Comparison of PhonePass™ Testing

with the

Educational Testing Service® Test of Spoken English™ (TSE®)

Brent Townshend and Ognjen Todic

Abstract

This study compares the test scores obtained by students on the Test of Spoken English (TSE) offered by Educational Testing Services (ETS), with an automated test based on PhonePass technology from ORDINATE Corporation. The measured correlation between the two tests over the observed population is 0.88.

PhonePass Testing Overview

The PhonePass system delivers automated tests that measure speaking and listening skills during a 10-minute interaction over the telephone. PhonePass tests are scored automatically by a computer-based system. Each test item requires the candidate to understand a spoken utterance and speak in response to it. Scores are based on the exact words used in spoken responses, as well as the pace, fluency, and pronunciation of those words in phrases and sentences. PhonePass tests measure facility in spoken English, which includes the ease and immediacy in understanding and producing basic conversational English. PhonePass tests measure core skills that enable a person to understand spoken language about everyday topics and respond intelligibly at a native conversational pace.

Studies of the PhonePass test [1] indicate an overall test score reliability of 0.93.

Test of Spoken English

The primary purpose of the Test of Spoken English (TSE) is to measure the ability of nonnative speakers of English to communicate orally in English. TSE is delivered in semidirect format, which maintains reliability and validity while controlling for the subjective variables associated with direct interviewing. The TSE score is a reflection of an examinee's oral communicative language ability on a scale from 20 to 60 (from "No effective communication" to "Communication almost always effective"). Raters evaluate speech samples and assign score levels using descriptors of communicative effectiveness related to task/function, coherence and use of cohesive devices, appropriateness of response to audience/situation, and linguistic accuracy. There are 12 items on the test, and each item receives an individual holistic score. The 12 scores are averaged across raters and reported in five-point increments (i.e. 20, 25, 30, 35, 40, 45, 50, 55, 60) [2].

ETS reports a test score reliability ranging from 0.87 to 0.92 and a standard error of measurement of 1.90 to 2.14 TSE scale points over several administrations of the TSE.
Method

The subjects of this test are adults located in Albany, NY who agreed to take both the TSE and the US-Talk implementation of PhonePass technology. The 58 subjects were chosen to cover a wide range of ability in spoken English. The native languages of the subjects were Russian (44%), Chinese (30%), and Spanish (26%). The gender of the sample was slightly skewed with 64% female subjects. Each subject took the standard ETS TSE within a few weeks of taking the US-Talk implementation of PhonePass testing.

Figure 1. Distribution of TSE scores, Current Sample, $\mu$=39.1, $\sigma$=10.7, N=58

Figure 2. Distribution of PhonePass scores, Current Sample, $\mu$=5.1, $\sigma$=1.6, N=58
Results

Figure 3. TSE vs. PhonePass, r=0.88, N=58

Figure 3 is a scatter plot of the TSE and the PhonePass test scores for the 58 subjects based on raw data. The measured correlation between the two tests over this population is 0.88.

The reported reliability of the TSE is approximately 0.89. However this reliability is for an examinee pool with a standard deviation of 6.2. The examinee pool in this study covers a wider range of TSE scores, having a standard deviation of 10.7. Although the reported reliability does not apply to the current sample, we can form a rough estimate of this reliability by applying a correction based on the observed and reported standard deviations of the examinee pools. This gives an estimated reliability for the TSE of 0.96 over the current sample.

Using the mean correlation of 0.88 we can then compute a corrected validity coefficient [3] using:

\[
    r_{xy(CV)} = \frac{r_{xy}}{\sqrt{r_{x(y)}}}
\]

where test \( x \) is PhonePass, criterion test \( y \) is TSE, \( r_{xy} \) is the observed validity coefficient (i.e. the correlation between the tests), \( r_{xy(CV)} \) is the validity coefficient corrected for
unreliability of the criterion measure (TSE), and $r_{xy_2}$ is reliability of the criterion measure, TSE. Applying Equation [1] to this data gives a corrected validity coefficient of 0.90. The interpretation of this coefficient is that one would expect the PhonePass score to correlate with the underlying true score, which TSE measures at a level of 0.90, over the population represented by the current sample.

In a separate analysis of results from 58 subjects, least square estimation was used to find an optimum linear mapping from PhonePass to TSE scores. The resulting fit is shown in Figure 3 as the dashed line. Using this mapping, the TSE scores predicted from PhonePass scores resulted in a mean square error equal to 5.1. This indicates that PhonePass can correctly predict a TSE score within one score step in most cases. In the current sample, PhonePass testing predicts 50/58 (86%) of the subjects' TSE scores within one score step.

Discussion

PhonePass testing demonstrates a significant level of concurrent validity with respect to the TSE exam. A raw correlation of 0.88 was measured between the TSE and PhonePass tests over a sample of subjects relatively evenly spread over the TSE score scale. A corrected validity coefficient of 0.90 was found with respect to TSE as a criterion measure. These results, coupled with the measured reliability of PhonePass, indicate that PhonePass testing can produce scores that measure a similar underlying construct to that of TSE. Furthermore, PhonePass scores can be used to infer TSE scores for the same subject with a mean square error of 5.1 TSE scale points. Since the TSE scale is quantized in 5 point steps, this indicates that a PhonePass score can predict a subject's TSE score within one score step in most cases.

However, it is important to note that the sample of examinees used for this study had TSE scores that were more widely spread and less concentrated at central scores than a typical sample of TSE examinees. The mean and standard deviation of the TSE scores shown in Figure 1 for the 58 subjects was 39.1 and 10.7 respectively. Test scores for the 1331 subjects that took the TSE test in October, 1998 have a mean and standard deviation of 45.1 and 6.3 with the overall distribution shown in Figure 4. It seems from this distribution that TSE candidates are self-selecting; lower ability candidates may elect not to take the test if their chances of success are slim. In studying the relation between PhonePass testing and TSE, it is unclear exactly what the “correct” candidate score distribution should be. As a screening tool for use prior to TSE testing, we would expect to see more candidates at lower abilities than shown in Figure 4.
Although the sample used in this study included native speakers of Russian, Spanish, and Chinese, it should be noted that the sample size is small and the distribution of first languages and gender in this sample does not match that of the total TSE test-taker population. For these reasons, the reliability of the TSE exam over this sample may be different than the average values used here. We would expect the reliability of the TSE over this particular population to be higher than the average reliabilities used in our calculations, due to the wider range of scores. Thus, the corrected reliability coefficient presented here is probably over-estimated.

It would be desirable to perform a similar study over a large, well-constructed sample of examinees for which the TSE individual rater scores are available. With this additional information, the reliabilities and concurrent validities of the two tests can be more accurately compared.

References