Consistency of Versant™ test scores over multiple administrations

This document describes a repeated-measures study conducted by the Versant Test Development team to examine the possibility of practice effects on the Versant English Test. The results demonstrate that overall performance is consistent across test occasions, regardless of proficiency level, and regardless of whether the candidate has first practiced for the test. This pattern of results is identical for both computer- and telephone-administered Versant English Test.

Design Overview

To determine whether the presence of a practice test influences performance on a subsequent pre- and post-test, a repeated-measures experiment was conducted. Three randomly generated test forms were administered in a single session to each participant. Test administrations are referred to as 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). In this way, each participant served as his or her own control for receiving a practice test. This design has the substantial strength of reducing variability between subjects and between testing sessions.

Finally, participants were assigned to receive the Versant English Test as a Computer-Delivered Test ("CDT") or telephone-delivered test ("Phone"), permitting an examination of the possible influence of administration modality on test-retest reliability.

Participants

A total of 140 participants with a mean age of 32 years (sd = 8.75) were drawn from five distinct populations of adult learners of English (see Table 1 for details). Participants came from 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. All participants provided informed consent and were compensated for their participation. The size and diversity of this sample were selected to ensure generalizability of findings.

Table 1. Demographic information on participants from each site.

Testing Site	n	Mean age in years *(sd)*	Number of Male:Female
College of San Mateo (CSM) *Northern California*	61	30.66 *(7.95)*	23:38
San Diego State University (SDSU) *Southern California*	18	25.56 *(4.97)*	5:13
San Jose Community College (SJCC) *Northern California*	30	37.53 *(10.13)*	4:26
South Piedmont Community College (SPCC) *North Carolina*	26	33.62 *(7.75)*	10:16
Community of Menlo Park (MP) *Northern California*	5	29.60 *(4.39)*	2:3
Total	140	31.99 *(8.75)*	44:96

n = number of participants; *sd* = standard deviation

Materials and Procedure

Participants were randomly assigned to a test Administration Modality: either CDT or Phone. Computer-administered tests were presented to participants through a microphone-equipped headset, while telephone-based tests were administered over a land line telephone. Each participant was administered three randomly generated Versant English Tests, one after another (back to back), all in the same modality (i.e., CDT or phone).

Analysis and Results

Test-retest reliability was estimated using Pearson's correlation coefficient applied to overall Versant English Test scores at the three different administrations. Results of correlation analyses are summarized in Table 2.[1]

[1] These correlations are remarkably high – considerably higher than is commonly found with human rating. Even so, they are consistent with correlations reported in previous Ordinate studies. Machine scoring of speech samples is unaffected by many of the biases implicit in human rating, permitting scores to be generated based on objective analysis of the features of interest only, without influence from extraneous, context-specific, or construct-irrelevant features. For that reason, Versant's machine-scored test-retest reliability values are consistently well above those seen in human-administered, human-scored assessments.

Table 2. Pearson's r-values for correlations between overall Versant English Test scores.

Condition	Administration Modality		
	CDT ONLY n = 64	PHONE ONLY n = 76	CDT & PHONE n = 140
Without a practice test (Test 1 vs. Test 2)	0.94	0.95	0.97
With a practice test (Test 2 vs. Test 3)	0.95	0.96	0.97
Repetition effects (Test 1 vs. Test 3)	0.94	0.96	0.97

Next, to determine whether there were any differences between scores on any of the 3 administrations, a separate single-factor Analysis of Variance (ANOVA) was performed for each group of participants (i.e., those who took the test in the CDT (n = 64) and Phone (n = 76) modalities), with Administration Order (Test 1, 2, or 3) as a factor. There were no statistically significant differences between administration order for either CDT [F(2, 189) = 0.12, n.s.] or Phone [F(2, 225) = 0.02, n.s.] administration groups. Descriptive results of the scores across administration modality and order are summarized in Table 3.

Table 3. Mean overall Versant English Test scores (on a scale from 20 to 80) and Standard Deviations across Administration Order (Tests 1, 2, and 3).

Administration Modality	Mean Score *(sd)*		
	Administration Order		
	1st	2nd	3rd
CDT only (n = 64)	44.80 *(12.53)*	45.78 *(12.82)*	44.88 *(12.64)*
Phone only (n = 76)	44.17 *(15.30)*	44.33 *(14.25)*	44.59 *(15.17)*
All together (n = 140)	44.46 *(15.30)*	44.99 *(14.25)*	44.72 *(15.17)*

n = number of participants; *sd* = *standard deviation*

The data show no appreciable differences between mean overall scores across administrations.

"Without a practice test" (Test 1 vs. Test 2)

Test-retest reliability for Versant English Test scores in the combined administration modality condition were correlated, with r = 0.95. Looking at computer-based and phone-based administration separately, the test-retest reliability was similarly high, with r = 0.94 and r = 0.97, respectively.

Figure 1 (top left) shows that Versant English Test scores on Test 1 and Test 2 demonstrate a strong, linear, positive relationship regardless of administration modality. This provides compelling evidence that test-retest reliability using random forms of the Versant English Test is high even when no practice test is administered.

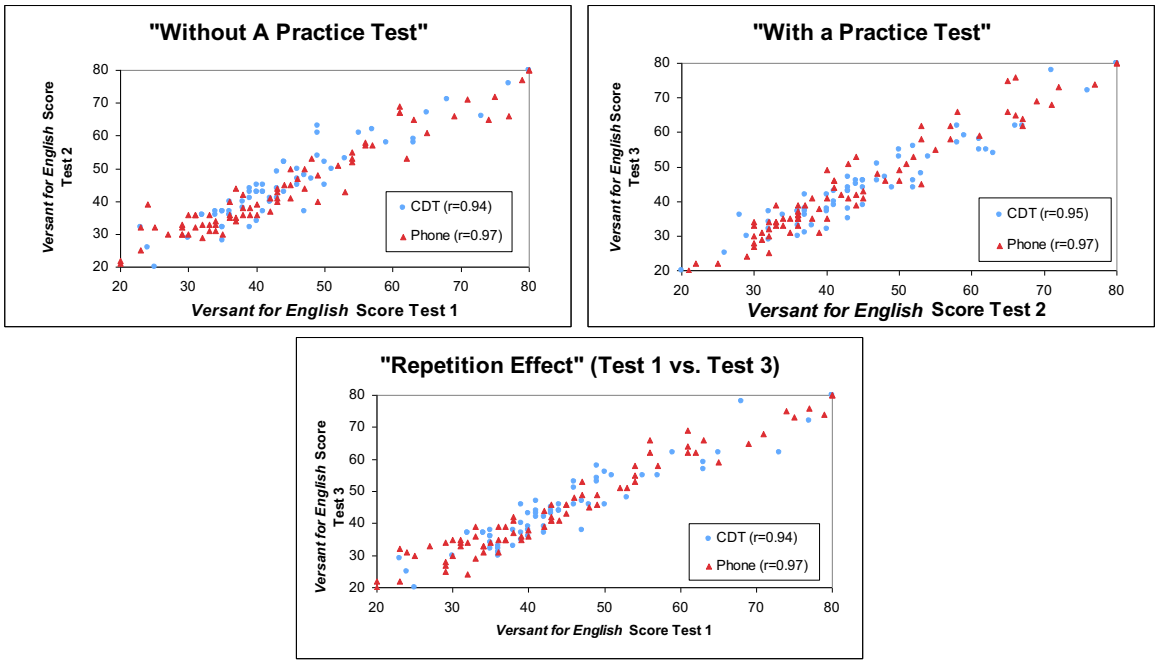

Figure 1. Comparison of Versant English Test overall scores obtained from two administrations, with CDT scores represented as blue circles and phone scores represented as red triangles. "Without a practice test" (top left) represents the comparison between Test 1 and Test 2; "With a practice test" (top right) represents the comparison between Test 2 and Test 3; "Repetition Effects" (bottom left) represents the comparison between Test 1 and Test 3.

When collapsed across both computer and phone administration modalities, mean test scores increased 0.53 Versant English Test points between Test 1 and Test 2 administrations, from 44.46 to 44.99, respectively. This pattern maintained for computer administered tests, with a mean score increase of 0.98 points (the standard error of the test is 2.9 points), from 44.80 to 45.78; and for phone administered tests, with scores increasing an average of 0.16 points, from 44.17 to 44.33. Paired t-tests revealed that none of these differences were significant (all ps > 0.05).

"With a practice test" (Test 2 vs. Test 3)

Overall test-retest reliability was high, with an r = 0.96. Computer-based and phone-based administrations showed reliability of r = 0.95 and r = 0.97, respectively. Thus, Versant English Test scores on Test 2 and Test 3 exhibit a strong, positive, linear relationship regardless of administration modality (Figure 1, top right).

When collapsed across both computer and phone administration modalities, mean test scores decreased 0.27 points between Test 2 and Test 3, from 44.99 to 44.72, respectively. Mean scores from the computer modality dropped from 45.78 to 44.88 (a difference of -0.91) between Test 2 and Test 3, respectively, while scores obtained from the phone administration increased 0.26 points between Test 2 and Test 3, from 44.33 to 44.59, respectively. No changes in score were statistically significant (all *p*s > 0.05).

These results support the claim that test-retest reliability is high when the Versant English Test is taken following a practice test.

"Repetition Effects"

Correlations between Tests 1 and 3 were strong, with the combined administration modalities yielding r = 0.96 (Figure 1, bottom left). Correlations of scores collected by CDT (r = 0.94) and phone (r = 0.97) were similarly strong.

To test the possibility that scores naturally improve with increased experience with the task due to so-called "repetition effects" (also referred to as "practice effects"), scores from Test 1 – in which participants were presumably naïve to the test's demands – were compared with scores from Test 3 – at which point participants had acquired experience and familiarity with the task. The scores from Versant English Tests 1 and 3 showed an overall mean increase of 0.26 points (from 44.46 to 44.72, respectively) when all participants were collapsed together. For CDT test-takers, this increase was 0.08 points (44.80 to 44.88), while for phone test-takers, the difference amounted to an increase of 0.42 points (44.17 to 44.59). None of these increases was statistically significant (all *p*s > 0.05).

This demonstrates that performance on a third administration is not significantly different than the performance on the initial administration, suggesting that repetition effects due to increasing familiarity with the task do not result in any consistent change in Versant English Test overall scores.

Conclusion

The data reveal that test-takers score comparably when taking Versant English Test on multiple test occasions. When present, mean score differences were typically on the order of <1 point (on Versant's 20 to 80 point scale) from test to test. Such a deflection is well within the standard error of measurement (2.9 points) of the Versant English Test. Further, there is no evidence that taking the Versant English Test following a practice test has any systematic effect on scoring. Thus, the test exhibits very high parallel-form reliability. Finally, this consistent test-retest pattern was identical for computer-based and phone-based delivery of the test, indicating no effect of presentation modality on test-retest reliability.