Test Comparisons

A comparison of three internationally recognized tests of spoken English

This document describes a study that compares test scores on three speaking tests:

- Versant English test
- TOEFL iBT (Speaking section)
- IELTS Interview tests

The study was conducted by a renowned researcher in language testing over a 6-month period in 2007. The research was financially supported by the Versant Test Development team but otherwise the methodology, data collection and analysis were entirely independent.

The data collection took place at a test centre in Iran. In previous studies the Versant Test Development team has collected data on speech samples of different nationalities and found no regional bias in test scores or accuracy of the voice recognizer. A similar study involving a mix of nationalities is desirable; however, we have every reason to believe that these results are generalizable outside of this regional population of test-takers.

Method

The participants were 130 male and female EFL learners from Iran. They represented a diverse sample of the population and had a variety of mother tongues and ethnic backgrounds, different levels of English proficiency, and came from various educational disciplines. All were genuine TOEFL iBT test-takers, who had applied to sit the test for their own educational or professional reasons. On applying to take the TOEFL iBT, the test centre administrators invited them to participate in additional Versant and IELTS speaking tests. They were financially compensated for this, but many participated for the extra practice and to find out what their scores would be.

Arrangements were made for participants to take the three tests in a counterbalanced form to remove any potential order of test effect as much as possible. However, in practice, in most cases the participants took iBT TOEFL first and then were scheduled for the other two tests on a different day. All tests were proctored and conducted under examination conditions.

Data

The following data was collected from the study participants across the three instruments used.

1. TOEFL iBT

This is intended to assess learners' English proficiency in an academic context. It is a 4-hour test and assesses the four skills of Reading, Listening, Speaking, and Writing (each reported on a scale 0-30). The Speaking section is based on the candidates' response to six questions. Responses are evaluated by human raters and receive a score of 0 (Low), 1, 2, 3, or 4 (High). The total score for the 6 questions is converted to a scale from 0 to 30.

The inter-correlations between the four subskills are shown in Table 1. Note that the iBT Speaking score correlates with the iBT Overall score at .71.

Table 1. Subscore Inter-correlation on the TOEFL iBT

TOEFL iBT	Reading	Listening	Speaking	Writing	Overall
Reading	1				
Listening	.67	1			
Speaking	.41	.59	1		
Writing	.62	.70	.59	1	
Overall	.85	.91	.71	.85	1

2. Versant English Test

This is an automated test that is designed to measure candidates' facility in spoken language. It yields an Overall score and four subscores: Sentence Mastery, Vocabulary, Fluency and Pronunciation. The inter-correlations are shown in Table 2.

Table 2. Subscore Inter-correlation on the Versant English Test

Versant English Test	Sentence Mastery	Vocabulary	Fluency	Pronunciation	Overall
Sentence Mastery	1				
Vocabulary	.72	1			
Fluency	.59	.55	1		
Pronunciation	.60	.62	.92	1	
Overall	.85	.81	.89	.90	1

3. IELTS Interview test

The IELTS speaking test is an approximately 15-minute interview consisting of several tasks: an interview, a long speaking turn, and a discussion. There are two examiners – one acts as the interlocutor and rater, and the other listens and rates only. This format is supposed to allow the examiners to probe the candidates' ability and to carefully score the resulting performance. There are four subscores: Fluency and coherence, Lexical resources, Grammatical range and accuracy, and Pronunciation.

Each candidate took two IELTS interviews, with two different examiners on each occasion (i.e. four examiners in total). These were not official IELTS tests, but rather simulations. Nevertheless, genuine IELTS past papers were used, the examiners were specially trained for their role, and all four examiners had PhDs in Applied Linguistics. Thus we can be confident that these tests approximated official IELTS Interviews in all but name.

Table 3 shows the inter-correlation of subscores on the IELTS Interview, as given by one of the four raters (it is much the same for the other three raters). The correlations are high – this possibly shows 'transfer of judgment', i.e. the raters were unable to distinguish among the distinct subskills and were influenced by the candidates' overall ability when assigning subscores.

Table 3. Subscore Inter-correlation on the IELTS Interviews

IELTS Interviews	Fluency	Lexis	Grammar	Pronunciation	Overall
Fluency	1				
Lexis	.90	1			
Grammar	.87	.90	1		
Pronunciation	.84	.84	.86	1	
Overall	.95	.94	.93	.91	1

Comparison of test scores

The inter-correlations between the four sets of test scores (i.e., the iBT Speaking, the Versant, the first IELTS Interview, and the second IELTS Interview) are given in Table 4. The highest coefficient in the table is, not surprisingly, the IELTS Interviews 1 and 2, i.e. the IELTS with itself. Nevertheless, note that this correlation is .78, and that the Versant correlates with IELTS 1 and 2 at .73 and .77 respectively. In other words, the Versant correlates with the IELTS almost as well as two IELTS tests themselves.

Looking at the relation between the Versant and iBT Speaking, we can see that the correlation is .75 (similar to the Versant and IELTS correlation). Interestingly, the IELTS and iBT Speaking do not predict each other as strongly: the correlations are .67 and .68.

Table 4. Inter-correlations on four sets of test scores

Test Correlations	Versant	TOEFL iBT Speaking	IELTS Interview 1	IELTS Interview 2
Versant	1			
TOEFL iBT Speaking	.75	1		
IELTS Interview 1	.73	.67	1	
IELTS Interview 2	.77	.68	.78	1

Summary

The data show that of the three speaking tests, the Versant correlates most highly with both the TOEFL iBT and with the IELTS. In other words, it has the most mutual variance with the other two test scores, and although the construct "speaking in English" can be elicited and measured in different ways, the Versant would appear to be the most overarching predictor of this ability. Furthermore, due to error variance in scoring the IELTS Interviews, the Versant is just as good a predictor of one IELTS test as another IELTS is.

Reference

This document was compiled from an internal report submitted by Hossein Farhady and presented at AAAL, 2008. Farhady, H. (2008). Human Operated, Machine Mediated, and Automated Tests of Spoken English. Paper presented at the American Association of Applied Linguistics, Washington DC, March 2008.