VERSANT™
A smarter way to test speaking skills

The Use of the Versant English Test as a Measure of Score Improvement

Knowledge Technologies Group, Pearson
December, 2010

Executive Summary

The Versant™ English Test is a fully automated assessment of spoken English used by leading government agencies, academic institutions, and corporations throughout the world. Many organizations and institutions have leveraged the Versant test as a measurement tool for assessing improvement over time. This document provides empirical evidence showing that the Versant test is sensitive to gains in spoken English skills from instruction and can be used to monitor progress of spoken English ability effectively.

A total of six studies are presented. The first four offer empirical support for the test's sensitivity to English proficiency improvements as a direct result of English language instruction. The method used for all four studies was a comparison of scores from a pre-test administered at the beginning of an English program and a post-test administered at the end of the program. For all four studies, the results showed statistically significant gains in performance between the pre- and post-tests. The increases ranged from two points for a three-week immersion program to over six points for a semester-long English course. That is, on the Versant score scale of 20 (beginner) to 80 (proficient), learners on a course of study consistently show average gains of between two and six points. These statistically significant score gains indicate that the Versant English Test can detect improvements in spoken English ability from focused English language instruction and can serve as an effective measurement tool for evaluating progress over time. The results of the first four experiments are summarized in Table 1.

Table 1. Summary of Average Point Gains from Four Pre- and Post-Test Studies

	Study 1	Study 2	Study 3			Study 4
	ESL course	ELL course	Immersion A	Immersion B	Immersion C	Corporate
Gains	6.2	6.2	2.6	2.2	3.3	4.5

To address a potential concern that growing familiarity with the test during the post-test session might have caused the observed score increases (as opposed to a true improvement in English proficiency), a study was designed to measure the magnitude of a potential practice effect. In the study, participants took the Versant test three times in one session. The results showed an average increase of only a half point. Given this small (not statistically significant) effect size, it is clear that performance improvements account for the sizable increases observed in the previous four studies.

In a final study, researchers from UC Davis used the Spanish version of the Versant test (Versant Spanish Test) to measure spoken Spanish ability of students taking Spanish courses of varying proficiency levels. On average, performances from students in advanced courses were more than 20 points higher than introductory students, with a clear separation between levels. This demonstrates how the test is useful for placing students into different classes, and how the scores can be used to characterize and quantify the difference in proficiency levels between classes.

Together, the findings from the research support the use of the Versant Test as an instrument for measuring gains in spoken language ability of individuals as well as for evaluating the success of instructional programs.

Introduction

The purpose of this document is to provide empirical evidence showing that the Versant English Test is sensitive to gains in spoken English skills as a result of formal language or classes instruction. The first section consists of four studies that show significant gains between tests taken at the beginning and end of various English language instruction programs. In the second section, concerns of potential practice effects are addressed. In the final section, data from the Versant Spanish Test are presented. Taken together, the results from the research confirm that the Versant Test can effectively measure gains in spoken language performance.

Evidence for Measuring Gains

Study 1: Oakland Community College, USA

In the first study, score differences were investigated for several hundred students enrolled in an English as a Second Language (ESL) program at Oakland Community College in Detroit, Michigan. Typical courses at Oakland Community College include four hours of English instruction per week, for a total of 56 hours per semester. Administrators at the college conducted a study in which 384 participants took a pre-test at the beginning of an ESL course and then took another post-test at the end. The mean Overall scores from the pre-test and post-test are summarized in Table 2.

Table 2. Pre- and Post-test Mean Overall scores for Oakland Community College

	Pre-test Overall Mean	Post-test Overall Mean	Difference
Oakland Community College (n=384)	43.8	50.0	6.2*

*$p<0.01$

The results show that students' scores improved in the post-test administration from 43.8 to 50.0, with a gain of over six points. The score difference was found to be statistically significant ($p<0.01$). Figure 1 graphically shows the score distribution patterns for the pre-test and post-test administrations.

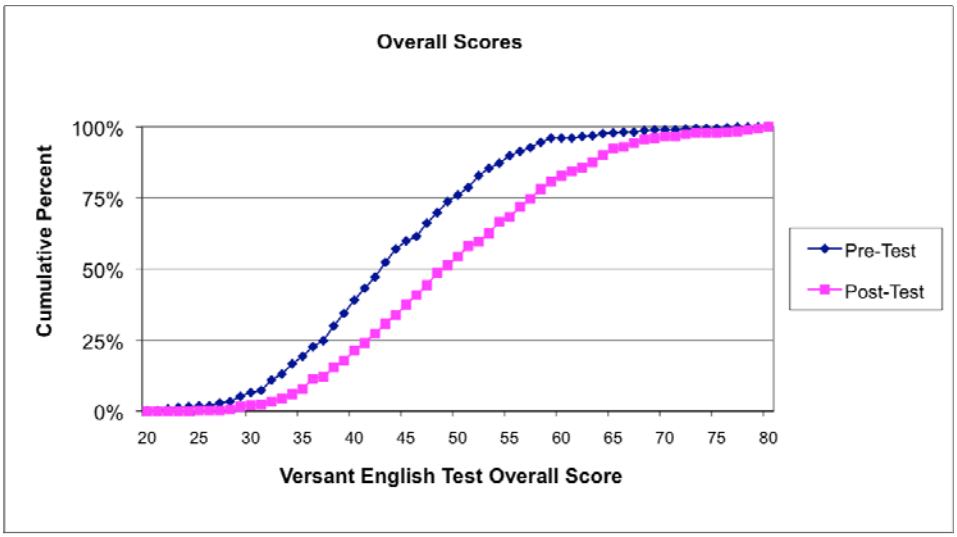

Figure 1. Cumulative Distribution Functions of Pre- and Post-test Administrations at Oakland Community College (n=384)

The statistically significant score increase and clear separation between the pre- and post-test distributions illustrates that the Versant English Test is sensitive to gains in spoken English performance.

Study 2: Waseda University, Japan

In Study 2, a similar experimental design was used but instead of evaluating performance of students in an ESL program in the US, this study investigated English proficiency improvements of students learning English as a Foreign Language (EFL) at Waseda University in Tokyo, Japan. The participants of the study were 75 first-year students enrolled in an English course that met for 1.5 hours per week, for a total of 40 hours of instruction per semester. The Versant test was administered to students at the beginning (April) and end (January) of the full instructional year in Japan.

The results from the pre-test and post-test are summarized in Table 3.

Table 3. Pre- and Post-test Mean Overall scores for Waseda University

	Pre-test Overall Mean	Post-test Overall Mean	Difference
Waseda University (n=72)	38.6	44.8	6.2

The mean Overall score increased from 38.6 at the beginning of the academic year to 44.8 at the end, for a total gain of over six points. Figure 2 shows the cumulative distribution functions for pre- and post-test scores.

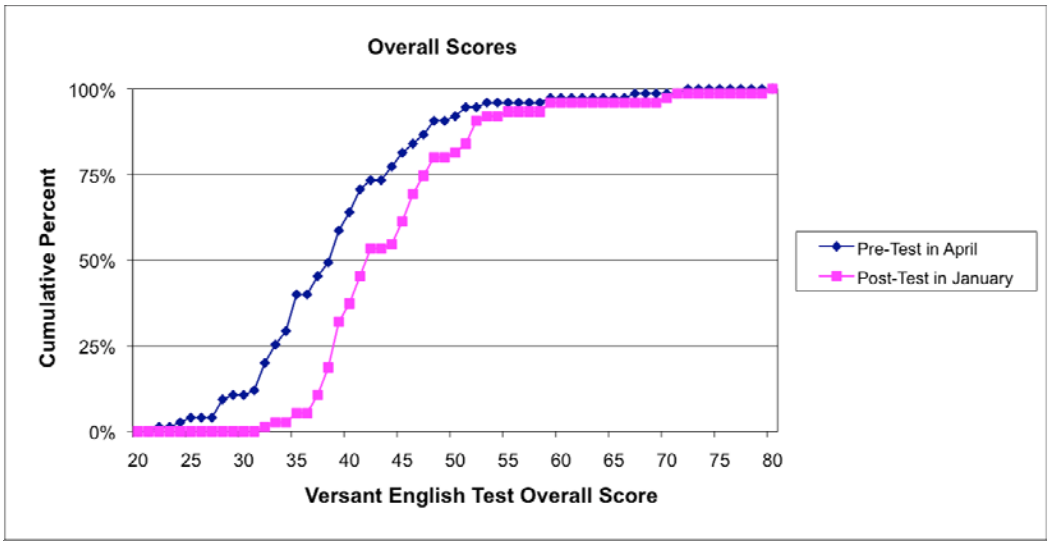

Figure 2. Cumulative Distribution Functions for students at Waseda University in Japan (n=72).

Figure 2 clearly shows the overall shift of scores to the right, which indicates improvement in spoken English ability by the end of the academic year.

The results from Study 2 with EFL students corroborate the findings from Study 1 and contribute further evidence that the Versant English Test is sensitive to gains in performance due to instruction.

Study 3: Performance Gains from Short Term Study Abroad Programs

Study 3 consisted of three substudies in which students learned English abroad in an English speaking country for three weeks. Two different universities in Japan used the Versant English Test as an instrument to assess students' improvement in spoken English as a result of the immersion program. A total of 76 students participated in the three substudies:

- Substudy A – 30 Japanese university students attended a program in Australia for three weeks
- Substudy B – 19 Japanese university students attended a program in Australia for three weeks
- Substudy C – 27 Japanese university students attended a program in Canada for three weeks

Students took the Versant English Test before and after their respective study abroad program. As summarized in Table 4, all three substudies showed increases in scores with improvements ranging from two to three points.

Table 4. Pre- and Post-test Mean Overall Scores from Three Study Abroad Programs

	N	Pre-test Mean Overall	Post-test Mean Overall	Difference
Substudy A (Australia)	30	36.3	38.9	2.6**
Substudy B (Australia)	19	36.2	38.4	2.2*
Substudy C (Canada)	27	37.3	40.6	3.3**

*p<0.05, **p<0.01

Compared to Study 1 and Study 2, the gains in performance were smaller because of the short amount of time between the pre- and post-test. The pre-post score differences were analyzed statistically and all were found to be statistically significant.

Study 3 demonstrated how sensitive the Versant English Test is to small changes in performance. Many other assessment instruments such as oral proficiency interviews, which have scales with only five or six proficiency levels, most likely would not have been able to detect a significant change in performance after such a short amount of time.

Study 4: English Training Program with Two Corporate Clients

In Study 4, a language training company selected Versant English Test as an assessment instrument to determine whether or not call center employees showed measurable gains in facility with spoken English as a result of an English training program. In this study, participants were assigned to a pilot group that took an English language training program or a control group that did not. Participants were recruited from three different companies in two separate locations in India. The pilot group consisted of 64 participants and the control group consisted of 74, for a total of 138 participants. Each participant took the test twice: once at the beginning of the experimental period, which lasted approximately three weeks, and once at the end of the period.

Table 5 presents the means before and after training and difference scores for both the pilot and control groups.

Table 5. Pre- and Post-test Mean Overall Scores After a Language Training Program

Group	Pre-Training Overall Mean	Post-Training Overall Mean	Difference
Pilot (n=64)	54.4	58.9	4.5**
Control (n=74)	57.2	58.6	1.4*

*p<0.05, **p<0.01

Improvement in the mean Overall score of the pilot group was 4.5 points. In contrast, the control group improved only by 1.4 points. Participants were expected to show some gains in performance given their daily use of English to perform their jobs at the call center. The data indicate that the English language training program resulted in a larger gain in performance.

Figure 3 presents cumulative distribution functions of the two groups.

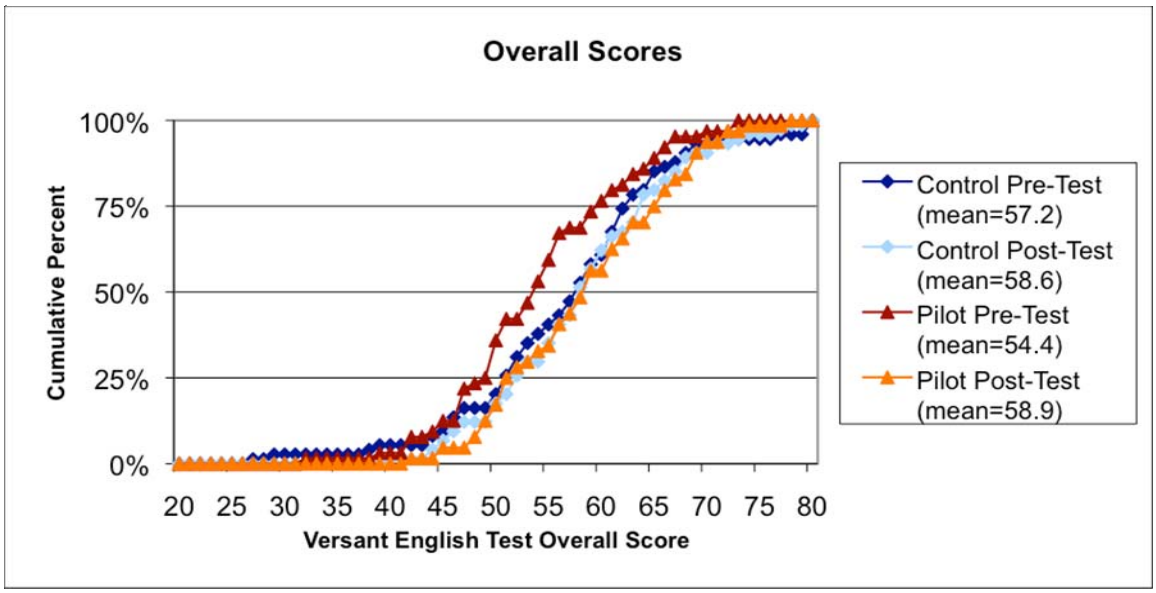

Figure 3. Cumulative Distribution Functions of Pre- and Post-tests for Control and Pilot Groups

Figure 3 shows a clear separation between pre- and post-test scores for the pilot group, which suggests that the language training program had a positive effect on the participants' facility with spoken English. The pre-test for the control group was about three points higher than the pilot group. This difference was simply an artifact of the sampling, such that participants in the control group were slightly more proficient on average at the beginning of the training period compared to the pilot group. In contrast to the pattern observed with the pilot group data, the post-test distribution for the control group did not shift to a higher score range. Instead, the post-test distribution remained similar to the pre-test distribution, which suggests that participants in this group did not exhibit consistent score increases. The fact that a clear difference can be observed between the patterns of the pilot and control groups lends further support that the Versant English Test is capable of monitoring performance gains due to language instruction.

Magnitude of Potential Practice Effects

Study 5: Repeated Measures

One concern regarding gains in test scores over time is the possibility that the improvements are due to the test-takers' increased familiarity with the test itself, commonly known as a "practice effect". To address this issue, Pearson conducted a repeated measures study to determine how much of a gain could be attributed to a potential practice effect.

A total of 140 English learners with a mean age of 32 years (sd = 8.75) participated in the study. Participants represented a wide range of native language backgrounds including Amharic, Arabic, Cambodian, Cantonese, Farsi, Indonesian, Italian, Japanese, Korean, Mandarin, Romanian, Russian, Somali, Spanish, Tagalog, Taiwanese, Turkish, and Vietnamese.

Experimenters administered three randomly-generated test forms to each participant in a single session (Test 1, Test 2, and Test 3). Comparisons between Test 1 and Test 2 represented test-retest reliability in the absence of a practice test, while comparisons between Test 2 and Test 3 represented test-retest reliability in the presence of a practice test (i.e., Test 1).

Descriptive results of the scores for each test are summarized in Table 6.

Table 6. Mean overall Versant English Test scores, standard deviations and score differences.

	Administration Order		
	1st	2nd	3rd
Mean	44.46	44.99	44.72
Standard Deviation	15.30	14.25	15.17
Score Difference (between 1st and 2nd)	0.53		
Score Difference (between 2nd and 3rd)		-0.27	

The results showed an average increase of only a half point between the first and second tests. Between the second and third administrations of the test, a slight decrease was observed.

To analyze whether or not there were statistically significant differences between scores, a single-factor Analysis of Variance (ANOVA) was performed. None of the score differences were found to be statistically significant.

Overall, the data showed no appreciable differences between mean Overall scores. These results emphasize the small effect size of a potential practice effect and support the conclusion that improved spoken language ability was the reason scores increased between pre- and post-tests in the previous studies.

Versant Spanish Test

Study 6: Versant Spanish Test at University of California, Davis

In Study 6, researchers at UC Davis used the Spanish version of the Versant Test (Versant Spanish Test) to discriminate the spoken Spanish ability of students enrolled in different levels of Spanish courses. The Versant Spanish Test follows the same set of principles as Versant English Test with regard to test development, test structure and scoring logic.

Study 6 was different from the previous studies in this document because it was designed as a cross-sectional study as opposed to a longitudinal study. Participants were 233 university students enrolled at UC Davis who were taking one of seven different Spanish courses. The courses are referred to as SP1, SP2, etc. with higher numbers denoting more advanced levels. In addition, a group of 15 heritage Spanish speakers, who spoke Spanish at home but who were not previously formally educated in Spanish, took the test as well. Figure 4 presents the cumulative distribution functions for students in each course and for heritage speakers.

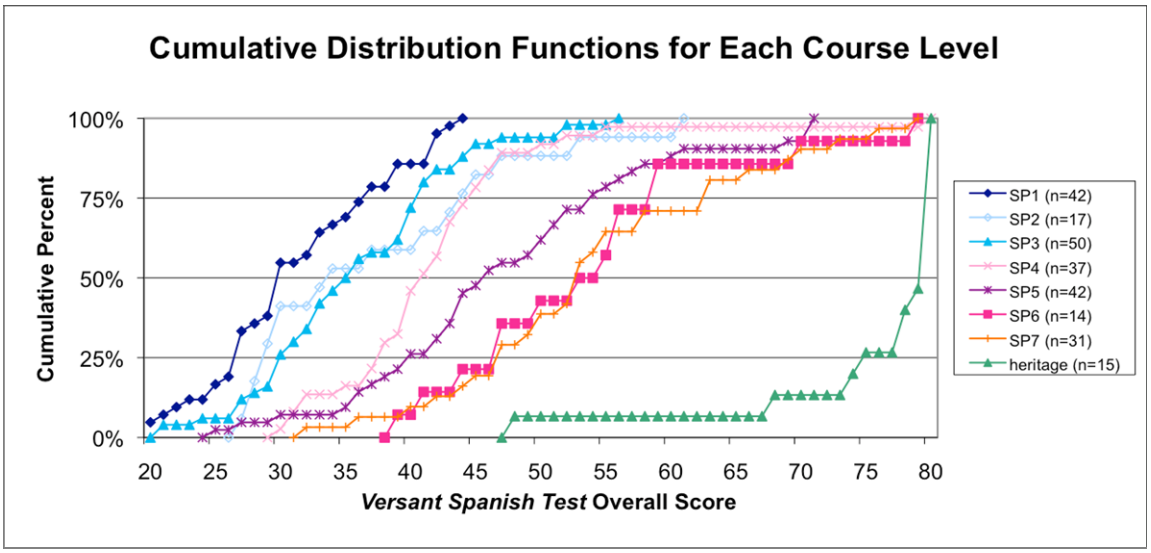

Figure 4. Cumulative Distribution Functions for Each Course Level at UC, Davis

In Figure 4, the distributions for students in lower level courses gravitated more toward the lower end of the scale, while distributions for students in higher level courses appeared in the middle of the scale. Performances from students in advanced courses were more than 20 points higher than introductory students, with clear separation between levels. The heritage speaker group showed a markedly different pattern from "typical" non-native learners. This was expected since heritage speakers tend to have strong spoken language skills from speaking the language at home with their parents.

The separation between the distributions indicates that the Versant Spanish Test was able to discriminate students with different spoken Spanish abilities. The findings suggest that the test can detect overall improvements as students progress through the offered courses.

Summary

In this report, four studies were presented that show statistically significant gains in spoken English ability as a direct result of English language instruction. For short, three-week programs, modest gains of two or three points were observed. For semester and year-long courses, the results showed larger increases of over six points on average. These studies provide evidence that the Versant English Test is sensitive to gains in performance from instruction.

A follow-up study on practice effects revealed that the significant score increases were not caused from growing familiarity with the test, but rather true improvements in English language proficiency.

A final study with the Versant Spanish Test, which is based on the same technology and scoring methods, confirmed the test's ability to discriminate students with different levels of spoken Spanish proficiency.

Together, the findings from the research presented in this document support the use of the Versant English Test as an instrument for measuring gains in spoken language ability of individuals as well as for evaluating the effectiveness of instructional programs.