Using The TrueNorth Speaking Test To Predict ACTFL Ratings

Executive Summary

Emmersion Learning has developed an automated test to rate the speaking proficiency of learners of English, named the TrueNorth Speaking Test. This research study, conducted in partnership with an independent organization, investigated the validity and reliability of the test by comparing and analyzing the results of 108 English Language Learners who took both this test and the Oral Proficiency Interview by Computer (OPIc).

The goal of the research was to investigate TrueNorth's ability to accurately predict the scores on the OPIc utilizing the American Council for Teaching Foreign Language (ACTFL) speaking proficiency scale. A secondary research goal also included a comparison of human scoring and automated scoring utilizing speech recognition provided by Carnegie Speech for the TrueNorth test. Participants, chosen from an English Language center in Korea, were all genuine OPIc test-takers, who had signed up to take the test for their own educational or professional purposes. The tests were taken on the same day.

Main results:

- The TrueNorth Speaking Test showed a high ability to predict ACTFL ratings from the OPIc scores for participants. Scores on the test were used to accurately predict the test-takers' ACTFL rating range for 90% of participants, and were within one sublevel on the ACTFL scale for 99% of the participants.
- The automated scores performed very well in comparison to the human-rated scores for the TrueNorth Speaking Test. The comparison of human-rated scores and automatic scores utilizing Carnegie Speech API resulted in an inter-rater reliability correlation of 0.89. There was no significant difference between the ability of the TrueNorth Speaking Test to predict OPIc ACTFL levels when utilizing human scoring and automated scoring.

Method

The 108 participants were male and female English Language Learners in Korea who had previously signed up to take the OPIc and chose to take the TrueNorth Speaking Test as part of this study. The participants took both tests in a proctored testing center in Korea. For the first research objective, the scores on the TrueNorth Speaking Test were independently scored and predictions of the ACTFL rating range were derived for each student. Then the predicted score ranges were compared to the participants' ACTFL ratings from the OPIc to find the number of matching instances. The TrueNorth Speaking Test uses innovative and technology-enabled methods to measure the construct of "speaking ability" which differ from the traditional method of most interview-based speaking tests, including the OPIc. The analysis of the ability to predict the ACTFL rating provides insights into the construct validity of the TrueNorth test.

For the secondary research objective, each item from the TrueNorth Speaking Test was rated twice: once through the speech recognition API provided by Carnegie Speech and once by a

human rater. Individual item scores for the two methods were compared using Pearson correlation as an indicator of inter-rater reliability. Additionally, ACTFL predictions were made using both methods and compared with the OPIc ACTFL ratings independently.

Results

Predicting OPIc ACTFL Ratings

The ACTFL scale for proficiency ratings utilizes an ordinal scale, while the TrueNorth Speaking Test raw score utilizes an interval scale. The TrueNorth Speaking Test used a polynomial regression algorithm based on the Rasch IRT model to take the raw performance score and use it to predict the latent trait of speaking proficiency. This made it possible to estimate an ordinal proficiency rating based on the interval-scale performance of the test. The result of the algorithm was a numerical predictor that was positioned between two levels on the ACTFL rating scale. In order to conduct a Pearson correlation, the ACTFL rating scale was represented on a scale from 1-10. The Pearson correlation of the TrueNorth Speaking Test predictor range and the OPIc ACTFL rating was r = 0.90, indicating a very strong relationship between both tests.

Because the ACTFL scale is an ordinal scale, it isn't appropriate to assume that the spacing between the levels is uniform. This means that simply rounding the predictor score to associate it to the nearest ACTFL level is inappropriate. For purposes of this analysis, the TrueNorth score was simply converted to a range on the ACTFL ordinal rating scale (ex. between Intermediate High and Intermediate Mid). Then the number of instances were counted in which the ACTFL level determined by the OPIc was included within the range predicted by the TrueNorth Speaking Test. For 98 of the 108 test takers (90%) the ACTFL rating from the OPIc was included in the predicted range by the TrueNorth test and for 107 of the 108 test takers (99%), the rating was within one.

	Total	Advanced – Intermediate High	Intermediate High – Intermediate Mid	Intermediate Mid – Intermediate Low	%
Advanced	8	6	2	0	75%
Intermediate High	39	9	26	4	90%
Intermediate Mid	49	1	35	13	98%
Intermediate Low	11	0	2	9	82%

Human Scoring vs. Automatic Rating

A major portion of the TrueNorth Speaking Test scoring method includes comparing test-takers' audio recordings to the prompt audio utilizing a method called Elicited Imitation (Burdis, 2014; Erlam, 2006; Vinther, 2002). This methodology began with human raters and has only recently begun to incorporate speech recognition for scoring successfully in place of human raters (Graham et al., 2008). For this study, each test item was scored both by a human rater and the speech recognition engine from Carnegie Speech. The analysis of the human and automated scores for each item produced a Pearson correlation of 0.89, representing strong inter-rater reliability.

The predictor algorithm was re-run using the human scores for the TrueNorth test to determine whether there was a difference in ability to predict ACTFL ratings between a test using human or automated scores. The Pearson correlation of the speaking test predictor range using

human scores and the OPIc ACTFL rating was r = 0.89. The difference between the correlations was not statistically significant (P > .05) indicating that both correlation coefficients indicate strong positive relationships between the TrueNorth Speaking Test scores and the ACTFL ratings.

There was also no difference in the number of instances in which the ACTFL level determined by the OPIc was included within the range predicted by the TrueNorth Speaking Test. For 98 of the 108 test takers (90%) the ACTFL rating from the OPIc was included in the predicted range by the TrueNorth test and for 107 of the 108 test takers (99%), the rating was within one. Both scoring methods produced the same, positive results.

	Total	Advanced – Intermediate High	Intermediate High – Intermediate Mid	Intermediate Mid – Intermediate Low	%
Advanced	8	7	1	0	75%
Intermediate High	39	12	23	4	90%
Intermediate Mid	49	3	29	17	98%
Intermediate Low	11	0	2	9	82%

Summary

The data indicate that the TrueNorth Speaking Test has a very strong ability to predict the ACTFL rating based on the OPIc. It also shows that the automated scoring process provided by Carnegie Speech and utilized in the TrueNorth test performs just as well as human scoring for predicting the ACTFL ratings. The study also provides further evidence that although the TrueNorth Speaking Test measures the construct of “speaking ability” differently than traditional interview-based tests, its output conforms well to the traditional measures’ outputs, such as the ACTFL rating from the OPIc.

References

Burdis, J. R. (2014). Designing and evaluating a Russian elicited imitation test to be used at the Missionary Training Center (Doctoral dissertation). Retrieved from BYU ScholarsArchive. Paper 4008.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. Oxford: Oxford University Press.

Graham, R., Lonsdale, D., Kennington, C., Johnson, A., & McGhee, J. (2008). Elicited imitation as an oral proficiency measure with ASR scoring. In N. Calzolari (Conference Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, (Eds.), Proceedings of the 6th International Language Resources and Evaluation Conference (LREC’08), Marrakech, Morocco.

Vinther, T. (2002). Elicited imitation: A brief review. International Journal of Applied Linguistics, 12, 54–73.