The Development of an Automated Online Italian Speaking Proficiency Assessment

Introduction

In today's more globalized world, communication is becoming increasingly important. Thus, being able to validly, reliably, and efficiently measure speaking ability is also becoming increasingly important. Many organizations have undertaken the challenge of assessing speaking ability, and they have done so with sufficient validity and reliability. However, most of these solutions are expensive, lack scalability, and, therefore, lack efficiency.

An alternative to these traditional assessment solutions is TrueNorth's Speaking tests with their TrueNorth (TN) Score. Unlike other assessments, TrueNorth's speaking tests are automated and eliminate the expenses and complications of highly trained human interviewers, proctors, or raters. A fully web-deliverable assessment, TrueNorth speaking tests use a task type whereby test-takers listen to audio prompts of target sentences and repeat them verbatim. This methodology referred in literature is Elicited Imitation (EI) has been well validated in its ability to predict speaking proficiency. By extrapolating the fundamental cognitive components of spoken language, EI can then be used to estimate the ability to extemporaneously use and produce spoken language.

Scoring of TrueNorth speaking tests utilizes automated speech recognition (ASR) technology to compare each response (recorded by the test taker) to its target audio prompt and returns a score. These scores are then used to estimate speaking ability via graded response modeling, which is part of the item response theory framework (Samejima, 1969; Bock & Mislevy, 1982). Using machine learning, this raw speaking ability estimate and its corresponding standard error are then converted to a consumable, easy-to-understand score of speaking ability. In contrast to other time- and resource-intensive measures, a TrueNorth speaking test can be administered and scored in about 15 minutes.

At the time of the writing of this paper, tests have been released for English, Spanish, Portuguese, French, German, and Japanese. The purpose of the current technical paper is to document the development of the TrueNorth Italian Speaking test as a valid and reliable measure of Italian speaking ability. Following this introductory section, the method used to develop the Italian assessment is reported including its validation. Finally, some concluding remarks are made in support of the EIS test as a valid and reliable measure of Italian speaking ability

Method

Participants, Instruments, and Procedure

In collaboration with a stakeholder who regularly trains its representatives to speak Italian, two rounds of pilot testing were conducted. First, a rough estimate of speaking ability was determined via data available to the stakeholder and a language background survey administered prior to any testing, which designated the ability of each test-taker as likely beginner, intermediate, advanced, or mastery. Next, all test-takers (n = 181) responded to 60 elicited imitation items in the first round of assessment. Test-takers then completed the second round of testing, where they responded to an additional 24 elicited imitation items and 3 open-ended items aligned in their difficulty to the test taker's likely ability.

Elicited Imitation and Open-Ended Item Scoring

All elicited imitation and open-ended items were processed both by ASR and manually by highly trained raters who spoke Italian as their first language. The number of phonemes spoken correctly was divided by the total number of possible phonemes in the audio prompt and then multiplied by 100, generating a percentage score. Next, the hand-rated percentage scores were then used to validate the machine-rated percentage scores produced via ASR. The machine-rated scores correlated very highly with the hand-rated scores ($r = .91$, $p < .001$), indicating parity between the scores (Dorans & Walker, 2007).

All responses to the open-ended items were processed by two independent ASR technologies, producing speaking fluency scores (e.g., words per minute, percentage silence, and number of syllables). The fluency scores from these two technologies showed a strong correlation ($r = .78$, $p < .001$), indicating that both ASR technologies measured something in common.

Elicited Item Calibration

To utilize item response theory, the machine-rated percentage scores were first transformed to polytomous values ranging from 0-3 informed by transformations in prior assessment developments. Using 60 elicited imitation items, a parallel analysis (PA) was performed to test for unidimensionality as well as the appropriateness of fitting a graded response model (Horn, 1965). The polytomous scores were then used to calibrate each item's discrimination parameter, difficulty parameters corresponding to each polytomous value, and overall fit relative to the other items via graded response modeling. Poorly fitting items were identified and removed via the S_X^2 statistics (Kang & Chen, 2007; Orlando & Thissen, 2000; 2003). From the remaining items, Maximized Fisher information (MFI) was used to select the 30 most informative items representing a full range of speaking ability.

Concurrent Validity

As has been addressed in other technical papers, EI items on a TrueNorth speaking test employ partial construct coverage instead of full construct coverage where speaking tasks strive for authenticity (e.g., conversational dialog). Previous validation work used a separate third party assessment to serve as a separate measure of speaking ability.

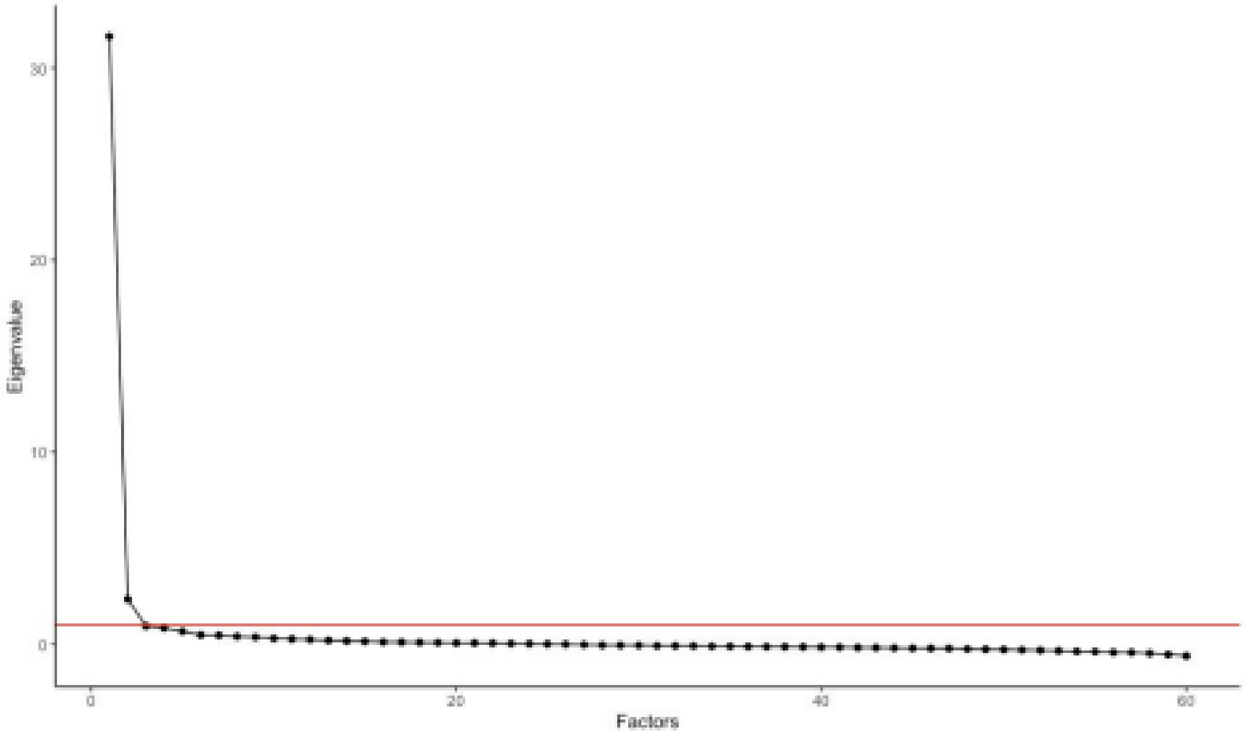
For the creation of the Italian assessment, Emmersion used a new approach. Based on data collected for the development of other language assessments, confirmatory factor analysis was conducted to model a latent trait theoretically similar to a latent trait approximated using full construct coverage. This latent trait was estimated using the speaking ability estimates derived from the elicited imitation items and speaking fluency scores. The fit of this model is reported in the results section.

Estimated scores derived from the latent trait alongside each test-taker's ability designation were used to create a 0-10 point scale that accords with established language frameworks. For example, a test-taker verified as an intermediate speaker could only receive a 4, 5, or a 6. Latent trait score ranges within each group of ability designations (except mastery) were used to score each test-taker within their respective score ranges.

Although there were progressive increases in latent trait scores across the ability designations, there was also some overlap. This, however, was not necessarily detrimental because it added some statistical noise seen previously between test-takers' scores on other assessments and elicited imitation speaking ability estimates.

For test-takers identified as beginner speakers, the range of elicited imitation estimates was divided into three quantiles. Those in the first quantile were given a 1, those in the second quantile were given a 2, and those in the third quantile were given a 3. Given the difficulty in discriminating between test-takers at the beginner level because of scarcity of produced speech, this rather crude but objective technique seemed appropriate.

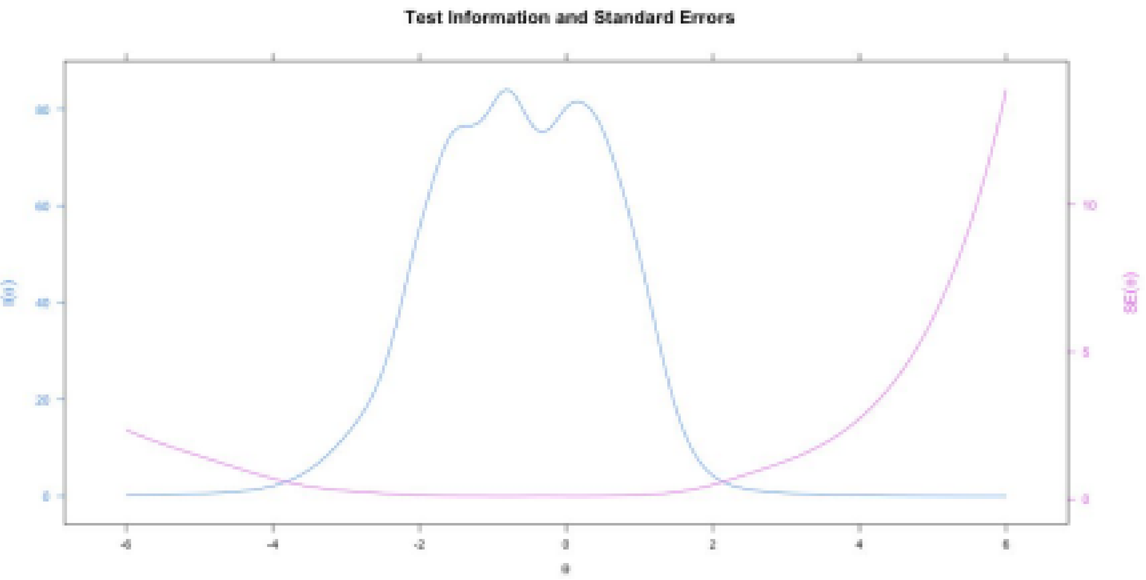
Figure 1



For intermediate speakers, those scoring at least one standard deviation below the mean were given a 4. Those scoring between one standard deviation below and above the mean received a 5. Those scoring at least standard deviation above the mean received a 6. For advanced speakers, those scoring at least one standard deviation below the mean were given a 7. Those scoring between one standard deviation below or above the mean received an 8. Those scoring higher than one standard deviation above the mean received a 9. All mastery speakers (confirmed educated native speakers) were given a 10.

These scores, denoted as Predicted_Ability, were used as the outcome variable in a supervised machine learning model. This model was trained using a subset of the elicited imitation ability estimates and their standard errors as features. The model was then tested on the remaining elicited imitation ability estimates and standard errors independent of those used to train the model. This predicted score is denoted as TNT_Score. The correlation between Predicted_Ability and TNT_Score was used as evidence of validity and is reported in the results section.

Figure 2



Results

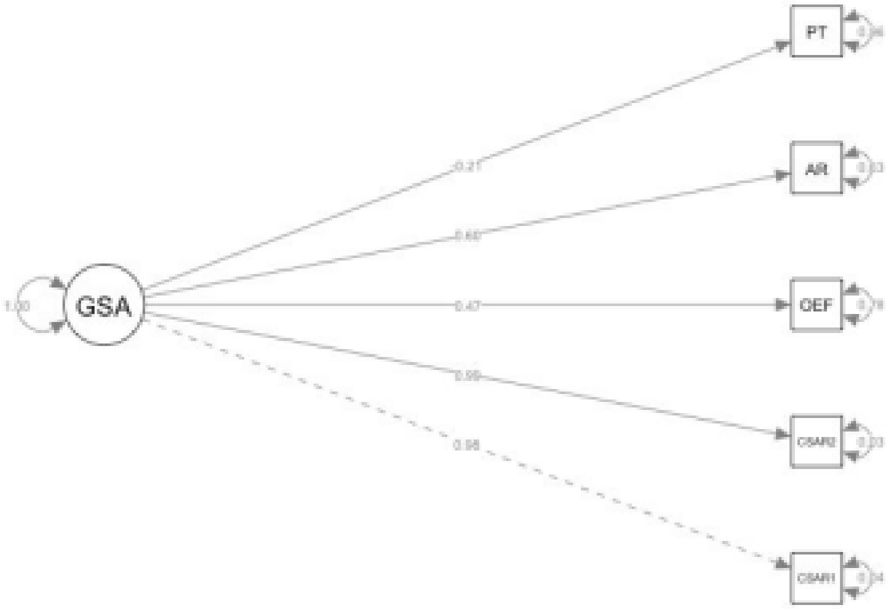
Unidimensionality

The Kaiser-Meyer-Olkin test (KMO) was used to determine the appropriateness of performing a parallel analysis using the polytomous scores from the elicited imitation items. This test indicated that the data were appropriate for parallel analysis, MSA = .97. The eigenvalues estimated by the parallel analysis for polychoric correlations indicated that two factors could be extracted from the data. However, examination of a scree plot of the eigenvalues and the ratio of the largest eigenvalue (i.e., 31.65) to the smallest eigenvalue (i.e., 2.33) indicated that the assumption of unidimensionality was not violated (Morizot, Ainsworth, & Reise, 2007). (See Figure 1.)

Item Response Theory Analysis

Using a dataset comprising only well-performing items, indices of fit indicated that a graded response model represented the elicited imitation items well, CFI = 1.0; TLI = 1.0; RMSEA = .00; SRMR = .05. Further, the discrimination parameters all ranged from moderate to very high (i.e., .76 to 5.70; Baker, 2001). Maximized Fisher information was then used to create a fixed form elicited imitation test comprising 30 items intended to measure across a range of ability. Figure 2 shows that the final fixed form TrueNorth's Italian speaking test does a reasonable job of providing information and precision across a range of speaking abilities.

Figure 3



Speaking Ability as a Latent Trait

Two elicited imitation speaking estimates from Round 1 and Round 2, an open-ended response speaking fluency score from one ASR technology, and phonation time derived from the other ASR technology loaded significantly on a latent trait representing speaking ability (see Figure 3). The fit of the latent trait measurement model representing speaking ability was excellent, $\chi^2(3) = 3.57$, $p = .61$; CFI = 1.00; TLI =1.00; RMSEA = .00, .12; SRMR = .03.

This model was used to calculate a score representing overall speaking ability. Discrepancies between Predicted_Ability and TNT_Score large enough to be considered statistical outliers were identified and removed. The remaining scores correlated strongly with Predicted_Ability, $r_s = .74$, $p < .001$. Thus, confidence was gained in the use of Predicted_Ability as an outcome variable for training a machine learning model for use as a scoring algorithm.

Predictive Machine Learning Model Training and Validation

For the predictive scoring algorithm, a stochastic dual coordinate ascent model was fitted by regressing Predicted_Ability onto the speaking ability estimate and its standard error as measured by the 30-item EIS test. As previously mentioned, a subset of the calibration data was used to train this model, while the remaining subset of data was used to test the model. The correlation between Predicted_Ability and TNT_Score was very strong, r_s = .94, indicating parity between the scores and that 88.4% of the variance in Predicted_Ability could be accounted for by TNT_Score (Dorans & Walker, 2007). (See Figure 4.)

Summary

The current technical report provides evidence in support of the EIS test as a valid and reliable measure of speaking ability. This evidence includes 1) analyses indicating the appropriateness of using a graded response model for estimating speaking ability 2) a confirmatory factor analysis indicating the appropriateness of measuring general speaking ability across constrained and unconstrained speaking tasks; 3) and a correlation analysis indicative of a strong relationship between scores derived from the elicited imitation items and the scores estimated by the general speaking ability measurement model. Thus, we are confident that the EIS test can be used as a valid and reliable assessment of Italian speaking ability.

References

Baker, F. (2001). The Basics of Item Response Theory. College Park: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6(4), 431-444.

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In Linking and aligning scores and scales (pp. 179-198). Springer, New York, NY.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179-185.

Kang, T., & Chen, T. T. (2007). An Investigation of the Performance of the Generalized SX 2 Item-Fit Index for Polytomous IRT Models. ACT Research Report Series, 2007-1. ACT, Inc.

Morizot, J., Ainsworth, A. T., & Reise, S. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), Handbook of Research Methods in Personality Psychology (pp. 407–423). New York, NY: Guilford.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. Applied Psychological Measurement, 27(4), 289-298.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement, 24(1), 50-64.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika, 34(S1), 1–97. https://doi.org/10.1007/bf03372160