The Development of an Automated Online Portuguese Speaking Proficiency Assessment

Introduction

As the world moves toward an increasingly globalized community, the ability to reliably and accurately measure speaking proficiency becomes increasingly imperative. Institutions who collaborate with international partners need to ensure that the professionals they hire to represent their mission can fluently speak a language common to both parties. To overcome this challenge, language assessment institutions have developed sophisticated measures of language ability that can reliably and accurately measure speaking proficiency. The problem with these measures, however, is that they can be expensive and difficult to scale.

To provide scalable and affordable alternatives to these assessments, Emmersion Learning has developed automated online speaking proficiency assessments for multiple languages. Using elicited imitation as the theoretical framework (Vinther, 2002; Erlam, 2006; Burdis, 2014), Emmersion Learning analyzes recorded oral repetitions of auditory prompts via speech recognition services and compares these records against the auditory prompt. This technology can accurately measure speaking proficiency in as little as 15 minutes, affording institutions the ability to inexpensively and accurately assess speaking ability at scale.

The purpose of the current document is to report on the development and validation of TrueNorth's Portuguese Speaking Test. The following section describes the method used to develop and validate this test. After that, the results of the development and validation are reported. Finally, this document concludes with some final comments.

Method

Test responses were collected in collaboration with a stakeholder who helped pilot a Portuguese speaking proficiency assessment comprising 60 elicited imitation test items. Test responses were also collected via Amazon's Mechanical Turk (MTurk), a crowdsourcing platform hosted online. Except for examinees recruited via MTurk, examinees were also given the American Council on the Teaching of Foreign Language (ACTFL) Oral Proficiency Interview-Computer (OPIc). The final sample size comprised 146 examinees. Examinees'

responses to the 60-item Portuguese pilot test were rated manually by trained raters. Responses were also rated by a third-party speech recognition analysis software and compared against the manually rated items to validate the software as a viable component of the test.

The Portuguese pilot test items were calibrated via partial credit modeling (Masters, 1982), which was used to estimate up to three location parameters per item corresponding to their difficulty. The results of these calibrations were then used to make further modifications to the model by removing misfitting items and response patterns.

Misfitting items and response patterns were identified by computation and analysis of their infit and outfit statistics (Linacre, 2002b). According to these statistics, 36 test records and 10 items were removed, resulting in a final dataset comprising 110 examinees and 50 items. From these 50 items, 30 items were selected to assemble the test by dividing the range of item difficulty values into 30 equidistant values, calculating item information across these values for all 50 items, and then judiciously selecting the most informative items across this range. The results of the following analysis are based on this final dataset.

Rasch reliability and separation statistics were computed and analyzed to determine if the 50 calibrated items were sensitive enough to distinguish between low and high performers and to determine if the sample size was large enough to confirm the range of item difficulties (Wright & Masters, 1982). Further, Spearman's rho was also computed to examine the reliability between examinees' estimated ability and their ACTFL levels. The results of this analysis are reported in the results section.

Because of the similarities between Portuguese and Spanish, differential item functioning (DIF) and differential test functioning (DTF) analyses were performed on the final version of the test to determine if the test was unfairly biased toward either examinees who were fluent in Spanish or examinees who were not fluent in Spanish. The results of these analyses are reported in the results section.

Lastly, examinees' scores on ACTFL's OPIc were used for two purposes: 1) to provide validity evidence and 2) to build a scoring algorithm for predicting ACTFL speaking proficiency levels of future examinees. The percentage agreement between actual ACTFL speaking proficiency levels and the predicted ACTFL speaking proficiency levels calculated using the predictive model provides additional evidence of validity. The final section summarizes and discusses the results obtained from these analyses.

Results

First, to estimate a polytomous partial credit model, examinees' response scores were converted from a continuous scale ranging from 0-100 to a categorical scale ranging from 0-3. The thresholds at which the values were transformed were modified until an approximately equal distribution between categories was obtained after calibration. Further, the Andrich threshold advances for adjacent categories were between the recommended 1.4 to 5 logits (see figure 1; Linacre, 2002a). This implies that each four-category polytomous item could theoretically be divided into three independent dichotomous items and that non-informative spacing between adjacent categories due to large Andrich threshold advances (greater than 5 logits) was minimized.

Figure 1

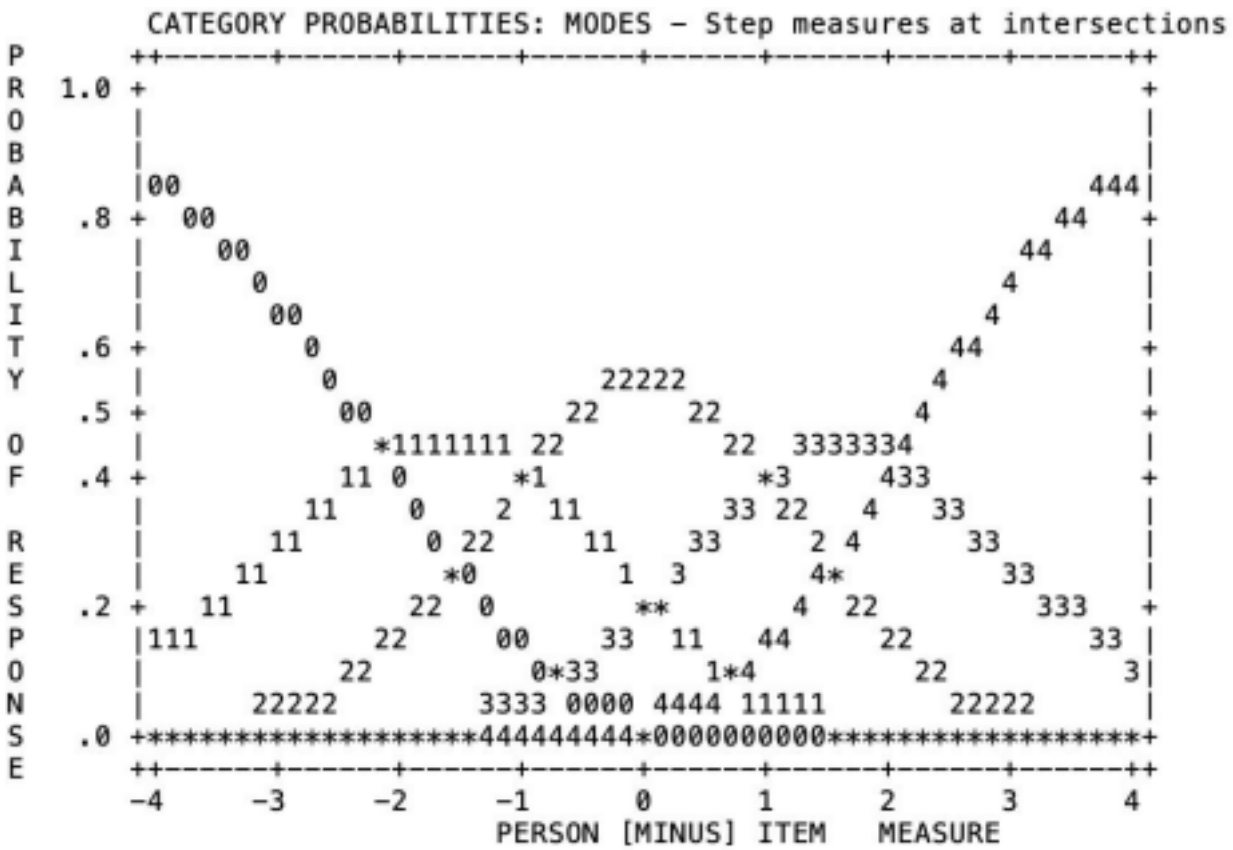


FIgure 2


PERSONS
MAP OF IT
<frequ>|<less>
<rare>|<more>
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
Q
S
M
30
29
28
27
26
22 23 24 25
19
15 20 21
14 16 17 18
12 13 9
10 11 7 8
6
5
3 4
2
1

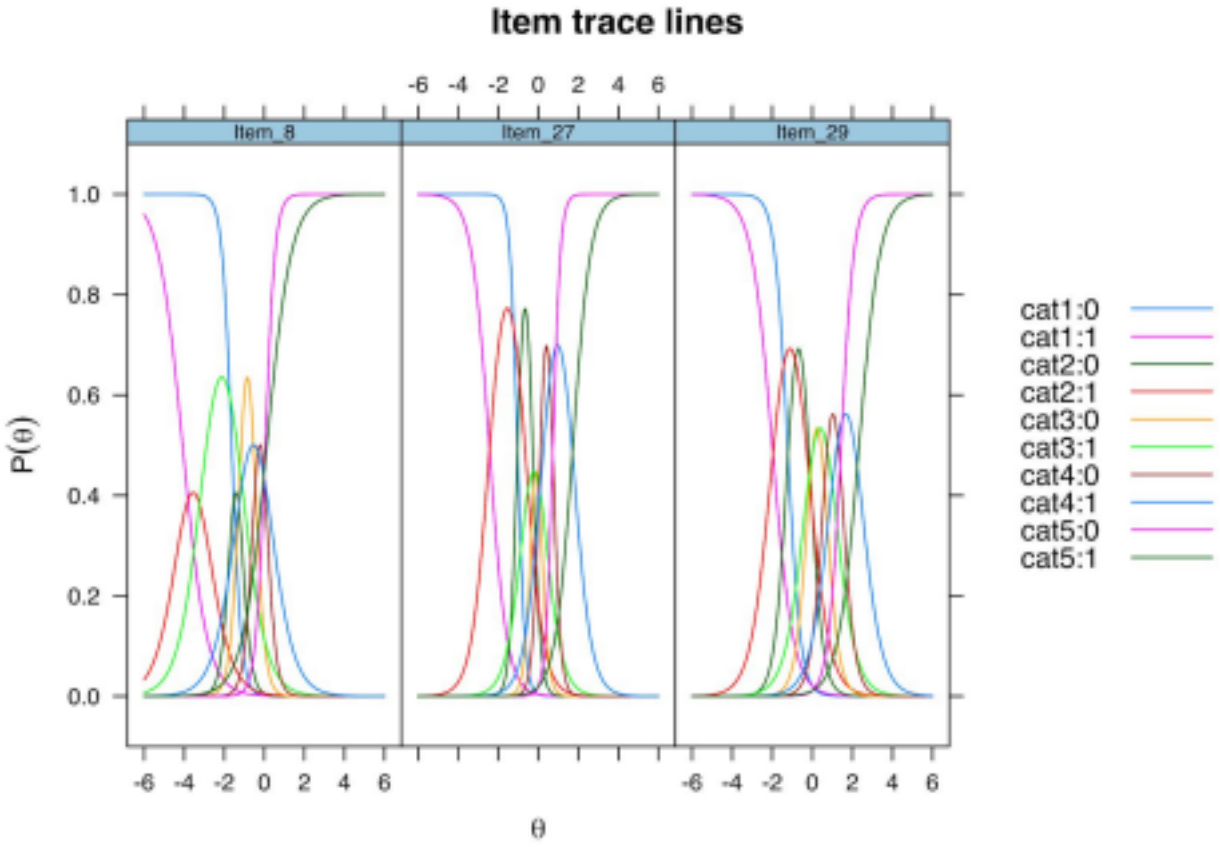
Person reliability, which is a measure of the reproducibility of person ability estimation, was .97, indicating that examinees who have lower or higher scores than other examinees, indeed, have lower or higher scores than other examinees. The person separation index is defined as the ratio between the true spread of ability and measurement error, and it is used to estimate the number of ability strata that are distinguishable within the range of item difficulties and person abilities (Wright & Masters, 2002). The person separation index was 5.76, indicating that approximately 8 ability strata were statistically distinguishable. Similar to person reliability, item reliability refers to the reproducibility of item difficulty estimation. Item reliability was .99, indicating that items with lower or higher difficulties than other items, indeed, had lower or higher difficulties than other items. Similar to the person separation index, the item separation index is defined as the ratio between the true spread of item difficulty and measurement error, and it is used to estimate the number of difficulty strata that are distinguishable within the range of item difficulties and person abilities. The item separation index was 8.46, indicating that approximately 12 difficulty strata were statistically distinguishable.

A Wright Map illustrated in figure 2 shows the distribution of examinee by ability (shown on the left) relative to the distribution of the 50 calibrated items by difficulty (shown on the right), which are on the same scale (Wright & Stone, 1979). This indicates that the examinees and items were appropriately distributed across the range of abilities and difficulties.

Treating the test and OPIc as separate raters of examinee speaking proficiency, the Spearman rank correlation between examinees' Portuguese estimated abilities and their ACTFL level scores was computed as a measure of interrater reliability. According to this analysis, there was an acceptably high degree of interrater reliability between the two speaking proficiency measures, $r_s = .909$, $p < .001$. Treating the manually rated and speech recognition software rated examinee responses as separate raters of Portuguese TNT speaking proficiency, the Spearman rank correlation between manual and speech recognition software ratings was acceptably high, $r_s = .869$, $p < .001$

Because of similarities between Portuguese and Spanish, and because several of the examinees were fluent in Portuguese, the test was examined for differential item functioning (DIF) and differential test functioning (DTF). According to this analysis, one item's DIF (i.e., Item 29) reached statistical significance, $\chi^2 = 4.55$, $p < .05$. Despite this result, the signed DTF (sDTF), which is a measure of the average directional bias, showed that the overall DTF of the test did not reach statistical significance, indicating that the test did not favor one group over the other, sDTF = -7.01, p = .388, 95% CI [-25.29, .83] (Chalmers, Counsell, & Flora, 2016). Unsigned DTF (uDTF), a measure of the average absolute bias across groups,

Figure 3



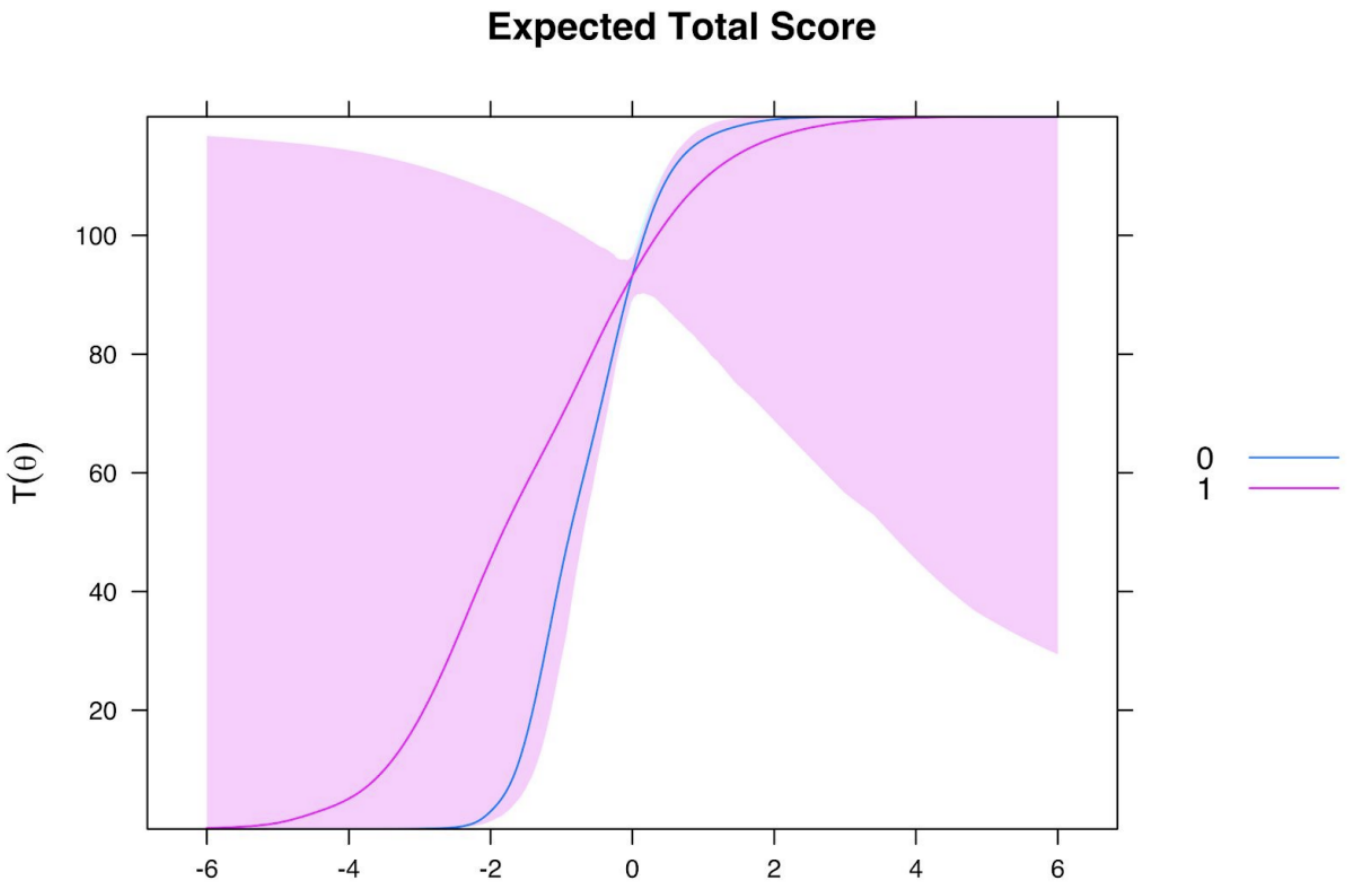
which could range from 0 to 150 (the total possible score of the test), indicated that the overall bias was not an issue, uDTF = 9.18, 95% CI [.42, 79.92] (Chalmers, Counsell, & Flora, 2016). Visual inspection of the items displaying the most DIF showed the similarity in response patterns between the groups (see figure 3). Further, visual inspection of each group's test characteristic curve suggested that the Portuguese TNT performed similarly between them (see figure 4).

A predictive model was developed by fitting a generalized additive model with OPIc ACTFL levels expressed as the sum of the smooth functions of the predictors (i.e., theta, or ability, and its standard error). This analysis showed that the generalized additive model explained 91% of the deviance, a model fit statistic that is a generalization of the coefficient determination (i.e., R^2) used in conventional regression. This indicated that the model had an acceptably high level of fit (see figure 5). Predicted ACTFL levels calculated using the estimated generalized additive model correlated highly with examinees' actual ACTFL levels at $r_s = 0.904$, $p < .001$.

In contrast to the ordinal actual ACTFL levels, the predicted ACTFL levels were continuous. Thus, to examine the percentage agreement between actual and predicted ACTFL levels, the

predicted ACTFL levels were rounded down and up to their nearest whole number values, and these two values were compared against the actual ACTFL levels. These values were then examined to see if they fell within a range of one level below or above their corresponding actual ACTFL levels and within a range of two levels below or above their corresponding actual ACTFL levels. According to this analysis, there was 73% agreement within the predicted range and 95% agreement within one level of the predicted range.

Figure 4



Summary

The current technical paper offers several lines of evidence in support of TrueNorth's Portuguese Speaking test as a reliable and valid measure of Portuguese language speaking proficiency: 1) person and item reliability and separation statistics indicated that the sample size and corresponding ability range were sufficiently large to determine the item difficulty hierarchy and that the number of items and their difficulty range were sufficiently large to

determine the person ability hierarchy; 2) the Spearman rank correlation between examinees' estimated Portuguese ability on the TrueNorth scale and their ACTFL levels indicated that the reliability between the speaking proficiency measures was acceptably high; 3) differential item functioning and differential test functioning statistics indicated that the test did not unduly favor non-Spanish speakers over Spanish speakers, or vice versa, and 4) the generalized additive model estimated for predictive purposes reached an acceptable level of accuracy in predicting examinees' ACTFL levels. Thus, the current study offers compelling evidence in support of TrueNorth's Portuguese Speaking test as a veritable measure of Portuguese speaking proficiency and of the third-party speech recognition software as a viable tool for rating examinee responses.

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.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. Educational and Psychological Measurement, 76(1), 114-140.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. Applied linguistics, 27(3), 464-491.

Linacre, J. M. (2002a). Optimizing rating scale category effectiveness. Journal of applied measurement, 3(1), 85-106.

Linacre, J. M. (2002b). What do infit and outfit, mean-square and standardized mean. Rasch Measurement Transactions, 16(2), 878.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149-174.
Vinther, T. (2002). Elicited imitation: A brief overview. International journal of applied linguistics, 12(1), 54-73.

Wright, B. D., & Stone, M. H. (1979). Best test design.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. MESA press.

Wright, B. D., & Masters, G. N. (2002). Number of person or item strata. Rasch Measurement Transactions, 16 (3), 888.