The Development of an Automated Online Spanish Speaking Proficiency Assessment

Introduction

The ability to reliably and accurately measure speaking proficiency is becoming increasingly imperative as the world moves toward an increasingly globalized community. Institutions who interact with international partners as part of their mission must ensure that their emissaries can communicate in a language common to both parties. To meet this goal, language testing organizations have developed sophisticated language measures that can reliably and accurately measure speaking proficiency. An issue with these measures, however, is that the resources required to administer them are neither affordable nor scalable.

Emmersion Learning has developed automated online speaking proficiency tests for multiple languages to provide scalable and affordable language assessment. Using elicited imitation as the theoretical framework (Vinther, 2002; Erlam, 2006; Burdis, 2014), Emmersion Learning analyzes recorded oral productions of auditory prompts via third-party natural language processing technology. These recordings are compared against prototypical auditory prompts to which examinees are instructed to listen to and repeat verbatim. Speaking proficiency is measured in as little as 15 minutes, affording institutions a scalable and affordable solution for measuring language ability.

The purpose of the current document is to report on the development and validation of the TrueNorth Spanish Speaking test. The following section describes the method used to develop and validate the test[1]. Following this, the results of the development and validation are reported. Lastly, this document concludes with some final remarks.

Method

Having collected test responses across several versions of TrueNorth's Spanish Speaking test, we used concurrent calibration using test questions overlapping across multiple test forms as anchor items. These data comprised 1,040 test records. Using the American Council on Teaching Foreign Languages' (ACTFL) Oral Proficiency Interview-computer (OPIc) assessment

[1]

as our gold standard for criterion validity, 332 of these test records also comprised examinees' scores denoting their level of speaking proficiency. It is against these latter test scores that the test was validated as a reliable and valid measure of Spanish speaking proficiency.

All of the Spanish test items were calibrated using a graded response model within an item response theory framework (Samejima, 1969). Misfitting response patterns were identified and removed via examination of the Zh statistic developed by Drasgow, Levine, and Williams (1985). Larger absolute values of the Zh statistic are indicative of aberrant response patterns. According to these values and the stringent criteria that was implemented, we identified and removed 30% of the response, resulting in a final dataset comprising 729 test records out of the initial 1,040.

Although a non-Rasch model was estimated for calibration, a partial credit model was estimated to compute Rasch reliability and separation statistics to determine if the test items were sensitive enough to distinguish low from high performers and to determine if the sample of examinees was large enough to confirm the range of item difficulties (Masters, 1982; Wright & Masters, 1982). Additionally, Spearman's rho was computed to examine the reliability between examinees' estimated ability and their ACTFL levels. These results are reported in the next section.

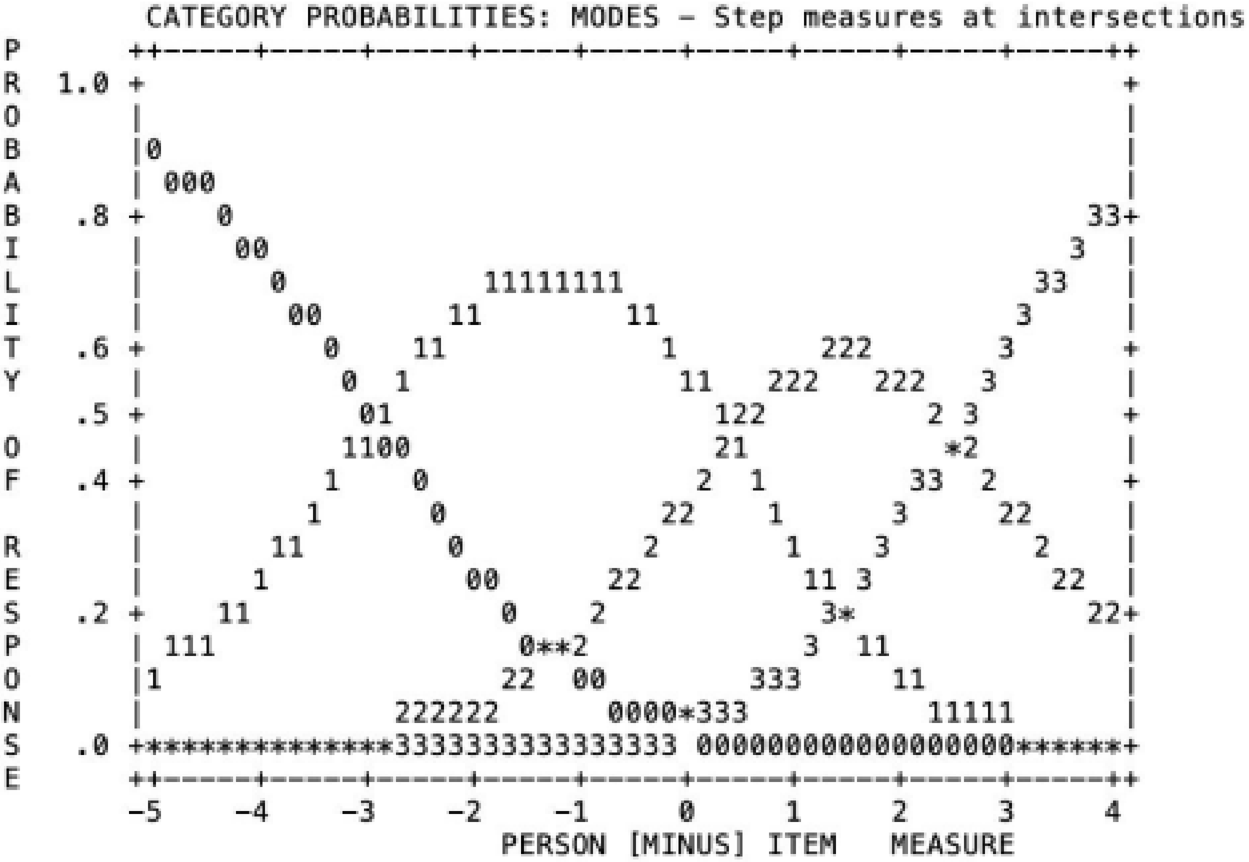
Lastly, examinees' scores on ACTFL's OPIc were used to provide evidence of validity and to build a scoring algorithm for estimating ACTFL speaking proficiency levels for future examinees. The percentage agreement between actual ACTFL speaking proficiency levels and the predicted ACTFL speaking proficiency levels calculated using the predictive model provides additional evidence of validity. The final section provides conclusions based on these results.

Results

First, to estimate a graded response model using polytomous data, 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. The distance between the adjacent apexes of these Andrich thresholds representing the polytomous categories were between the recommended 1.4 to 5 logits (see figure 1; Linacre, 2002a). Thus, each four-category polytomous item could theoretically be divided into three independent dichotomous items, and non-informative spacing between adjacent categories due to large Andrich threshold advances (greater than 5 logits) was minimized.

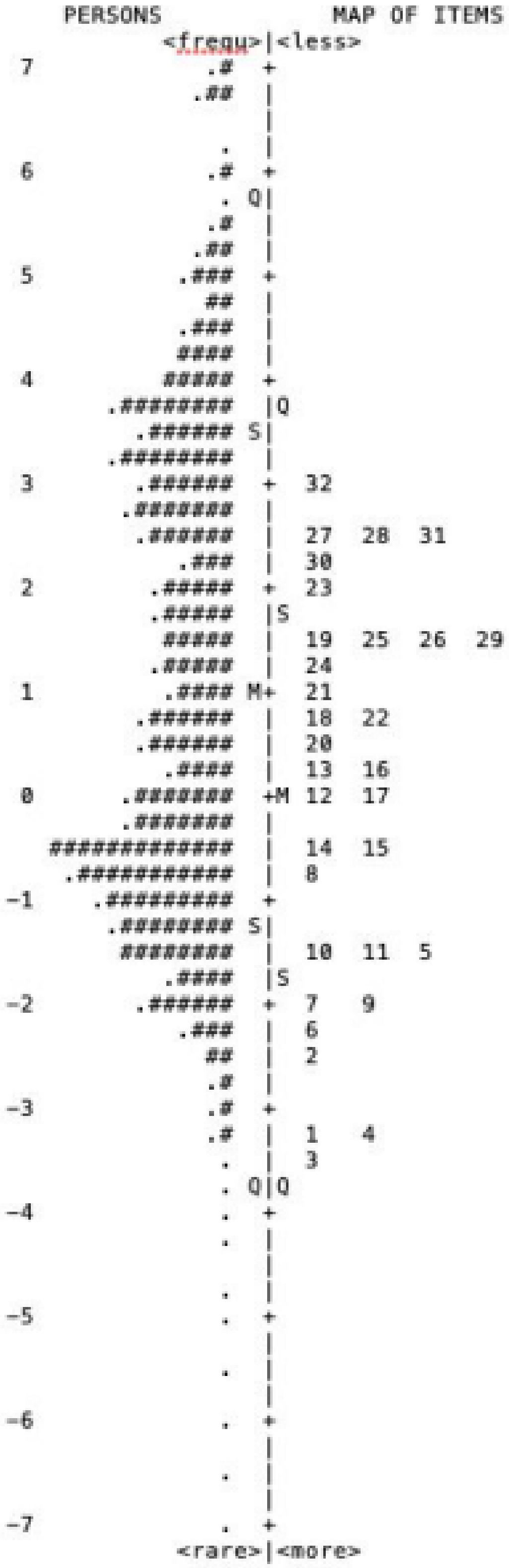
Person reliability, which is a measure of the reproducibility of person ability estimation, was .97, indicating that examinees who have lower or higher scores than other examinees, indeed, have lower or higher scores than other examinees. The person separation index is defined as the ratio between the 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 5.49, 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 1.0, 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 19.16, indicating that approximately 26 difficulty strata were statistically distinguishable.

Figure 1



```
        CATEGORY PROBABILITIES: MODES - Step measures at intersections
P       ++-----+-----+-----+-----+-----+-----+-----+-----+-----++
R  1.0 +                                                       +
O      |                                                       |
B      |0                                                      |
A      | 000                                                   |
B   .8 +    0                                                33+
I      |     00                                             3  |
L      |       0           11111111                       33   |
I      |        00       11        11                    3     |
T   .6 +          0    11            1        222       3      +
Y      |           0  1               11   222   222   3       |
    .5 +            01                  122         2 3        +
O      |           1100                 21           *2        |
F   .4 +          1    0               2  1        33  2       +
       |         1      0            22    1      3     22     |
R      |       11        0          2       1    3        2    |
E      |      1           00      22         11 3          22  |
S   .2 +    11              0    2            3*             22+
P      | 111                 0**2            3  11             |
O      |1                   22  00        333     11           |
N      |              222222      0000*333          11111      |
S   .0 +**************3333333333333333 000000000000000000******+
E       ++-----+-----+-----+-----+-----+-----+-----+-----+-----++
       -5    -4    -3    -2    -1     0     1     2     3     4
                            PERSON [MINUS] ITEM   MEASURE
```

Figure 2



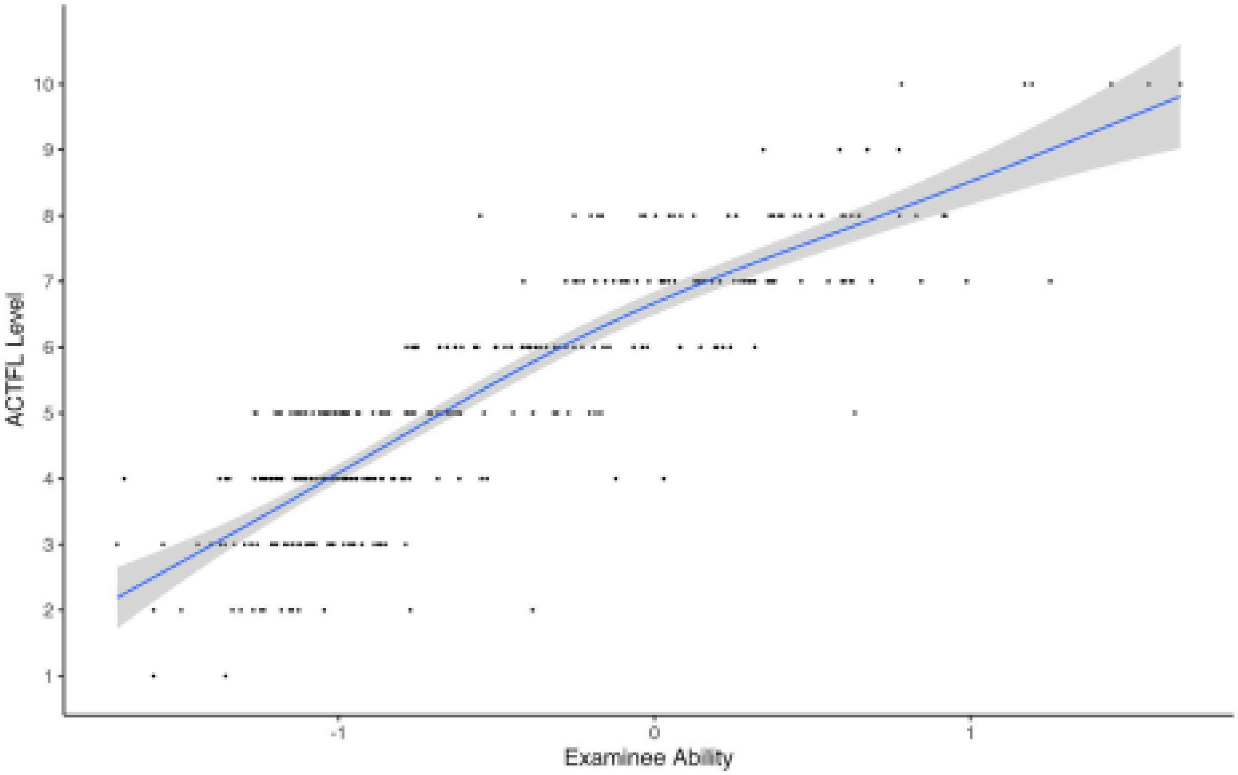
```
    PERSONS            MAP OF ITEMS
           <frequ>|<less>
 7            .#  +
             .##  |
                  |
               .  |
 6            .#  +
               . Q|
              .#  |
             .##  |
 5          .###  +
              ##  |
            .###  |
            ####  |
 4         #####  +
       .########  |Q
         .###### S|
       .########  |
 3       .######  +  32
        .#######  |
         .######  |  27  28  31
            .###  |  30
 2        .#####  +  23
          .#####  |S
           #####  |  19  25  26  29
          .#####  |  24
 1         .#### M+  21
         .######  |  18  22
         .######  |  20
           .####  |  13  16
 0      .#######  +M 12  17
        .#######  |
   #############  |  14  15
    .###########  |  8
-1    .#########  +
       .######## S|
        ########  |  10  11  5
           .####  |S
-2       .######  +  7   9
            .###  |  6
              ##  |  2
              .#  |
-3            .#  +
              .#  |  1   4
               .  |  3
               . Q|Q
-4             .  +
               .  |
                  |
               .  |
-5             .  +
                  |
               .  |
                  |
-6             .  +
                  |
               .  |
                  |
-7             .  +
            <rare>|<more>
```

A Wright Map illustrated in figure 2 shows the distribution of examinee by ability (shown on the left) relative to the distribution of the 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 alternate forms measuring examinee speaking ability, the Spearman rank correlation between examinees' Spanish 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, $r_s = .854, p < .001$.

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 89% of the deviance, which is a generalization of the coefficient determination (i.e., R^2) used in conventional regression. This indicated that the model had an acceptable level of fit (see figure 3). Predicted ACTFL levels calculated using the estimated generalized additive model correlated highly with examinees' actual ACTFL levels at $r_s = 0.863, p < .001$.

Figure 3



In contrast to the ordinal 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 were within 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 75% agreement within the predicted range and 95% agreement within one level of the predicted range.

Summary

The current technical paper provides evidence in support of the TrueNorth Spanish Speaking test as a reliable and valid measure of Spanish language speaking proficiency. Person and item reliability and separation statistics indicated that the examinee sample and its ability range were sufficient for determining the item difficulty hierarchy and that the number of items and their difficulty range were sufficient for determining the person ability hierarchy. The Spearman rank correlation between examinees' estimated ability and their corresponding ACTFL levels indicated that the reliability between the speaking proficiency measures was acceptably high. The precision of the estimated generalized additive model for predicting examinees' ACTFL levels was acceptable. Thus, the current report offers compelling evidence in support of the TrueNorth Spanish Speaking test as a reliable and valid measure of Spanish speaking proficiency.

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.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

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.