TRUENORTH

Technical Report

Japanese TrueNorth Test (TNT)

March 19, 2019

The Development of an Automated Online Japanese Speaking Proficiency Assessment

Introduction

Being able to communicate with others in languages other than a native one is becoming an increasingly important skill as the world moves toward a more globalized community. Institutions who interact with international partners need to ensure that their emissaries can communicate in a language common to both parties. One way these institutions can facilitate this communication is by having the language speaking proficiency of each of its emissaries assessed. Measuring language speaking proficiency is a challenging undertaking, which only a few testing and assessment organizations have attempted. Those that have taken the challenge have developed sophisticated tests, some of which require proctors to administer the test and highly trained raters to evaluate examinee responses. Others have even developed computer-based automated language proficiency assessments, which, unlike the proctored tests, immediately returns a score upon completion. Like the proctored tests, however, these tests can be prohibitively expensive, making them difficult to scale.

To overcome these issues, Emmersion Learning has developed affordable and scalable alternatives to these assessments for various languages, and it is actively increasing the number of languages that it measures. Using elicited imitation as its methodological framework (Vinther, 2002; Erlam, 2006; Burdis, 2014), Emmersion Learning has developed automated tests that have examinees' listen to and repeat sentences spoken to them in a target language. Using third-party speech recognition technology, Emmersion Learning rates the quality of examinees' spoken responses relative to their corresponding audio prompts, and then it estimates examinee speaking proficiency using sophisticated psychometric models and predictive algorithms. The speaking proficiency of an examinee can be estimated in as little as 15 minutes.

The purpose of the current report is to document the development and validation of the Japanese TNT. The following section describes how the Japanese TNT was developed and validated. Next, the results of these development and validation processes are discussed. Finally, the document concludes with a discussion on the implications of these findings.

Method

Test responses were collected in collaboration with a stakeholder who regularly trains and dispatches emissaries to various countries across the world to achieve its goals. Examinees were administered a 60-item pilot test and the Oral Proficiency Interview – Computer (OPIc), which is a reliable and accurate measure of speaking proficiency for multiple languages developed by the American Council on the Teaching of Foreign Languages (ACTFL).

The final sample comprised test records from 129 examinees. Examinees' responses to the 60-item Japanese pilot test were rated manually by trained raters and by a third-party speech recognition software. Manual ratings were used in the development of a Portuguese speaking proficiency assessment to ensure that the speech recognition software was functioning properly.

To identify misfitting items and response patterns via infit and outfit statistics (Linacre, 2002b), item and person parameters were first estimated using a partial credit model (Masters, 1982). According to infit and outfit statistics, 23 test records were removed and the 30-item Japanese TNT was assembled. Based on this estimated partial credit model, Rasch reliability and separation statistics were computed and analyzed to determine if the 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).

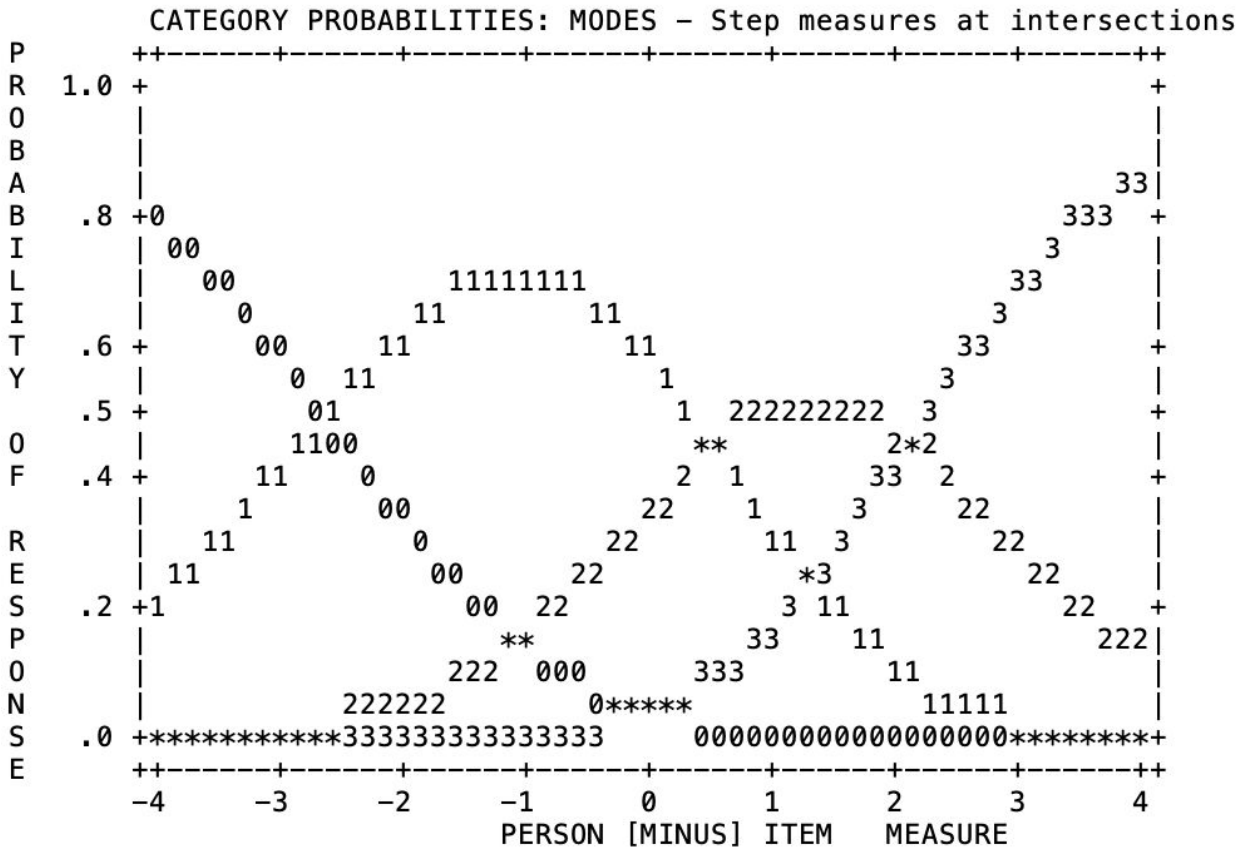
The Japanese pilot test items were then calibrated via graded response modeling (Samejima, 1969), which was used to estimate up to three threshold parameters per item corresponding to the percentage of accuracy obtained by the examinees. Results from this calibration were used to develop a statistical model for predicting examinees' ACTFL levels. The percentage agreement between actual ACTFL speaking proficiency levels and the predicted ACTFL speaking proficiency levels model provided additional evidence of validity. Finally, the last section concludes by discussing the implications of these results. Finally, Spearman's rank correlation was calculated to examine the reliability between examinees' estimated ability and their actual ACTFL levels. The results of this analysis are reported below.

Results

First, to estimate a polytomous partial credit model to identify misfitting items and response patterns and calculate reliability and separation statistics, 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 appropriate distribution between categories was obtained. Further, the Andrich threshold advances for adjacent categories were

between the recommended 1.4 to 5 units on the abscissa (see figure 1; Linacre, 2002a). Two implications for these statistics is that 1) each four-category polytomous

Figure 1. Japanese TNT rating category distribution



item could theoretically be divided into three dichotomous items and that non-informative spacing between adjacent categories due to excessive Andrich threshold advances (greater than 5 units) was minimized.

Person reliability, a measure of the reproducibility of the person ability hierarchy, was an acceptably high .96. The person separation index, defined as the ratio between the true spread of ability and measurement error and is used to estimate the number of statistically distinctive ability strata, was 4.99, indicating that there were approximately seven distinguishable ability strata. Similar to person reliability, item reliability, a measure of the reproducibility of the item difficulty hierarchy, was an acceptably high .99. 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 is used to estimate the number of statistically distinctive difficulty strata, was 10.27, indicating that there were approximately 14 distinguishable difficulty strata.

The Wright Map shown in figure 2 illustrates the distribution of examinees by ability (shown on the left) relative to the distribution of items by difficulty (shown on the right) on the same scale (Wright & Stone, 1979). This shows that examinees and items were appropriately distributed across their respective ability and difficulty.

Next, a generalized additive model expressed as the sum of the smooth functions of the predictors (i.e., theta, or ability, and its standard error) was estimated to predict ACTFL levels. This model explained 98.2% of the deviance, which is a generalization of the coefficient determination (i.e., R^2) used in more conventional regression. This indicated that the model explained nearly all of the variance (see figures 5). Using Spearman rank correlation, predicted ACTFL levels calculated using the estimated generalized additive model correlated highly with examinees' actual ACTFL levels at $r_s = 0.913$, $p < .001$.

In contrast to the ordinal, actual ACTFL levels expressed as whole number integer values, the predicted ACTFL levels were continuous. Thus, to calculate the percentage agreement between actual and predicted ACTFL levels, the predicted ACTFL levels were rounded down and up to their nearest whole number integer values. These values were examined to determine if they fell within one level below or above their corresponding actual ACTFL levels and within two levels below or above their corresponding actual ACTFL levels. According to this analysis, there was 77% agreement within one the predicted range and 94% agreement within one level of the predicted range. These results were computed using manually rated examinee responses.

Summary

The current technical paper offers evidence in support of the Japanese TNT as a reliable and valid measure of Japanese language speaking proficiency. For example, person and item reliability and separation statistics indicated that the sample size and corresponding ability range were sufficiently large to determine the person ability and item difficulty hierarchies; the Spearman rank correlation between examinees' estimated Japanese TNT ability and their actual ACTFL levels indicated that the reliability between the speaking proficiency measures was acceptably high; and the generalized additive model estimated reached an acceptable level of agreement in predicting examinees' ACTFL levels. Thus, the current report offers compelling evidence supporting the Japanese TNT as a valid measure of Japanese speaking proficiency.

Figure 2. Wright Map

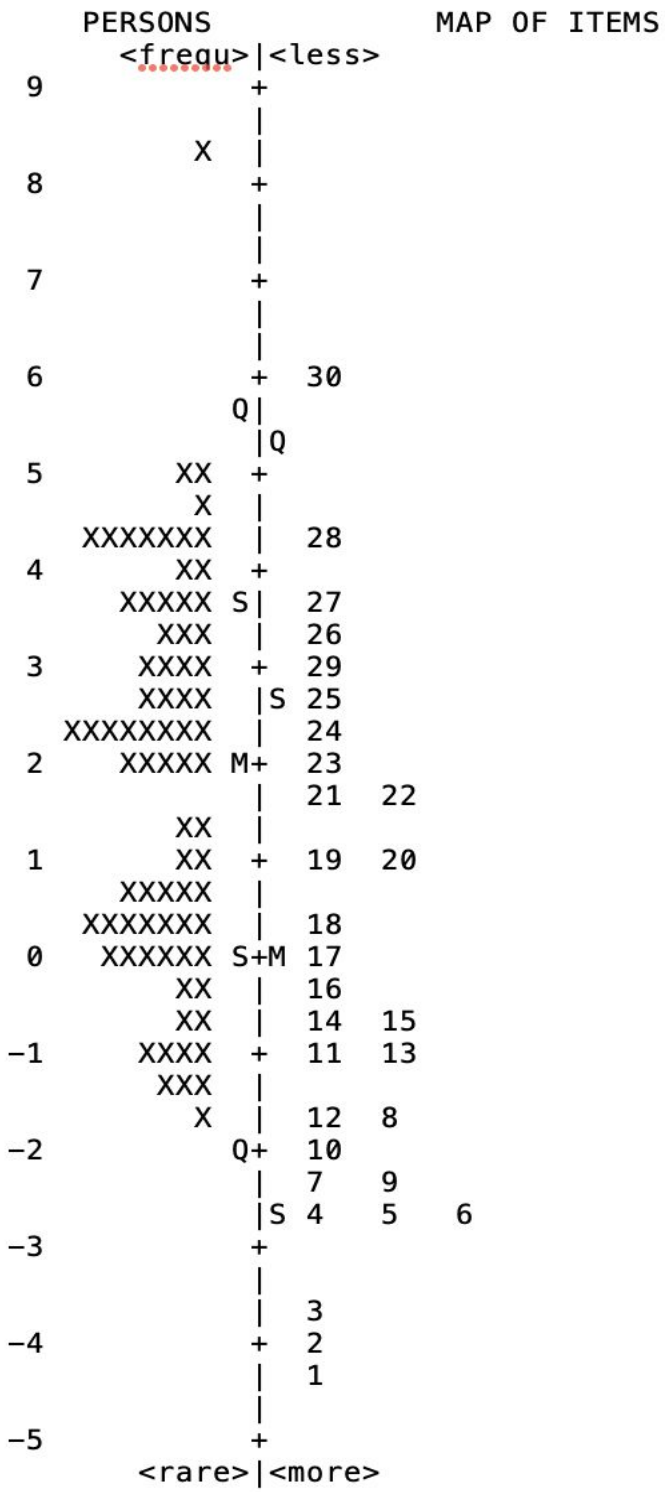
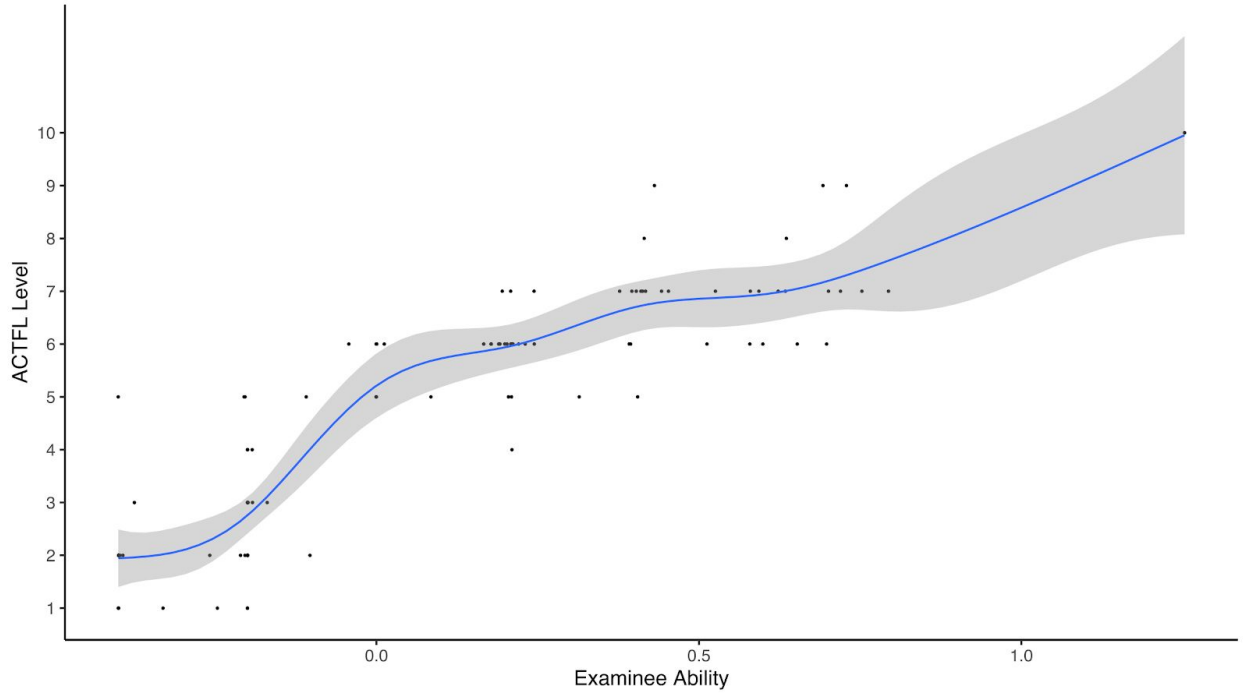


Figure 3. Scatterplot illustrating the relationship between ACTFL levels and Japanese TNT ability estimates



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.