outside their homeroom for English instruction. One teacher, originally listed as teaching a bilingual classroom, was found to be working out of a resource room with several different clusters of students. (We determined that these arrangements were set up prior to the random assignment and were not done in response to the availability of *OWE*.) In California, we also found a high level of mixed-grade classrooms such that a group designated as second grade may contain first graders. Third graders are often combined with either second or fourth graders. Because our measures of English proficiency formed a continuous scale across grades, this blurring of grade-level distinction did not impact our results.

### Test Scores

As noted below, we had measures of English reading proficiency and oral language or listening skills in both districts. In California we also had a writing score and an overall English language development score (a combination of the listening, reading, and writing sections of CELDT).

**California:** California English Language Development Test (CELDT), a state-wide test developed and scored by CTB/McGraw-Hill, provided the primary outcome as well as pre-test measure for the California experiment. Students in our sample took this test between August and October of 2003 and again in the same window in 2004. Student performance is reported as a scale score which, as a test of English language proficiency, can be used as a single scale for students through this age range. The results are reported in three subscales: listening, reading, and writing. The overall score is a weighted combination with listening counting for 50% and the other two subtests counting for 25% each.

**Texas:** STAR Reading, a computer-adaptive English language reading test given to all students at the beginning and end of the school year, provides a measure of growth through a single year. STAR Reading provides a scale score and a grade equivalent score. Because only grade equivalent scores were available from the district, we used the published conversion tables to convert them back to scale scores for use in our analysis (Renaissance Learning, 2003). (Some students entering school with no English language skills scored below the lowest level of the test. School staff members verified that these students had been exposed to the test so as to distinguish these cases from missing data, as might result when a student was absent.)

**Texas:** The IPT (Idea Proficiency Tests) Oral is individually administered and addresses vocabulary, comprehension, grammar/syntax, and verbal expression (Ballard, Dalton, & Tighe, 2001). Students are tested through a maximum of six levels and scored at the highest level completed. For students in grades 2 to 6, scoring in the first three levels is considered "Non-English Speaking." The next two levels are considered "Limited English Speaking." The sixth level is designated "Fluent English Speaking." The district administers the IPT Oral at the beginning and end of the school year and uses the test as a placement tool for incoming students.

In Texas, a variety of other tests are given but, unlike in California, the criteria for identification for inclusion in an English language development (ELD) program are not based on test results. Student identification for an ELD program is based on a district oral test, whereas graduating from the program is determined by achievement on written standardized English language tests. Even with adequate scores on such tests, students are not pushed out of bilingual programs, which are viewed as providing other advantages such as first language literacy. Because the state-mandated Reading Proficiency Test in English (RPTE), which is administered to Texas students in grades 3 through 12 each April, has a much lower standard for passing than the district English learners program, many district-identified English learners do not take the RPTE. In fact, students in bilingual programs typically take different tests from those taken by the immersion students. The Texas state achievement test (Texas Assessment of Knowledge and Skills, or TAKS) is given both in English and Spanish, with the immersion students taking the English version and the bilingual students generally taking the Spanish version. Similarly, the Developmental Reading Assessment (DRA) is given in English and Spanish (Evaluación del Desarrollo de la Lectura), depending on the program. Conversely, the students in English immersion programs do not take tests that measure their progress in Spanish literacy.

### Implementation

We were able to track the use of the *OWE* materials in treatment group classrooms and the use of alternative products in the control classrooms, as well as potential contamination, through periodic web-based surveys and interviews with teachers and principals. Five web-based surveys were conducted of all participating teachers during the experiment. These continuous updates allowed us to identify teachers leaving the project and to take corrective actions.

### Statistical Analysis

In each district, we had outcome measures for different aspects of English language development, particularly reading and oral proficiency. Each of these outcomes was addressed in a separate analysis. Across these outcomes, the basic question for the statistical analysis was whether students in the *OWE* classrooms had higher scores than those in the control classrooms. Recognizing that the strength of the impact of *OWE* (treatment effect) could depend on levels of other factors, we developed statistical models that took into account each student's incoming proficiency level as well as information about the type of classroom implementation (bilingual vs. immersion). An analysis of covariance including these variables potentially increases the precision of the estimated treatment effect. Models that also examine the interaction of these factors with treatment give separate estimates of the impact for the different subgroups. The statistical models were multi-level because they accounted for the clustering of students in classes, which provides a more accurate, and often more conservative, assessment of the confidence we should have in the findings. We based our decisions about which covariates to include on the goal of maximizing precision and the need to look at results for certain subgroups. There were several other potential factors that were not found to have an impact that made a difference to our interpretation of

the results and were not used in the models. Beyond these, we construct exploratory models to better understand unexpected results. We use SAS PROC MIXED (SAS Institute, 2003) as the primary tool for this work.

## Results

We begin with an examination of the groups that resulted from the randomization in order to establish our starting point. We then consider the observations of the implementations and the responses to our interviews as the context for the more quantitative analyses, which are reported separately for each district.

### *Formation of the Experimental Groups*

The number of teachers and students in the experimental groups in the two districts are given in the following tables. These are broken down by school, class, experimental condition, and whether or not the classrooms were designated as bilingual or English immersion. These counts of students include only those for whom we had pre- and post-intervention test data. We performed statistical tests to determine the equivalence of the *OWE* and control groups that are reported here.

#### California Groups

In California, the teacher selection and randomization procedure resulted in 29 classrooms and 379 students in the sample. Of these students, 68 did not have a score on the pretest. While these students were distributed quite evenly between the *OWE* and control groups, there were two teachers, one *OWE* teacher in the immersion group and one control teacher in the bilingual group, for whom none of their students had pretest scores. These teachers could not be used in the analysis. Table 1 represents the sample of students and their teachers for whom we had pretest scores.

**Table 1: Participants in California *OWE* and control groups**

| Implementation | Bilingual | | Immersion | | Total | |
|---|---|---|---|---|---|---|
| | Number of Teachers | Number of Students | Number of Teachers | Number of Students | Number of Teachers | Number of Students |
| *OWE* | 5 | 60 | 9 | 88 | 14 | 148 |
| Control | 3 | 49 | 10 | 119 | 13 | 168 |
| **Total** | **8** | **109** | **19** | **207** | **27** | **316** |

The almost equal number of students (and classrooms) in the two conditions – *OWE* and control – was assured by the randomization process. The main goal of random assignment is to get close to equal distributions between the pilot and the control of other factors that affect the outcome. With randomization we expect equal distributions on average, but in any single randomization there may be discrepancies between the distributions due to chance. This was the case here – the two groups differed on the pretest scores on the CELDT as shown in Table 2. This imbalance is not critical, provided that the pretest score is included in the analysis of covariance. To limit bias it is also desirable for students with missing data to be as similar as those who are included in the analysis on the covariates that matter. This was the case in this analysis.

**Table 2: Independent t test of the difference between *OWE* and control groups on the California district CELDT overall pretest scores**

| Descriptive statistics: CELDT overall outcomes | Raw Group Means | Standard Deviation | Number of Students | Standard Error | |
|---|---|---|---|---|---|
| *OWE* | 468.750 | 50.184 | 148 | 4.125 | |
| Control | 486.679 | 48.550 | 168 | 3.746 | |
| ***t* test for difference between independent means** | **Difference** | | | ***t* value** | ***p* value** |
| Condition (*OWE* – control) | -17.929 | | | --3.22 | 0.0014 |

This test shows that there is a substantial initial difference between the *OWE* and control group; this difference was controlled for statistically in our analysis.

### Texas Groups

There were fewer student participants in the Texas district. As in California, the number of *OWE* vs. control classrooms were fairly evenly matched. In each condition, bilingual vs. English immersion were evenly represented. As shown in Table 3, there were a total of 169 students in 20 classrooms in the study.

**Table 3: Participants in the Texas control and *OWE* Groups**

| Implementation | Bilingual | | Immersion | | Total | |
|---|---|---|---|---|---|---|
| | Number of Teachers | Number of Students | Number of Teachers | Number of Students | Number of Teachers | Number of Students |
| *OWE* | 4 | 62 | 6 | 20 | 10 | 82 |
| Control | 3 | 55 | 7 | 32 | 10 | 87 |
| **Total** | **7** | **117** | **20** | **52** | **20** | **169** |

In comparison to the Texas English immersion classes, the California classes were on average much larger in terms of the number of English learners (but not in the total number of students). This is consistent with the policy in California to place English learners in regular classrooms and with the fact that the California district had a much larger population of students from non-English-speaking ethnicities.

## *Attrition*

In California we documented two cases where teachers left the district or were reassigned. In both cases, the use of *OWE* was taken up by the replacement teacher. We assumed that as long as the student experience was continuous (the classroom continues to use or not use the materials), this teacher attrition would not substantially affect the results. There were no cases of teacher attrition in the Texas district.

Among the California students, 30 students took the CELDT pretest but not the posttest. The distribution of these students was not related to condition. In Texas, 11 students were lost between the STAR pretest and posttest. These were also unrelated to condition.

## *Program Implementation*

Because teachers were not given strict guidelines as to implementation approach and because they represented very different classroom settings, we expected variation in the patterns of usage. According

to the surveys and interviews, the variations in the English immersion classrooms were of greatest interest. In two of the Texas classrooms, teachers reported that they did not often find a need to use the *OWE* program since their English learners were already advanced and participated with the rest of the class in using the regular reading text. In several other immersion classes, especially in California, an opposite pattern was observed. Teachers found the program to be useful for their whole class and felt it benefited their native English speaking students who were weak readers.

Surveys and interviews also revealed that most teachers were very enthusiastic about the program and reported that it was well designed and very attractively presented, provided great variety, and was highly motivating to students. Most teachers were very positive about the program's integrated approach to oral language development, literacy, and content area learning. Many identified the leveled readers as an important strength of the program.

Our informal survey of student engagement indicated that the bilingual teachers in the *OWE* group described their students as very engaged. The English immersion classes and the bilingual control classes rated their students, on average, as somewhat engaged. The amount of time spent on English instruction did not differ between the *OWE* and control groups.

### Differences in Approach to Bilingual Instruction

The interviews of teachers and district administrators revealed differing approaches to the role of bilingual education. The distinction between bilingual and immersion classrooms in Texas was much stronger than in California. In Texas, many of the students labeled as English language learners in the English immersion classrooms were reasonably fluent in English to start with and were often integrated into English instruction either by being grouped with native-English-speaking poor readers or by being mainstreamed in the use of the standard reading text. These students were in the English immersion classes either because their native language was not Spanish or because their parents had specifically requested that they not be placed in a bilingual program. These students tended also to be of a higher socio-economic status than the students in the bilingual classrooms. Their classrooms generally had only a small number of English learners.

While the two implementations for English learners were qualitatively distinct in Texas, in California, many of the immersion classrooms actually contained very high proportions of non-English speakers so that student day-to-day exposure to English and the percentage of students in need of English instruction formed a gradation rather than a clear distinction. As a check on the impact of this non-dichotomous situation in California, we developed a new variable that placed all the classrooms on a continuous scale according to percentages of students with good proficiency in English. Using the ELD status designations, we categorized students at the Intermediate or Advanced level and above (and native English speakers) as English speakers and the three lower ELD categories: Early Production, Pre Production, and Speech Emergence as non-English speakers. Thus we had a new variable representing the extent that students were immersed in an English-speaking context. The t test in Table 4 shows that the control group was slightly more immersed in English than the *OWE* group.

**Table 4: Independent t test for the difference in percent of English speakers in California *OWE* and control classrooms**

| Descriptive statistics: Percent of English speakers | Raw Group Means | Standard Deviation | Number of Students | Standard Error | |
|---|---|---|---|---|---|
| *OWE* | 0.373 | 0.272 | 168 | 0.0198 | |
| Control | 0.418 | 0.257 | 148 | 0.0224 | |
| ***t* test for difference between independent means** | **Difference** | | | ***t* value** | ***p* value** |
| Condition (*OWE* – control) | 0.045 | | | 1.52 | 0.128 |

### Year-round Schools with No Summer Break

A complicating factor in California was the fact that some of the schools were year-round and some of their classes were in session over the summer just prior to the CELDT testing period. Classes in traditional schools and most of the year-round classes had a summer break prior to the test. We expected students in class all summer to score higher on the CELDT than those who had been on break. For each class in the year-round schools, we determined their track or schedule of vacations, thus we determined which students were in school in the months immediately prior to the outcome measure and which students were on vacation and likely immersed in their home language. Table 5 shows the distribution of these students between *OWE* and control. Students in the control group were slightly more likely to have no break before the test but the difference is not significant.

**Table 5: Chi square of the distribution of students with no summer break between *OWE* and control**

| Condition | Break before the CELDT | | Totals |
|---|---|---|---|
| | Summer Break | No break | |
| *OWE* | 124 | 24 | **148** |
| Control | 131 | 37 | **168** |
| **Totals** | **255** | **61** | **316** |
| **Chi-square statistics** | | **value** | ***p* value** |
| | | 1.35 | 0.245 |

## *Outcomes for California*

We first report the results for the California site and then the results for the Texas site. The discussion section addresses commonalities found in the results.

For each setting, we have multiple outcome measures that allow us to separate the impact of *OWE* on different aspects of English language development. In California, the CELDT provides an overall score as well as sub-scores for listening/speaking, reading, and writing.

### CELDT Overall Score

Table 6 presents the results of our statistical modeling of the overall CELDT score. The first part of the table shows the contrast between *OWE* and control groups in terms of their mean scores and includes

other descriptive details. Our statistical analysis of these scores used a model that took into account the pretest score as well as clustering of the students in classes. The bottom rows of this table contain the technical details needed for review. The row in the table labeled "Condition" gives us the information about whether *OWE* made an overall difference in test scores. The coefficient associated with the treatment is 4.675, which is the difference between the average student in both groups after we adjust for the pretest score. This is a small difference and the p value, which gives us the probability of finding a difference this large or larger simply by chance, indicates that we cannot distinguish this difference from zero. In conventional terms, this result is not statistically significant.

**Table 6: Multi-level mixed model for the CELDT overall score—results for the condition controlling for the pretest**

| **Descriptive statistics for CELDT overall outcomes** | **Raw Group Means** | **Standard Deviation** | **Number of Students** | **Number of Teachers** | |
|---|---|---|---|---|---|
| *OWE* | 492.91 | 47.88 | 130 | 14 | |
| Control | 503.63 | 36.38 | 139 | 13 | |
| **Mixed model: Fixed factors related to CELDT overall outcomes** | **Estimate of coefficient** | **Standard error** | **DF** | ***t* value** | ***p* value** |
| Intercept | 494.877 | 4.928 | 25 | 100.410 | <.0001 |
| Pretest score (centered at the mean) | 0.704 | 0.043 | 241 | 16.203 | <.0001 |
| Condition (OWE = 1; control = 0) | 4.675 | 6.950 | 241 | 0.672 | 0.502 |
| **Mixed model: Technical details for random components** | **Estimate of variance component** | **Standard error** | | ***z* value** | ***p* value** |
| Class mean achievement | 246.891 | 88.219 | | 2.798 | 0.003 |
| Within class variation | 573.971 | 52.072 | | 11.022 | <.0001 |

*Note: 287 students had both pre and post tests. 18 cases were removed as outliers or influential points.*

We present a simple model here because including additional factors or interactions among the factors did not give us different results.

To help visualize this result, we present in Figure 1 a scatterplot that shows where all the students fell in terms of their starting point (on the horizontal x-axis) and their outcome measure (on the vertical y-axis). This graph therefore shows the differences in growth among the students. On top of the scatterplot, we then superimpose lines representing *OWE* and control groups in terms of what the statistical model predicts a student's outcome score will be, given where he or she started on the pretest scale. (For this graph, we modeled the interaction between pretest and condition because we wanted to show how the two lines cross and diverge slightly, even though it was not enough to impact our results.) In any case, differences in the height of these lines represent the very small difference we found between *OWE* and control.

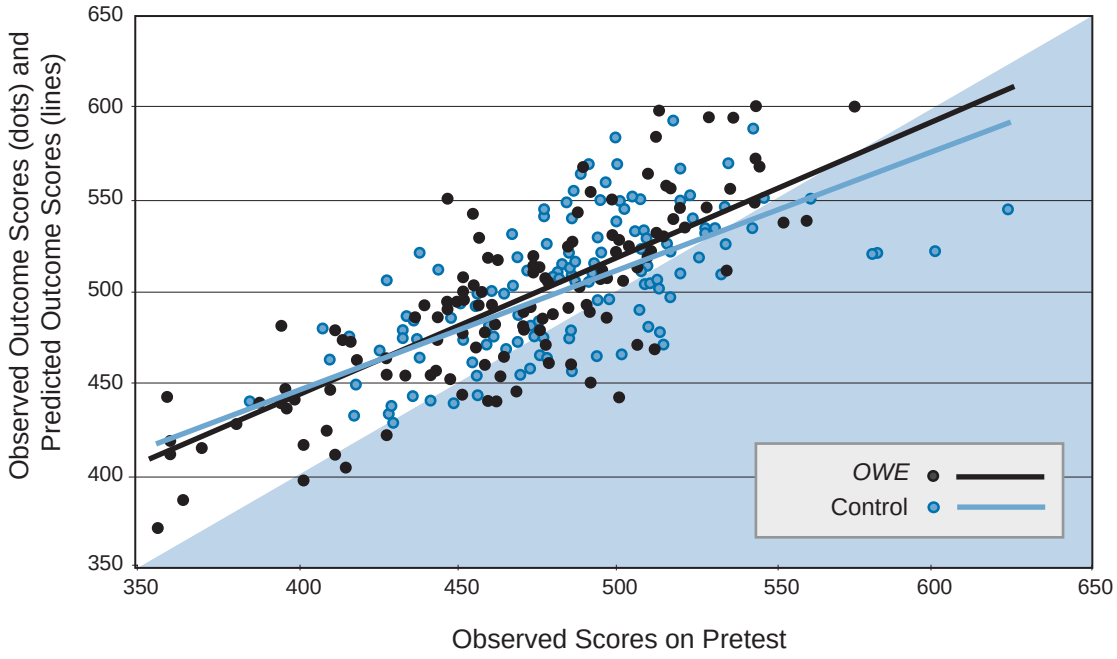



**Figure 1: CELDT overall score—scatterplot of *OWE* and control students with lines showing the predicted values based on pretest score**

An additional element in the graph is the shaded area in the bottom right. Students in this area demonstrated negative growth over the course of a year.

The overall score is a composite of three subtests: reading, writing, and listening/speaking. We can look more closely at the subtests that compose this overall score to get a better sense of where *OWE* has the greatest impact.

### Reading

The reading score constitutes 25% of the overall score. Table 7 presents the statistical model for the reading test. These results are very similar to what we found for the overall score. The statistical model adjusts for the pretest and, when that is taken into consideration, the results do not indicate any difference between *OWE* and control.

**Table 7: Multi-level mixed model for the CELDT reading score—results for the condition controlling for the pretest**

| Descriptive statistics for CELDT overall outcomes | Raw Group Means | Standard Deviation | Number of Students | Number of Teachers | |
|---|---|---|---|---|---|
| *OWE* | 479.65 | 46.82 | 119 | 14 | |
| Control | 494.23 | 41.02 | 137 | 13 | |
| **Mixed model: Fixed factors related to CELDT overall outcomes** | **Estimate of coefficient** | **Standard error** | **DF** | ***t* value** | ***p* value** |
| Intercept | 488.226 | 4.928 | 25 | 141.878 | <.0001 |
| Pretest score (centered at the mean) | 0.673 | 0.043 | 228 | 16.198 | <.0001 |
| Condition (OWE = 1; control = 0) | -2.767 | 6.950 | 228 | -0.558 | 0.578 |
| **Mixed model: Technical details for random components** | **Estimate of variance component** | **Standard error** | | ***z* value** | ***p* value** |
| Class mean achievement | 68.250 | 45.939 | | 1.486 | 0.069 |
| Within class variation | 779.127 | 72.661 | | 10.723 | <.0001 |

*Note: 287 students had both pre and post tests. 31 cases were removed as outliers or influential points.*

As with the overall CELDT score, including additional factors in the model does not alter the result.

### Writing

The writing sub-test accounted for 25% of the overall CELDT score. Table 8 shows the results of our statistical model. Overall, *OWE* gives positive results; that is *OWE* has a 7-point advantage with a p value of .177, which means there is about an 18% chance that an impact with an absolute value this large or larger would happen by chance. We found two other factors (that did not play a role in our analysis of reading) to be important to consider. First, it made a difference whether the student was in a bilingual or immersion class. Second, the impact of the treatment (*OWE* or control) was different depending on the implementation. This is shown in the row labeled “Condition by Implementation interaction”.

**Table 8: Mixed-level model for the CELDT writing score—results for the condition controlling for the pretest**

| Descriptive statistics for CELDT writing outcomes | Raw Group Means | Standard Deviation | Number of Students | Number of Teachers | |
|---|---|---|---|---|---|
| *OWE* | 491.79 | 50.86 | 139 | 15 | |
| Control | 504.49 | 42.28 | 146 | 13 | |
| **Fixed factors related to CELDT writing outcomes** | **Estimate of coefficient** | **Standard error** | **DF** | ***t* value** | ***p* value** |
| Intercept | 495.293 | 3.479 | 23 | 142.348 | <.0001 |
| Pretest score (centered at the mean) | 0.583 | 0.034 | 252 | 16.960 | <.0001 |
| Condition (OWE = 1; control = 0) | 7.077 | 5.231 | 252 | 1.353 | 0.177 |
| Implementation (Bilingual = 1; Immersion = 0) | 13.671 | 6.962 | 252 | 1.964 | 0.051 |
| Condition by Implementation interaction | -22.306 | 9.189 | 252 | -2.427 | 0.016 |
| **Mixed model: Technical details for random components** | **Estimate of variance component** | **Standard error** | | ***z* value** | ***p* value** |
| Class mean achievement | 22.072 | 31.284 | | 0.706 | 0.240 |
| Within class variation | 960.979 | 84.287 | | 11.401 | <.0001 |

*Note: 287 students had both pre and post tests. 2 cases were removed as outliers or influential points.*

The low p values for implementation and the interaction between this and the condition suggest that value of *OWE* cannot be understood without considering how these factors work together. Figure 2 is a picture of the model without considering implementation. The lines are close to parallel regardless of the pretest score the student started with.

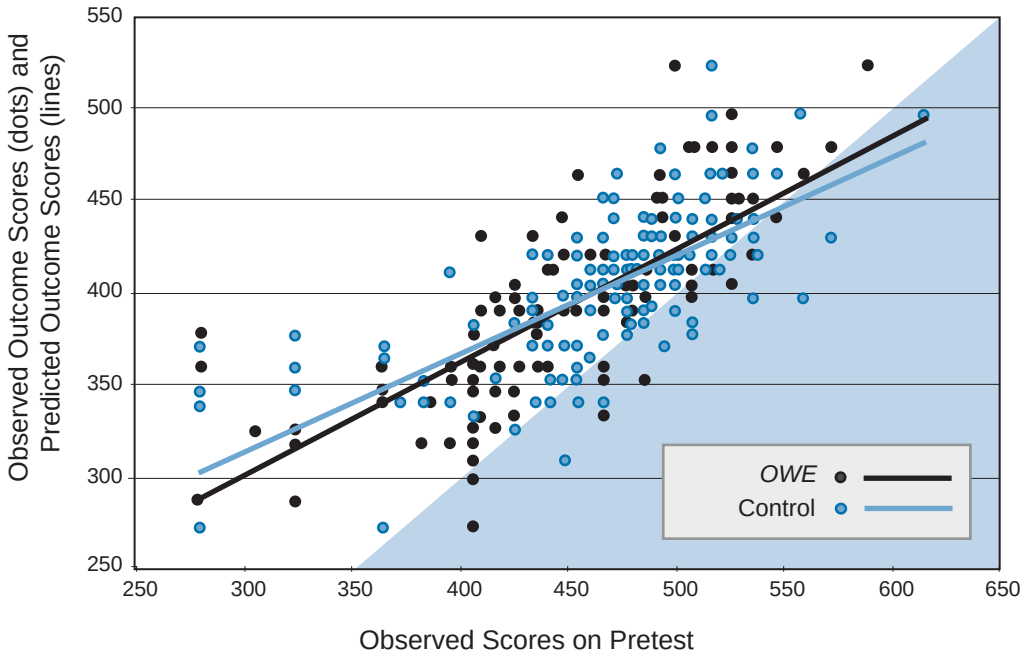



**Figure 2 CELDT Writing score—scatterplot of *OWE* and control students with lines showing the predicted values based on pretest score**

Figure 3 presents the condition by implementation interaction from Table 8 as a set of bar graphs (for values at the mean of the pretest score). On the left are the mean writing scores for *OWE* and control bilingual classes. On the right are *OWE* and control immersion classrooms. We can see that the impact for *OWE* in the immersion classrooms was in the positive direction, whereas the existing program had a stronger effect in the bilingual classrooms.

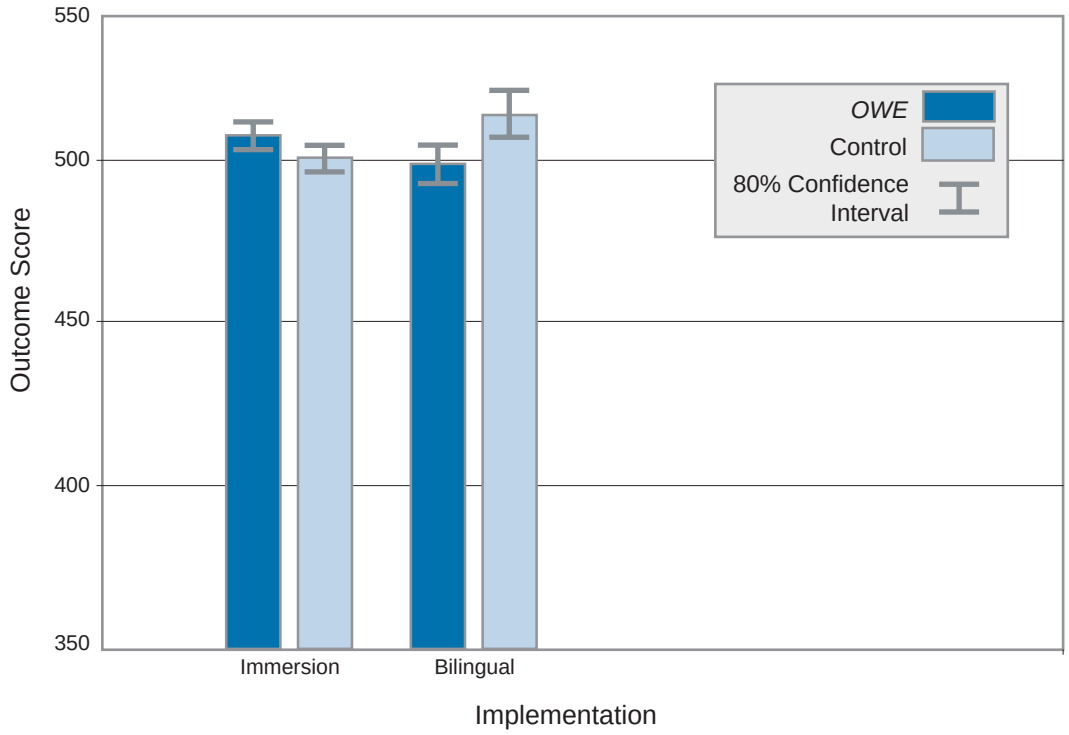



**Figure 3: CELDT Writing—bar graphs showing the difference between *OWE* and control for the immersion and bilingual groups**

The confidence we can have in these differences is indicated by the markers on the top of each of the bars. We know from Table 8 that the p value for the interaction is quite low, indicating that it is unlikely that an interaction this large would have happened by chance without there being a real difference. The markers show the 80% confidence interval. (In other words, there is a 4 in 5 chance that the correct value falls within that interval.) The intervals for *OWE* and control groups in the immersion classes overlap, indicating that there is a reasonable probability that the difference we see may not actually be different from zero. For the bilingual group, because the intervals do not overlap, we have confidence that there is a disadvantage of *OWE* for that group. (The reader is reminded, however, that ignoring immersion- or bilingual-status, the average impact of *OWE* is slightly positive).

### Listening/speaking

The final sub-test of the CELDT was a test of proficiency in oral English. This score counted for 50% of the overall score. In Table 9, the statistical model that best describes the results is more complex than the previous model used to explain the results for writing. In this model, we used all the factors (pretest, condition, and implementation). We also examined each of the interactions among pairs of these factors. Finally, we included the three-way interaction for which the p value is low enough to consider further investigation.

**Table 9: Multi-level mixed model for CELDT Listening—results for condition including pretest, implementation, and interactions**

| Descriptive statistics for CELDT listening outcomes | Raw Group Means | Standard Deviation | Number of Students | Number of Teachers | |
|---|---|---|---|---|---|
| *OWE* | 507.38 | 47.50 | 138 | 15 | |
| Control | 502.92 | 50.63 | 130 | 13 | |
| **Fixed factors related to CELDT listening outcomes** | **Estimate of coefficient** | **Standard error** | **DF** | ***t* value** | ***p* value** |
| Intercept | 498.697 | 10.105 | 23 | 49.353 | <.0001 |
| Pretest score (centered at the mean) | 0.750 | 0.112 | 237 | 6.694 | <.0001 |
| Condition (OWE = 1; control = 0) | 14.416 | 14.690 | 237 | 0.981 | 0.327 |
| Implementation (Bilingual = 1; Immersion = 0) | -5.823 | 21.075 | 237 | -0.276 | 0.783 |
| Pretest score by condition interaction | -0.315 | 0.139 | 237 | -2.265 | 0.024 |
| Implementation by condition interaction | -8.436 | 27.469 | 237 | -0.307 | 0.759 |
| Pretest score by implementation interaction | -0.170 | 0.232 | 237 | -0.733 | 0.465 |
| Three-way interaction | 0.452 | 0.277 | 237 | 1.630 | 0.104 |
| **Mixed model: Technical details for random components** | **Estimate of variance component** | **Standard error** | | ***z* value** | ***p* value** |
| Class mean achievement | 826.640 | 295.506 | | 2.797 | 0.003 |
| Within class variation | 1144.805 | 105.203 | | 10.882 | <.0001 |

*Note: 287 students had both pre and post tests. 19 cases were removed as outliers or influential points.*

In the presence of a three-way interaction, the clearest way to understand the results is in terms of a set of graphs that allow us to illustrate all the effects simultaneously. We have done this by dividing the information into two sets of graphs, one for the bilingual and one for the English immersion classrooms. In both sets of graphs lines are plotted representing *OWE* and control groups. In Figure 4, we can see that, for the bilingual classrooms, there is very little difference between *OWE* and control groups. For the English immersion students, however, as shown in Figure 5, there is greater separation between the lines, indicating that, for these classes, the *OWE* students generally out-performed control group students.

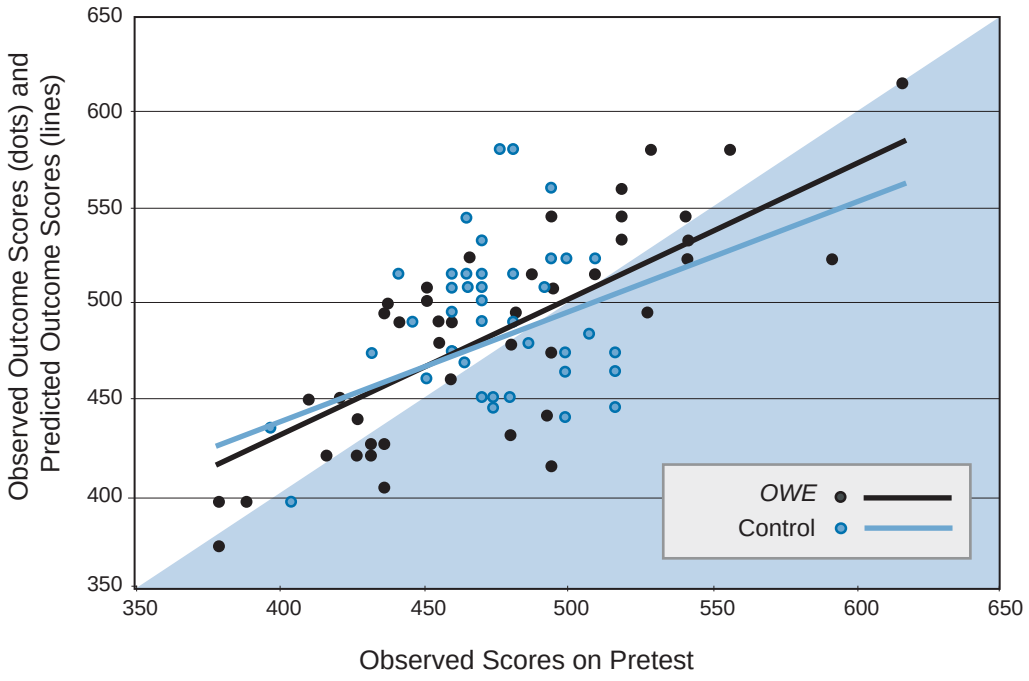



**Figure 4: CELDT Listening score for the bilingual group—scatterplot of *OWE* and control group students with lines showing the predicted values based on pretest score**

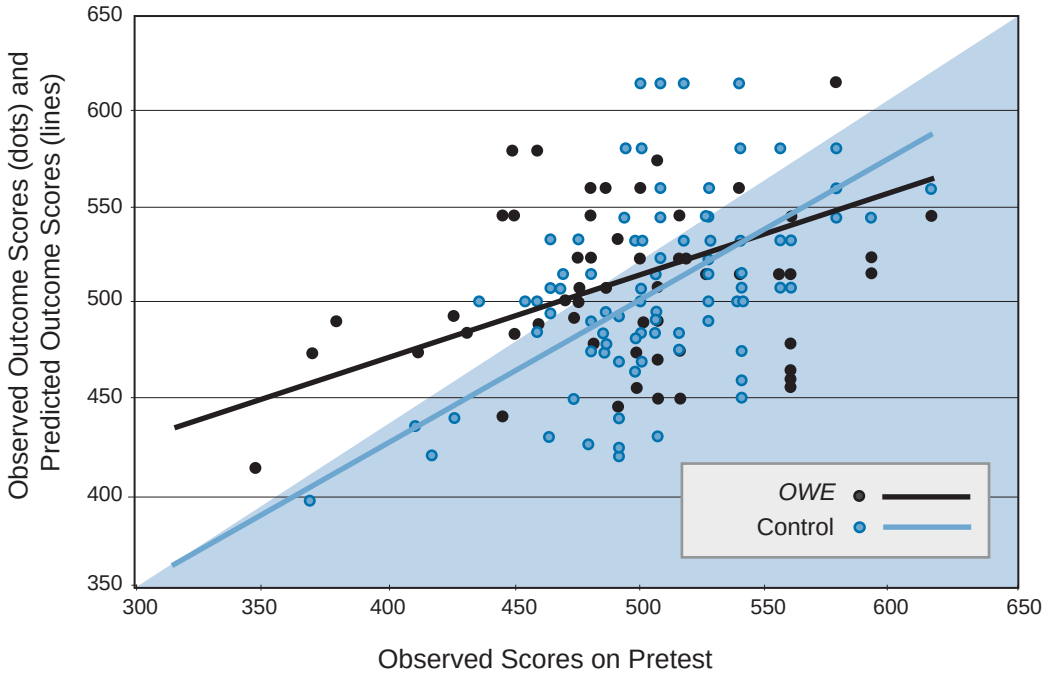



**Figure 5: CELDT Listening score for the immersion group—scatterplot of *OWE* and control students with lines showing the predicted values based on pretest score**

From these graphs we can see that the impact of *OWE* that resulted in the three-way interaction is confined to the immersion group where *OWE* appears to have had the largest impact on students initially scoring low on the listening portion of the CELDT.

A second graph, Figure 6, is a representation of this separation as a difference, shown by the dark line. This difference represents the predicted outcome for an *OWE* student minus the outcome for a control student in an immersion classroom. Around the difference line, we provide gradated bands representing confidence intervals. The shaded bands represent how likely the difference indicated by the black line could have happened just by chance. These confidence intervals are an alternative way of expressing what is often called statistical significance. The band with the darkest shading surrounding the black line is the "50-50" area, where the difference is considered equally likely to lie within the band as not. As we move out to the lighter bands, the likelihood increases that the true difference exists within the bands. The outer band represents conventional significance for which there is only a 5% chance that the true value of the difference lies outside the band. We can be reasonably confident that, for the students in the lower part of the listening scale, there was a measurable difference.

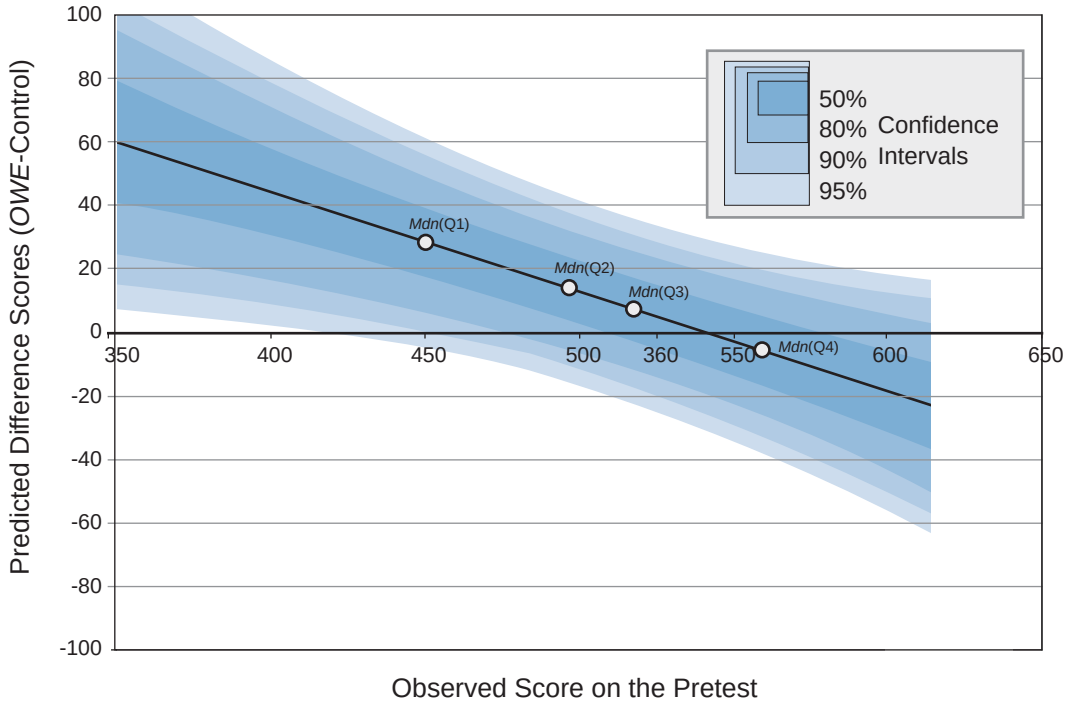



**Figure 6: CELDT Listening score for the immersion group—difference between *OWE* and control showing the values for the median student at each quartile of the pretest**

Figure 6 also indicates the locations of the median student in each of the quartiles of initial listening/ speaking skills.

We can represent the impact of *OWE* for the median student in the bottom quartile of incoming oral language skills when placed in an immersion class as a bar graph in Figure 7. As with Figure 3, we include the 80% confidence interval. In this case, this interval is simply an alternative way of representing the confidence bands in Figure 6.

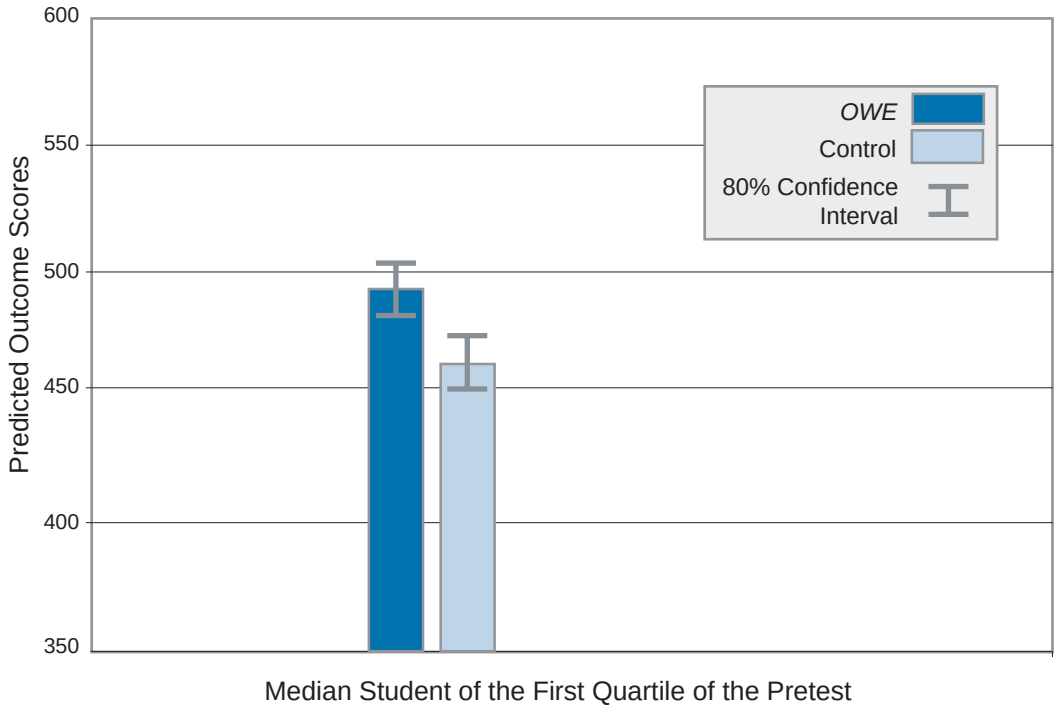



**Figure 7: CELDT Listening for immersion group—bar graph showing the difference between *OWE* and control for the median student of the bottom quartile**

Overall, we find that when there is a difference, it tends to be more positive for the immersion implementation than for the bilingual. In bilingual classes, the program either makes no discernable difference or the program already in place performs better.

## *Outcomes for Texas*

Our outcome measures in Texas were based on tests given to all students in our sample at the beginning of the year and at the end. STAR Reading tested reading ability, whereas IPT Oral tested oral proficiency in English. These skills parallel the measures of reading and listening in the California site.

### Reading

The pre- and post-experiment reading scores were scale scores from the computer-adapted test, STAR Reading, which makes it possible to include students from different grades on the same scale. The test was given to all students, regardless of English learner status. STAR Reading provides a scale score and a grade equivalent score. As previously noted, because only the grade equivalent scores were available from the district, we used the published conversion tables to convert these back to scale scores for use in the analysis.

Students who had little or no knowledge of written English and were unable to complete the practice pages for any reason were treated as missing data after closer inspection of their performance. To compare these students with the others, we assigned them a scale score equivalent of the first percentile for the grade. Most of this group of 59 students showed very little change through the year. Nine of them were still unable to take the test at the end of the year. The exceptions were four students who made greater than a three-grade equivalent gain in the year. Students with such massive gains were considered outliers and were removed from the analysis. The remaining 46 of the 59 students who did register a

score on the posttest averaged a gain (posttest score minus pretest score) of only 4.7 scale score points, compared to a gain of 93.8 points for the students who scored non-zero in the initial test. We concluded that we did not know enough about these students to include them in the analysis. In addition, as shown in Table 10, the randomly selected control classrooms were somewhat more likely to contain students for whom we had no score for STAR Reading. Knowing that these students made very little gain in the year, we can assume that eliminating them would be a conservative choice with respect to showing an effect of treatment.

**Table 10: Chi square test of the distribution of readers in *OWE* and control conditions**

| Condition | Reading test performance | | Totals |
|---|---|---|---|
| | Did not complete the practice | Got a non-0 score | |
| *OWE* | 24 | 55 | **79** |
| Control | 35 | 44 | **79** |
| Totals | 59 | 99 | **158** |
| **Chi-square statistics** | | **value** | ***p* value** |
| | | 2.71 | 0.10 |

*Note: 158 students had post test scores including scores of 0.*

Thus we built our statistical model for the STAR Reading analysis using the results only for the students who scored greater than zero on the pretest. Of the original 169 students, our sample for this model is only 105 as shown in the descriptive section of Table 11. These students had a non-zero pretest score and took the post test. Even with this small sample, we were able to support a fairly complex model that included not only pretest, condition, and implementation but the interactions among them. The three-way interaction was not used in the model since it did not add any explanatory value beyond the three two-way interactions.

**Table 11: Multi-level mixed model for STAR Reading—results for condition including pretest, implementation, and interactions**

| Descriptive statistics for STAR Reading outcomes | Raw Group Means | Standard Deviation | Number of Students | Number of Teachers | |
|---|---|---|---|---|---|
| *OWE* | 276.08 | 149.51 | 60 | 10 | |
| Control | 322.74 | 163.10 | 45 | 10 | |
| **Mixed model: Fixed factors related to STAR Reading outcomes** | **Estimate of coefficient** | **Standard error** | **DF** | ***t* value** | ***p* value** |
| Intercept | 306.546 | 16.071 | 15 | 19.075 | <.0001 |
| Pretest score (centered at the mean) | 1.013 | 0.074 | 74 | 13.667 | <.0001 |
| Condition (OWE = 1; control = 0) | 29.233 | 21.399 | 74 | 1.366 | 0.176 |
| Implementation (Bilingual = 1; Immersion = 0) | 2.868 | 22.593 | 74 | 0.127 | 0.899 |
| Pretest score by condition interaction | -0.234 | 0.103 | 74 | -2.276 | 0.026 |

| | | | | | |
|---|---|---|---|---|---|
| Pretest score by implementation interaction | 0.222 | 0.109 | 74 | 2.038 | 0.045 |
| **Mixed model: Technical details for random components** | **Estimate of variance component** | **Standard error** | | ***z* value** | ***p* value** |
| Class mean achievement | 318.945 | 300.222 | | 1.062 | 0.144 |
| Within class variation | 2757.337 | 434.750 | | 6.342 | <.0001 |

*Note: Of the 169 students in the sample, 59 had zero scores on the pretest and were removed. Of the remainder, 5 did not have a posttest score.*

Given the interactions revealed by this model, the results are most readily interpreted through inspection of graphs. Since the bilingual and English immersion classrooms were expected to be different, we constructed two graphs, one for each of these settings. Figure 8 shows the scatterplot for the bilingual students.

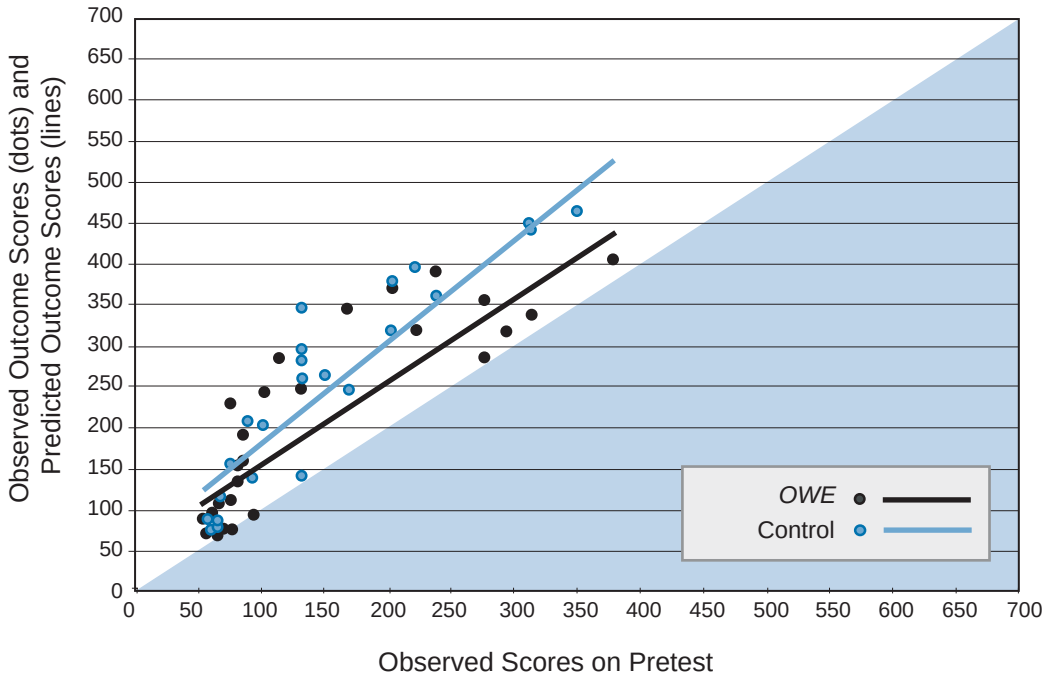



**Figure 8: STAR Reading for the bilingual group—scatterplot of *OWE* and control students with lines showing the predicted values based on pretest score**

We can see that, among students in bilingual classrooms, a large cluster of them began with low performance on STAR Reading and made very little progress during the year. This pattern is similar to those students initially getting a score of zero. However, we found the conclusion from this analysis is the same even when these low scoring students are removed. It is clear that many students make almost no progress as measured by this test. For the progress that is made, the materials used in the control classrooms outperformed *OWE*.

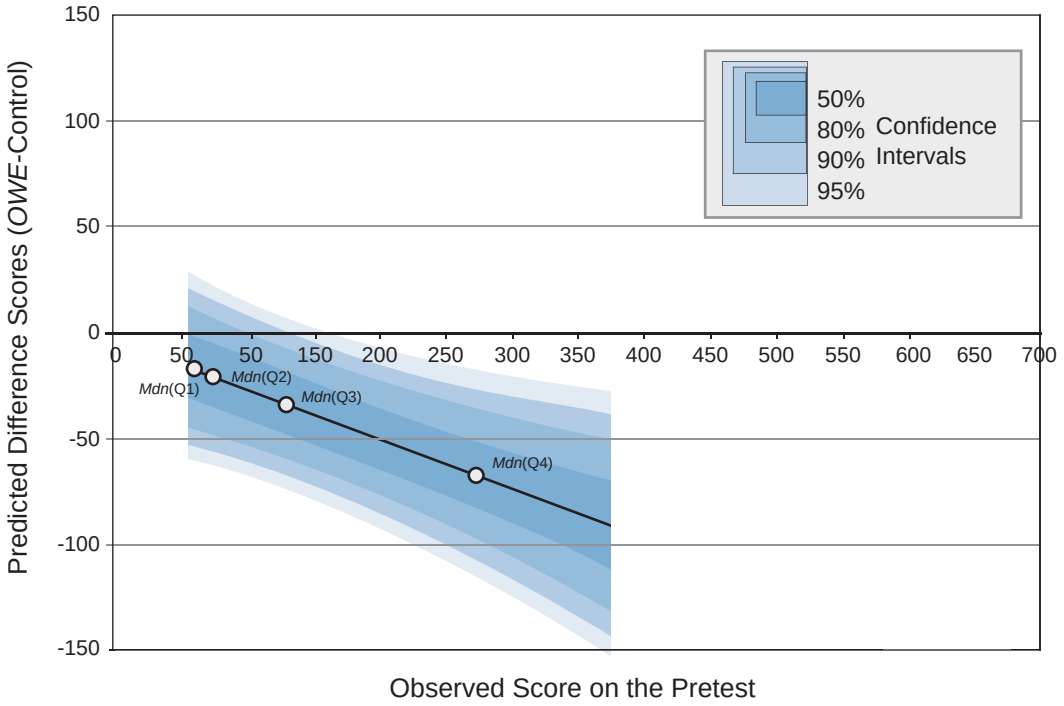



**Figure 9: STAR Reading scores for the bilingual group—difference between *OWE* and control showing the values for the median student at each quartile of the pretest**

Figure 9 represents the difference between *OWE* and control as a function of the pretest score. The advantage for the control program is strongest for the students starting out with higher reading scores. As this figure indicates, the median student in the top quartile performs substantially better with the control program.

Figure 10 represents the same information from Figure 9 as a bar graph just for the student at the median of the top quartile. Translating this difference back into the grade equivalent score corresponds approximately to a seven-month advantage for the control group.

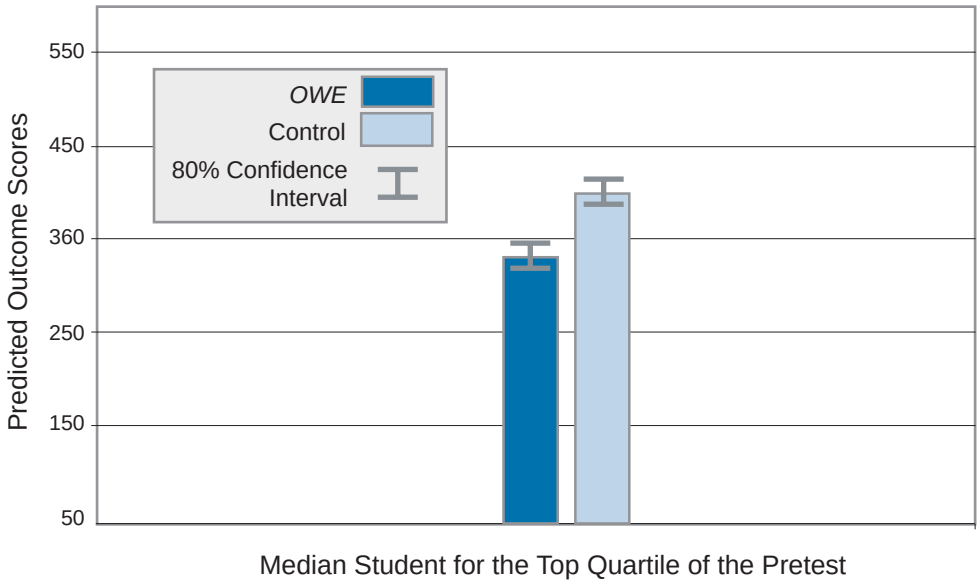



**Figure 10: STAR Reading for the bilingual group—bar graph showing the difference between *OWE* and control for the median student in the top quartile on the pretest**

In Figure 11, we use a scatterplot to examine student performance under the other implementation, immersion. Here we observe that the immersion students' scores are spread out more evenly and over a wider range than the scores of students in bilingual classes.

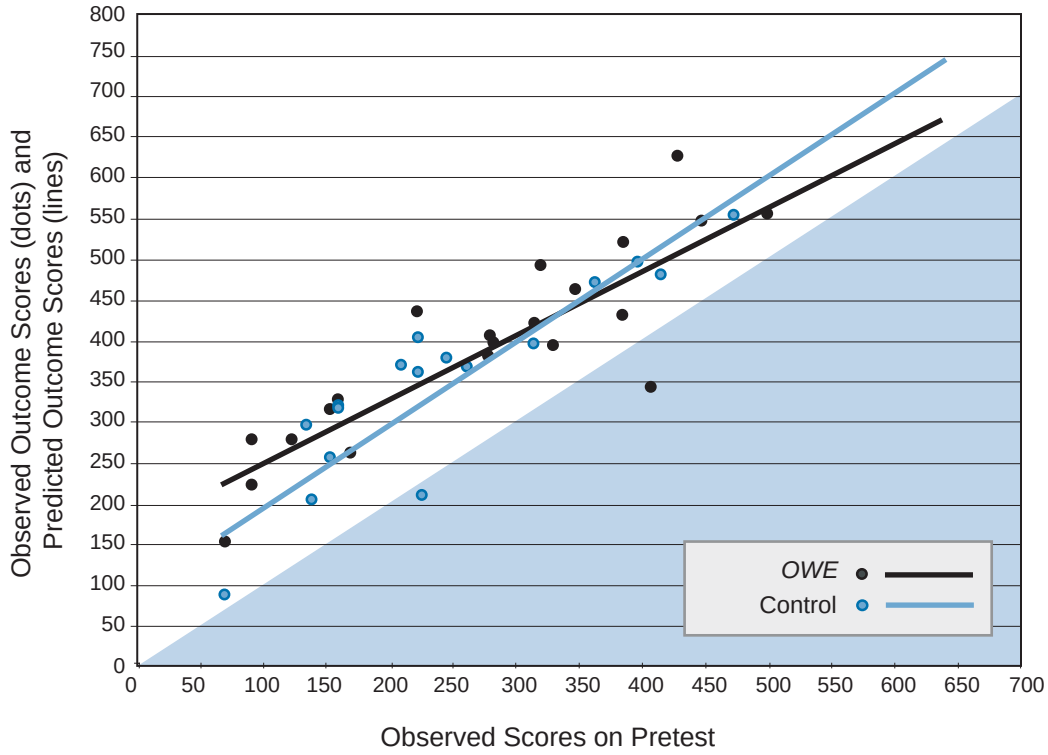



**Figure 11: STAR Reading for the immersion group—scatterplot of *OWE* and control students with lines showing the predicted values based on pretest score**

In this case, the lines representing the *OWE* and control groups cross toward the higher end of the reading proficiency range. When represented as a difference line in Figure 12, we can see that, although the median students in the bottom two quartiles may stand to gain from *OWE*, the programs performed similarly across the rest of the initial reading levels.

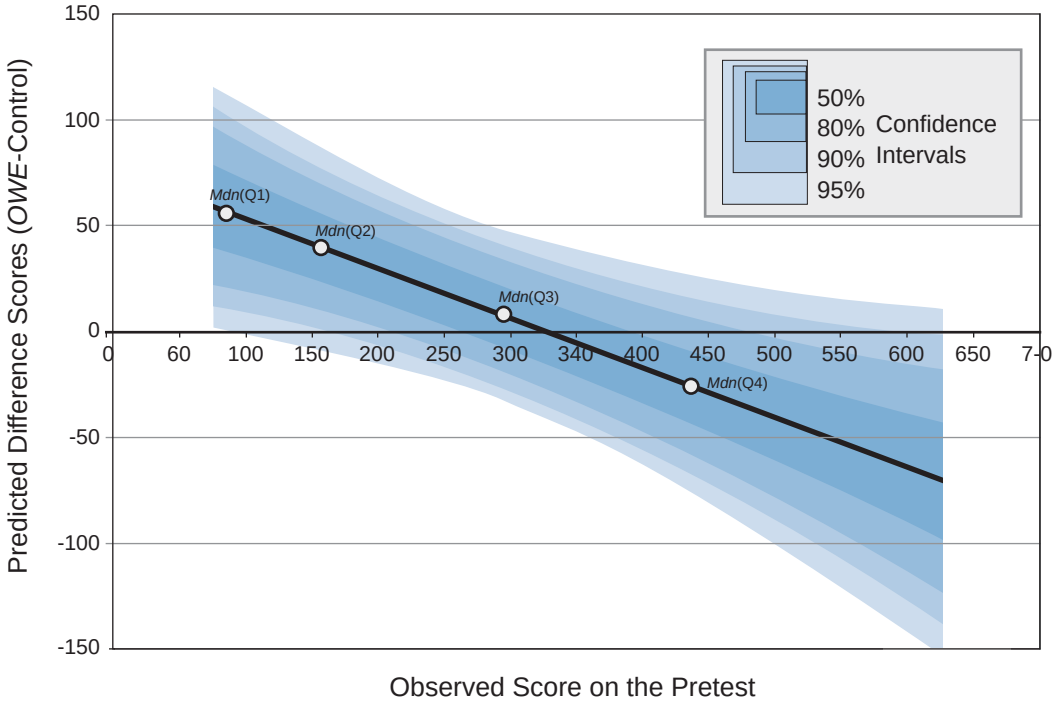



**Figure 12: STAR Reading scores for the immersion group—difference between *OWE* and control showing the values for the median student at each quartile of the pretest**

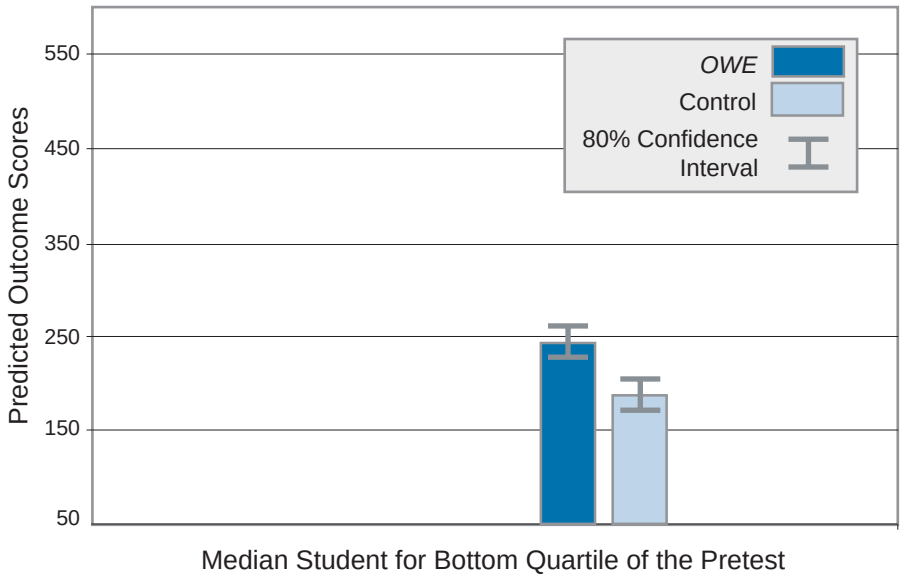



Figure 13: STAR Reading for immersion group—bar graphs showing the difference between *OWE* and control for the median student in the bottom quartile on the pretest

Figure 13 shows the prediction for the median student in the bottom quartile. The confidence interval markers are an alternative representation of the 80% confidence bands in Figure 12. Translating this difference back into the grade equivalent scale reveals an advantage of approximately two months for the *OWE* group.

The two implementations were quite different in terms of the students served. The immersion classes had a wider range of students and more students at the higher end of the reading ability scale. The bilingual classes had a substantial number of students at the very bottom of the scale who showed very little progress in reading during the year as measured by STAR Reading. Neither program was successful with these students. It is clear that the two settings function differently, and that a program that succeeds in improving reading skills for one will not necessarily be successful for the other.

### Oral proficiency

The district also tested all the English learners for oral proficiency at the beginning and end of the year. This test provides outcomes in terms of a six-point scale, where a student scoring at level six is considered a proficient English speaker. The bilingual and English immersion classrooms were also very different with respect to oral proficiency. Table 12 shows that many of the students identified as English learners in the immersion classes were already proficient speakers before the experiment began. By the end of the year, all but five students tested as proficient.

**Table 12: Chi square test of the distribution of initial oral proficiency between the implementations**

| Oral proficiency on pretest | Implementation | | Totals |
|---|---|---|---|
| | Bilingual | Immersion | |
| Less than fully proficient | 95 | 28 | **123** |
| Fully proficient | 12 | 20 | **32** |
| Totals | 107 | 48 | **155** |
| **Chi-square statistics** | | **value** | ***p* value** |
| | | 16.94 | <.0001 |

Thus we concluded that the English immersion sample was not an appropriate population for an experiment on oral proficiency development. Because the bilingual classes were assigned randomly and equally represented across conditions, we analyzed the results from those classes as a separate experiment.

The oral proficiency test results formed an ordinal scale, so we converted the outcomes to a dichotomous variable. We subtracted the student's pretest level from his or her posttest level to obtain an indication of growth. Students were then divided between those who moved up in proficiency (first, second, or third levels) and those who stayed the same or dropped a level (two students in the sample). Table 13 shows that *OWE* students were far more likely to have improved than students in the control group.

**Table 13: Comparison of the number of bilingual students gaining greater oral proficiency by condition**

| Condition | Change in Oral Proficiency | | Totals |
|---|---|---|---|
| | Growth | No Growth | |
| *OWE* | 39 | 9 | **48** |
| Control | 26 | 33 | **59** |
| Totals | 65 | 42 | **107** |
| **Chi-square statistics** | | **value** | ***p* value** |
| | | 13.83 | <.0002 |

We can represent this relationship graphically in terms of the percentage of students in each condition who improved in their oral English proficiency. This is shown in Figure 15.

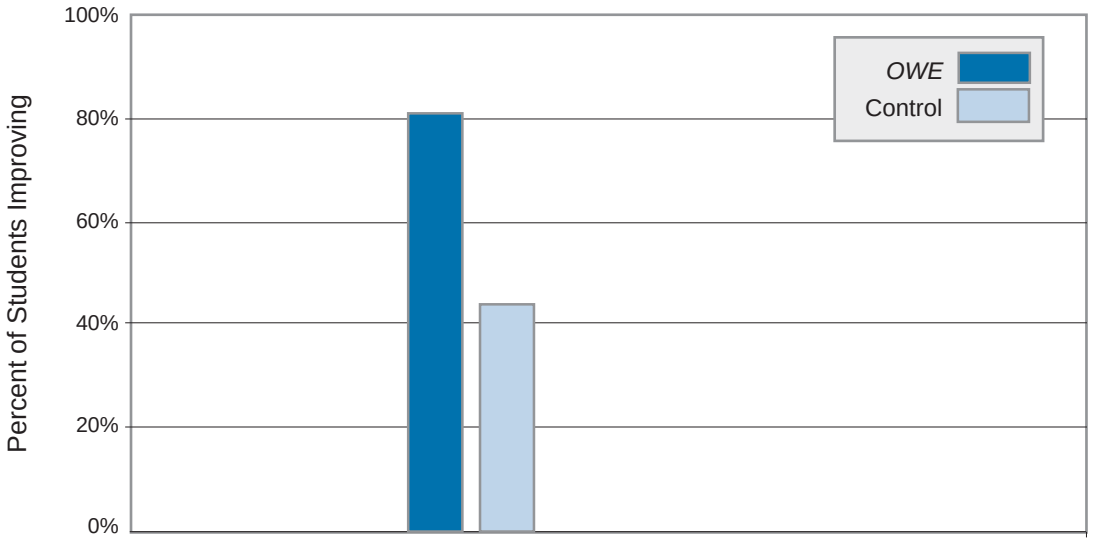



**Figure 15: Difference between *OWE* and control in percentage of students who improved in oral proficiency on the IPT Oral**

Using this dichotomous outcome, we find a substantial difference between the two conditions. In order to get a more detailed analysis of these results, we treated the ordinal scale as a continuous scale using the same modeling techniques used for the other scales. (Prior to undertaking this approach, we modeled the ordinal outcome using methods that assume a single underlying continuous latent trait. HLM was used to perform this analysis and the results yielded the same conclusions as when we treated the outcome measure as continuous. The results of the latter method are easier to explain and therefore presented here.) Table 14 shows the result of the statistical model for oral proficiency. As with the model for reading, we are interested in the interaction of treatment with the pretest, the difference here being that the analysis is restricted to the bilingual group.

**Table 14: Multi-level mixed model for IPT Oral for bilingual group—results for condition including pretest and treatment by pretest interaction**

| Descriptive statistics for IPT Oral outcomes for the bilingual group | Raw Group Means | Standard Deviation | Number of Students | Number of Teachers | |
|---|---|---|---|---|---|
| *OWE* | 4.638 | 1.405 | 47 | 4 | |
| Control | 3.305 | 1.773 | 59 | 3 | |
| **Mixed model: Fixed factors related to IPT Oral outcomes** | **Estimate of coefficient** | **Standard error** | **DF** | ***t* value** | ***p* value** |
| Intercept | 3.538 | 0.145 | 5 | 24.448 | <.0001 |
| Pretest score (centered at the mean) | 0.968 | 0.065 | 97 | 14.778 | <.0001 |
| Condition (OWE = 1; control = 0) | 0.926 | 0.218 | 97 | 4.237 | <.0001 |
| Pretest score by condition interaction | -0.264 | 0.104 | 97 | -2.549 | 0.012 |
| **Mixed model: Technical details for random components** | **Estimate of variance component** | **Standard error** | | ***z* value** | ***p* value** |
| Class mean achievement | 0.038 | 0.058 | | 0.658 | 0.255 |
| Within class variation | 0.641 | 0.092 | | 6.930 | <.0001 |

*Note: of the 117 bilingual students in the sample, 11 did not have both pre and posttest results for IPT Oral.*

The analysis shows a very low p value for condition (*OWE* or control), indicating that the difference of the size we observed is unlikely to occur by chance. Because there is an interaction between condition and prior score, it is best to interpret these results with the help of a graph.

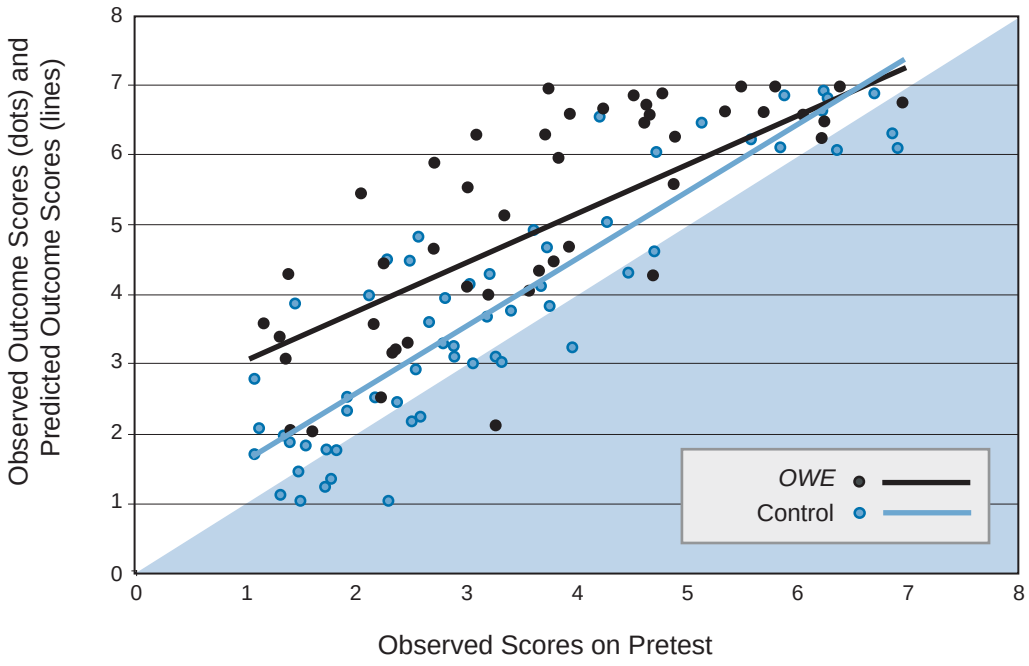



**Figure 16: IPT Oral for bilingual group—scatterplot of *OWE* and control students with lines showing the predicted values based on pretest score**

In this graph, the points representing the students are “jittered,” i.e., dispersed randomly for ease of viewing. The lines of best fit (i.e., the lines representing the predicted values) are based on the unjittered results.) The dark line again represents the bilingual students in the *OWE* group and the light line, those in the control group. While the control group shows consistent progress across the range of pre-test scores (that is, parallel to the “no growth line”), the *OWE* line shows that *OWE* was substantially more effective for the students at the early stages of learning to speak English.

Figure 17 puts confidence intervals around the difference between *OWE* and control groups. Here we see that, except for the students at the highest level where the programs were equally effective, the difference is unlikely to have occurred by chance.

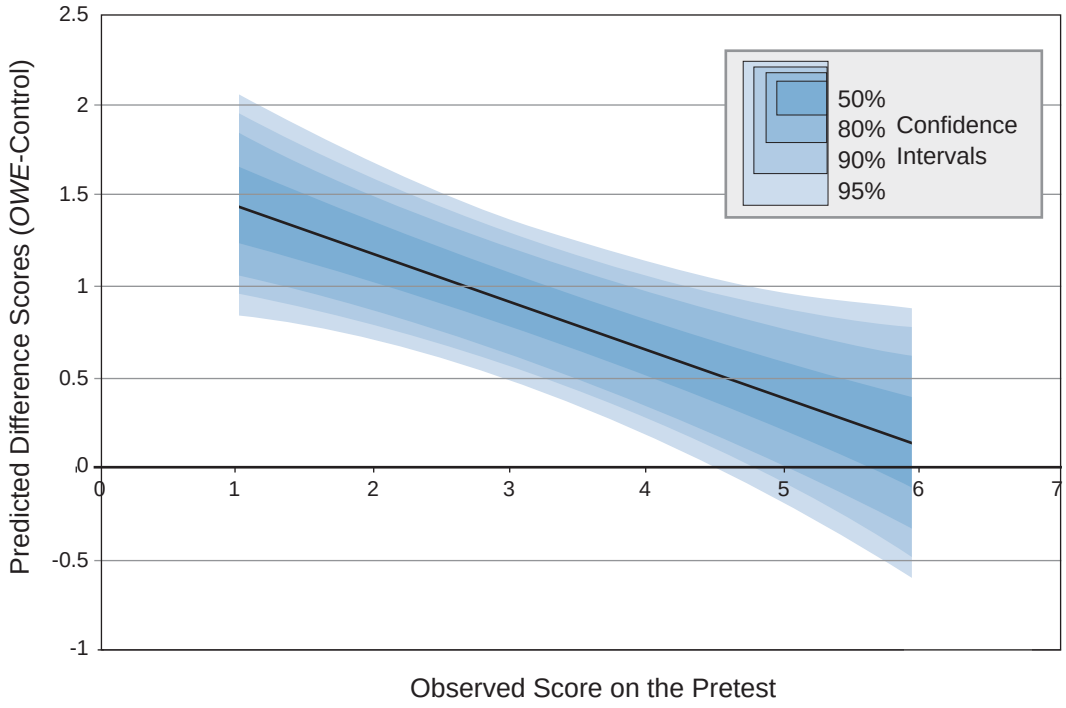



**Figure 17: IPT Oral for the bilingual group—difference between *OWE* and control showing the values for the median student at each quartile of the pretest**

## Discussion

The results in California and Texas were consistent. In both settings, we had tests of reading and of oral proficiency. *OWE* made a substantial difference for achievement of oral proficiency compared to the control conditions. In California, the results for reading and writing were less favorable. Under some conditions, the control program was substantially more effective. In many cases, there was no difference between *OWE* and the control conditions for student achievement in reading.

A finding of no difference does not mean that the product is ineffective. In all cases, it is important to interpret these results in relation to what teachers were using in the control condition classrooms and usage patterns, implementations, and applications in the *OWE* classrooms. School districts such as those participating in this study must look at the incremental benefit of a new program, given what is already in place. In both California and Texas, all classrooms had English language learner materials and

an English reading text prior to the intervention. We know that in some of the Texas English immersion classrooms, because of their students' high level of proficiency, teachers found the standard texts to be preferable. In California, the adopted basal reading text already had an English learner component, and in the control classrooms, this was supplemented by other English learner programs.

Districts considering a new program must also carefully consider the very different needs of programs serving the newest English learners and those serving students already making progress toward English proficiency. The very different pattern of success in bilingual and immersion classrooms observed here also suggests that one product is not effective overall for every situation. Even in the area of oral fluency, where *OWE* was consistently successful, it turned out to be more than was needed for many of the students in the Texas immersion classrooms.

Oral proficiency is the area where *OWE* made the greatest difference compared to the control group programs. In Texas, we did not include the English immersion classrooms in the analysis because their students were proficient or near proficient in speaking English at the outset. In California, the level of English proficiency was more evenly distributed between the bilingual and English immersion classrooms, so all the students were included in the analysis. It is striking that in California, the students in the control classrooms on average made no progress from one year to the next, whereas students in the *OWE* classrooms increased by about 16 points on the CELDT scale, which represents a considerable gain when we recognize that each of the five proficiency levels (beginning through advanced) spans an average of about 37 points. In Texas, fewer than half the students in the control classrooms made any progress from the beginning of the year to the end. In the *OWE* classrooms, more than 80% of the students made notable progress. All the students who moved three proficiency levels were in the *OWE* group.

*OWE* is shown to be an effective program in comparison to programs used by a randomly assigned control group. In reading and writing, it is generally as effective as other common programs. In developing oral proficiency, it is shown to be more effective than the programs used in the control classrooms.

We designed the experiments to provide useful information to the participating districts, not by themselves to provide widely generalizable results. Our recommendation to the participating districts is to focus the use of *On Our Way to English* in the area of oral proficiency. For reading, it is one of many potentially effective programs to consider. We found that the program was often more successful in immersion implementation rather than in bilingual classes; therefore, when making decisions, educators should carefully consider the needs of these different kinds of classes.

## References


Ballard, W. S., Dalton, E. F. & Tighe, P.L. (2001) *Examiners Manual: IDEA Oral Language Proficiency Test English*. Brea, CA: Ballard & Tighe, Publishers.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *HLM: hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs*. Chicago, IL: Scientific Software, Inc.

Renaissance Learning, Inc. (2003) *STAR Reading: Understanding Reliability and Validity*. Wisconsin Falls, WI: Renaissance Learning Inc.

SAS Institute (2003). *SAS/STAT Software: Changes and Enhancements through Release 9.1.3*, Cary, NC: Author.

Shadish, W.R, Cook, T.D. & Campbell, D.T (2002) *Experimental and Quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin Co.

US Department of Education (2003) *Identifying and implementing educational Practices Supported by Rigorous Evidence: A User Friendly Guide.* Washington DC: Institute of Education Sciences. (Available at: http://www.ed.gov/rschstat /research/pubs/rigorousevid/rigorousevid.pdf