The Development of an Automated Online French Speaking Proficiency Assessment

Introduction

As the world becomes more globalized, it is becoming increasingly imperative that its denizens can effectively communicate with each other. Organizations need emissaries who can interact with their stakeholders and partners, who oftentimes come from different parts of the world and speak different languages. To ensure that their emissaries can, indeed, communicate with stakeholders and partners who speak different languages, sophisticated assessments that measure speaking ability or proficiency have been developed. These assessments, however, are oftentimes expensive and difficult to scale.

TrueNorth's speaking tests, which are currently available in English, Japanese, Portuguese, and Spanish, have been developed to overcome these limitations. More assessments in other languages have been slated for development in the near future. All of TrueNorth's speaking tests have been developed using elicited imitation as its methodological framework, which simply requires its examinees to repeat verbatim sentences presented to them auditorily (Vinther, 2002; Erlam, 2006; Burdis, 2014). How closely examinees' responses resemble the auditory prompt is determined via powerful automated speech recognition services that utilize natural language processing. These assessments can measure speech ability in as little as 15 minutes without the need of language experts to score examinee responses. Further, these assessments are relatively inexpensive and highly scalable.

The current document's purpose is to report on the development and validation of Emmerson's French Speaking Test (EFS test). It is divided into four consecutive sections: an introduction section (this section), a method section, a results section, and summary section. The following section describes the method used to develop the EFS test.

Method

A preliminary version of the EFS test comprising 60 elicited imitation items was piloted in collaboration with a stakeholder who regularly sends its emissaries around the world to achieve its mission. Before sending them to countries where the predominant language spoken is different from their native language, each emissary is given extensive, immersive training and experience in the target non-native language. The stakeholder requires valid and reliable language assessments for 1) evaluation of their language training programs and 2) measurement of the speaking ability of the emissaries as well as their growth in the target language.

Response data were also collected via Amazon's Mechanical Turk (MTurk), an online crowdsourcing platform. Except for examinees recruited via MTurk, the Oral Proficiency Interview-computer (OPIc), developed by the American Council on the Teaching of Foreign Languages (ACTFL), was administered to examinees to glean evidence of criterion and convergent validity and to develop a scoring algorithm. The final calibration dataset comprised 150 examinees, 95 of which had OPIc scores.

Prior to item calibration, each item's response score, a value returned by expert human raters (which are used to create a beta version of the assessment before a final version is released) ranging from 0-100, was transformed to a polytomous value ranging from 0-4. Scores ranging from 0-10 were converted to 0s, scores ranging from 10-50 were converted to 1s, scores ranging from 50-90 were converted to 2s, and scores greater than or equal to 90 were converted to 3s.

The French pilot test items were then calibrated via a graded response model (Samejima, 1969). The calibration results were used to assemble an assessment representative of the full range of speaking ability and to remove misfitting items and response patterns. Misfitting items were identified and removed via the S-X^2 statistic (Orlando & Thissen, 2000; 2003; Kang & Chen, 2007). Misfitting response patterns were identified and removed via the Zh statistic (Drasgow, Levine, & Williams, 1985). After item and response pattern removal, the final dataset comprised 126 examinees, 78 of which comprised OPIc scores, and 50 items.

A Rasch reliability analysis was conducted to determine if the sample size was sufficient for confirming the range of item difficulties and if the items were sensitive enough to distinguish between low and high performers (Wright & Masters, 1982). Thus, the pilot data were also fitted to a partial credit model, but this model was not directly used to assemble an assessment or developing a scoring algorithm (Masters, 1982). Additionally, Spearman's rho was computed to examine the reliability between estimated ability and their corresponding OPIc scores, which is reported in the results section.

The quality of the scoring algorithm was confirmed by checking the reliability and agreement between its predicted OPIc scores and the examinees' actual OPIc scores. The percentage of perfect agreement was checked by dividing the number of predicted OPIc scores that matched their corresponding actual OPIc score by the total number of actual OPIc scores. Additionally, the percentage of agreement where the predicted OPIc score was no less or more than one from the actual OPIc score was checked by dividing the number of predicted OPIc scores that met this criterion by the total number of actual OPIc scores. Lastly, Spearman's rho was calculated to check the reliability between the predicted and actual OPIc scores.

The following section discusses the results derived from the method used to develop and validate the EFS test.

Results

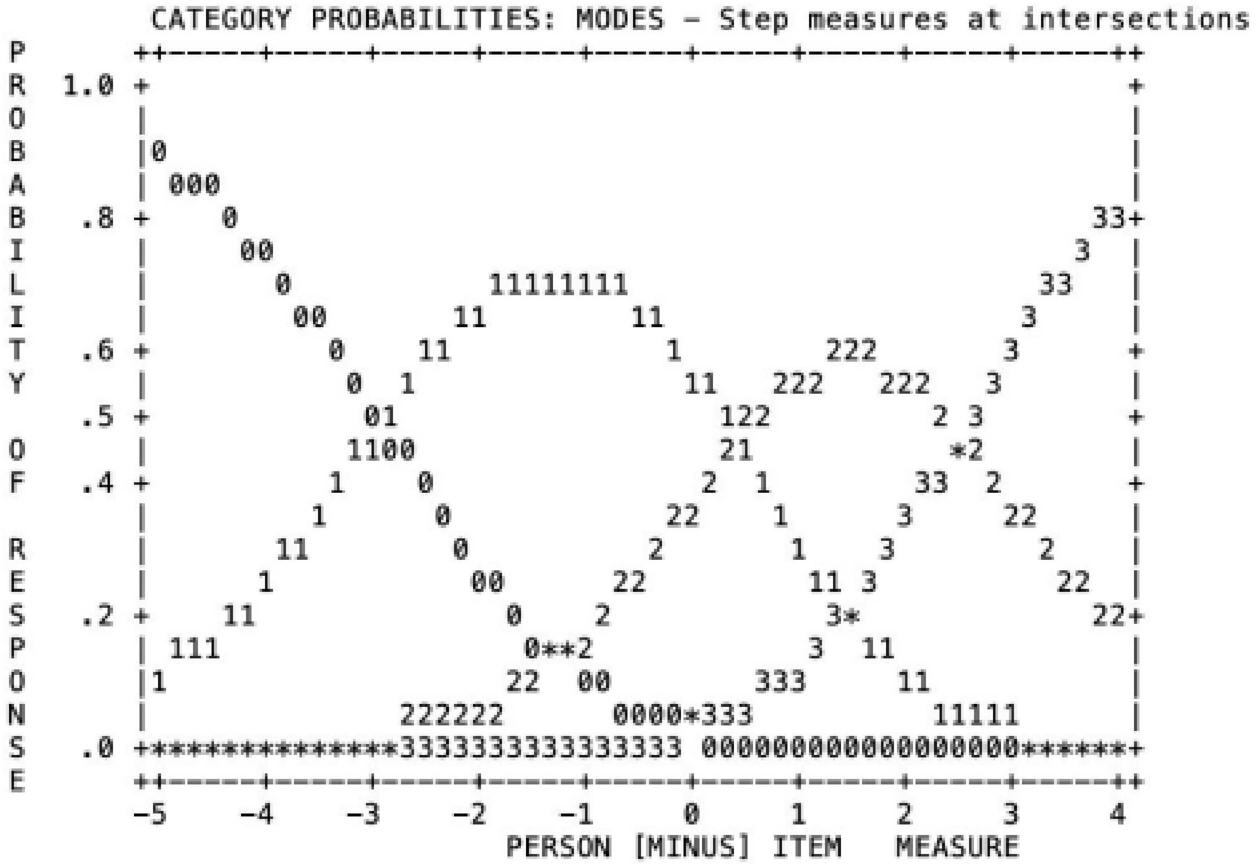
To ensure that the thresholds at which the expert human rater scores were converted to polytomous categories resulted in an approximately equivalent distribution, the Andrich threshold advances for the adjacent categories were verified to be between the recommended 1.4 to 5 logits (see figure 1; Linacre, 2002). This implied that each item could be divided into three dichotomous items and that non-informative spacing between the adjacent categories, indicated by large Andrich threshold advances (greater than 5 logits), was minimized.

Person reliability, a value ranging between 0-1 that indicates how well an assessment can distinguish between low and high performers, was .98. This indicated that the items comprising the French TNT could reliably distinguish between examinees across the ability range. The ratio between the true spread of ability and measurement error, defined as the person separation index, is used to estimate the number of strata that are available to categorize examinees by ability (Wright & Masters, 2002). The person separation index was 8.98, indicating that approximately 12 ability strata were statistically distinguishable.
Item reliability, a value ranging from 0-1, indicates the reproducibility of item difficulty hierarchy.

Item reliability for the French items was equal to .99, indicating a high likelihood that the item difficulty hierarchy could be reproduced using another sample of examinees. The ratio between the true spread of item difficulty and measurement error, defined as the item separation index, is used to estimate the number of difficulty strata that are distinguishable within the range of item difficulties. The item separation index was 12.46, indicating that approximately 17 difficulty strata were statistically distinguishable.

The Wright Map in Figure 2 shows the distribution of examinees by ability on the left side relative to the distribution of items by difficulty on the right side, which are on the same measurement scale (Wright & Stone, 1979). This shows that the examinees and items were well distributed across the entire range.

Figure 1

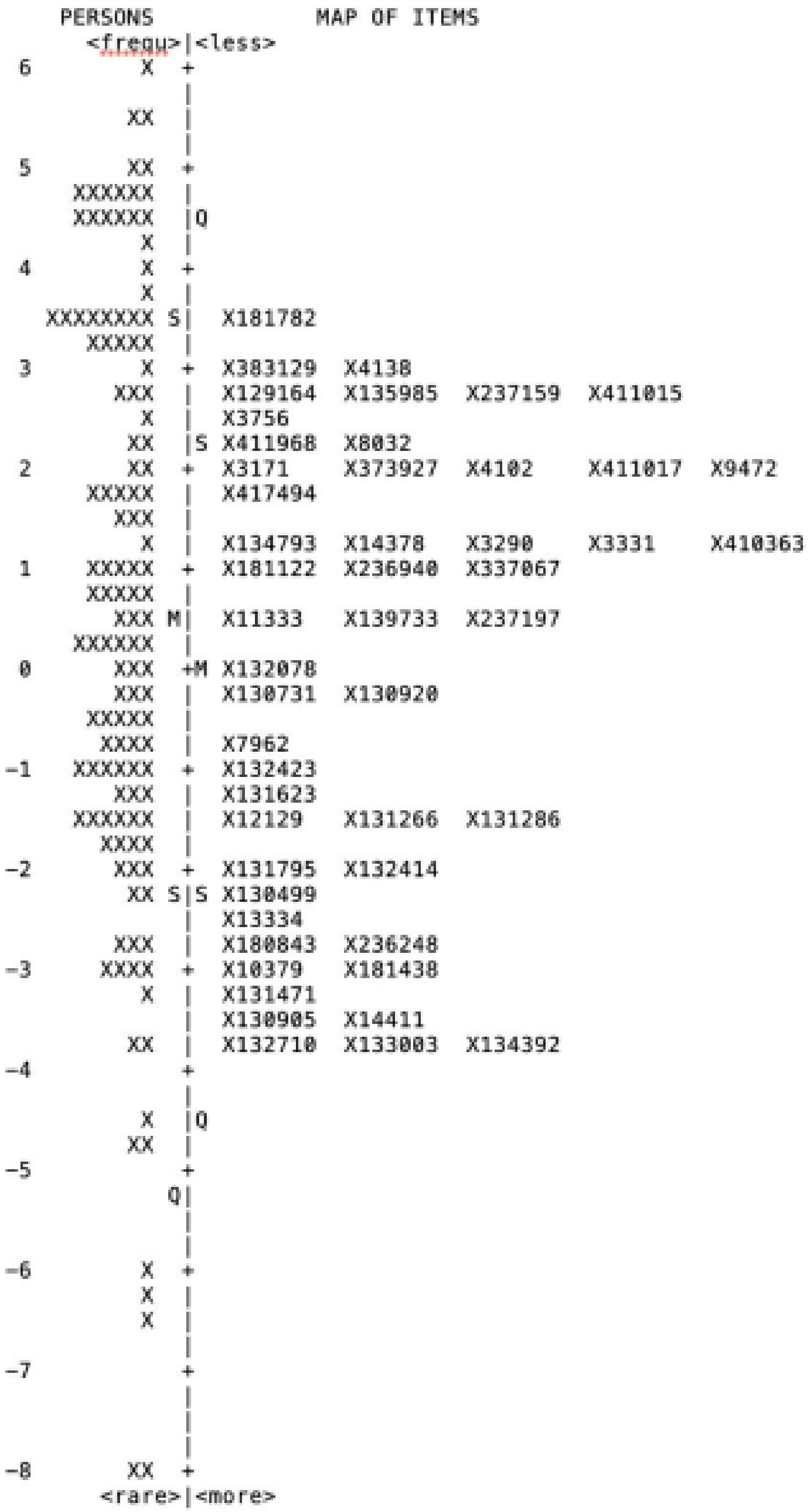


```
        CATEGORY PROBABILITIES: MODES - Step measures at intersections
P      ++-----+-----+-----+-----+-----+-----+-----+-----+-----++
R  1.0 +                                                       +
O      |                                                       |
B      |0                                                      |
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
```

Treating the EFS test and OPIc as alternate measures of examinee speaking ability, the Spearman rank correlation between examinees' scores on the pilot French assessment and the OPIc was computed to ensure an appropriate level of reliability. Spearman's rho indicated that there was an excellent level of reliability between speaking ability estimated by the EFS test's TrueNorth score and the actual OPIc score, $r_s = .900$, $p < .001$.

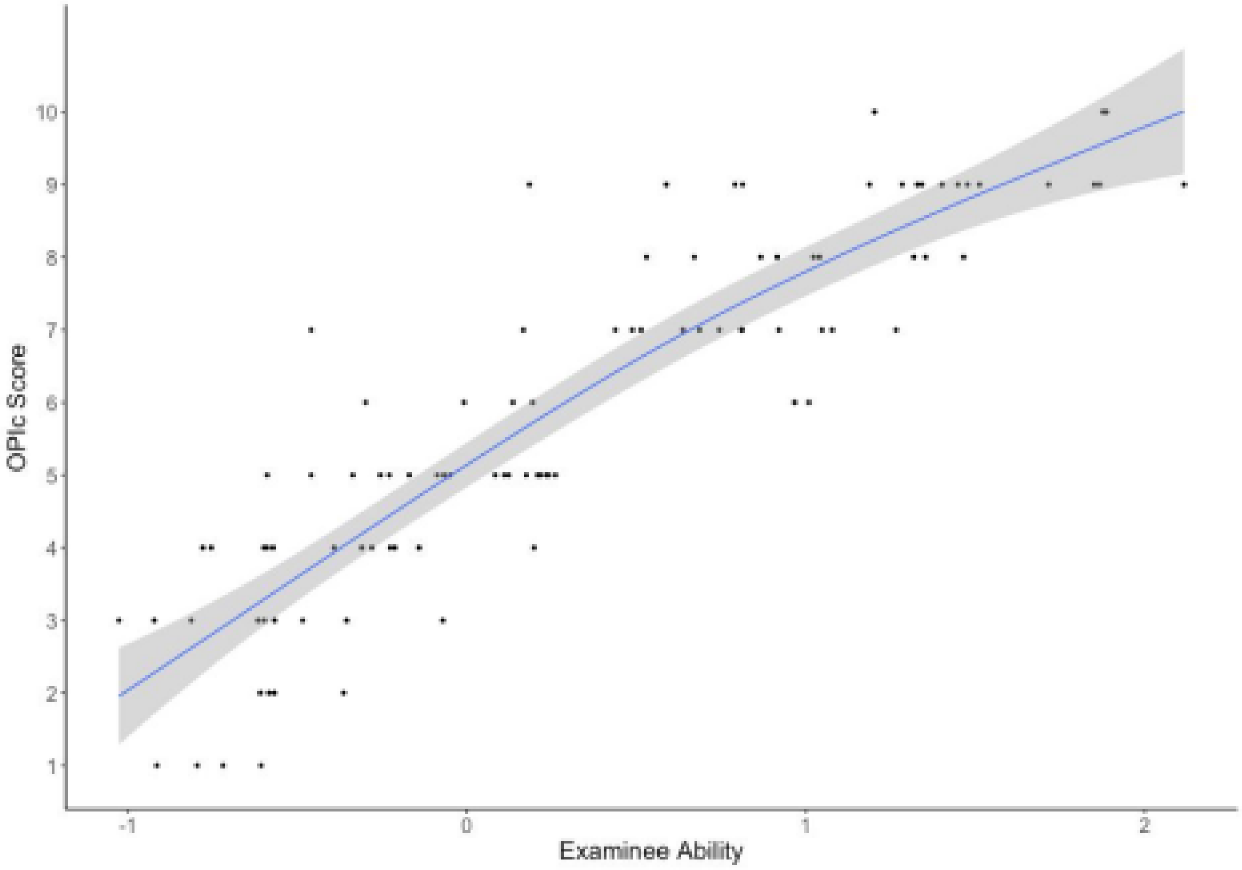
A predictive model using generalized additive modeling was fitted by expressing OPIc scores as the sum of the smooth functions of the predictors (i.e., theta, which represents ability, and its standard error). The model fit the data well, explaining 91.3% of the deviance, which is a generalization of the coefficient determination (i.e., R^2) used in conventional regression to generalized additive modeling (see Figure 3). The predicted OPIc scores calculated using generalized additive model correlated highly with examinees' actual OPIc scores at $r_s = 0.909$, $p < .001$. Further, agreement between predicted OPIc scores and actual OPIc scores met our internal guideline of at least a 70% perfect match and a 90% match within 1 in either direction.

Figure 2



```
    PERSONS            MAP OF ITEMS
      <frequ>|<less>
 6        X  +
             |
         XX  |
             |
 5       XX  +
     XXXXXX  |
     XXXXXX  |Q
          X  |
 4        X  +
          X  |
   XXXXXXXX S|  X181782
      XXXXX  |
 3        X  +  X383129  X4138
        XXX  |  X129164  X135985  X237159  X411015
          X  |  X3756
         XX  |S X411968  X8032
 2       XX  +  X3171    X373927  X4102    X411017  X9472
      XXXXX  |  X417494
        XXX  |
          X  |  X134793  X14378   X3290    X3331    X410363
 1    XXXXX  +  X181122  X236940  X337067
      XXXXX  |
        XXX M|  X11333   X139733  X237197
     XXXXXX  |
 0      XXX  +M X132078
        XXX  |  X130731  X130920
      XXXXX  |
       XXXX  |  X7962
-1   XXXXXX  +  X132423
        XXX  |  X131623
     XXXXXX  |  X12129   X131266  X131286
       XXXX  |
-2      XXX  +  X131795  X132414
         XX S|S X130499
             |  X13334
        XXX  |  X180843  X236248
-3     XXXX  +  X10379   X181438
          X  |  X131471
             |  X130905  X14411
         XX  |  X132710  X133003  X134392
-4           +
             |
          X  |Q
         XX  |
-5           +
            Q|
             |
             |
-6        X  +
          X  |
          X  |
             |
-7           +
             |
             |
             |
-8       XX  +
       <rare>|<more>
```

Figure 3



Summary

Evidence has been presented in support of the EFS test as a valid and reliable measure of speaking ability: 1) reliability and separation statistics as well as a Wright Map showing that the range of item difficulties and examinee ability was sufficient for assessment development; 2) the Spearman rank correlation between speaking ability and OPIc scores indicated that the reliability between the speaking ability measures was excellent; 3) the model used to predict OPIc explained 91.3% of the deviance; 4) the Spearman rank correlation between predicted OPIc scores and actual OPIc scores indicated that their reliability was excellent; and 4) the generalized additive model predicted; and 5) the predictive model met our guideline for an acceptable level of agreement between predicted OPIc scores and actual OPIc scores. Thus, we believe the current paper offers compelling evidence supporting the reliability and validity of the EFS test as a measure of speaking ability.

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.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1), 67-86.

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

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.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.

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 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. (2002). Number of person or item strata. *Rasch Measurement Transactions*, 16 (3), 888.