The Development of an Adaptive, Automated Online English-Speaking Assessment

Introduction

The purpose of this white paper is to highlight the technical work behind the version update of the TrueNorth English Speaking test from Version C.19.03 to Version D.20.09. This was a significant update in test form[1]. Earlier forms of test including Version C.19.03 utilized a fixed form for Part 1 - Listen and Repeat. In this form, 30 items were presented in a set order from items that targeted ability at Novice/Beginning, Intermediate, Advanced, Superior/Mastery levels. Thus, regardless of ability each test taker would be presented with the full battery of items in a set order.

Items are scored using an automated speech recognition (ASR) solution and informed the final calculation of an ability estimate. This ability estimate drives the selection of open response questions for Part 2 that provided opportunity to demonstrate extended discourse ability at this predicted ability level. These early semi-adaptive versions of the test serve well their intended purpose of providing a fast, reliable, accurate, scalable measure of speaking ability. The version has been used to measure the speaking ability of tens of thousands of individuals around the world.

However, from the outset of the test's development, the creators of the test have had a vision of fully exploiting the unique characteristics of elicited imitation items (i.e., list-and-repeat items) that allow for the immediate automatic scoring of each test item at a level of precision that begins to quickly reveal insight about the test takers ability. Instead of inefficiently presenting a rigid fixed battery of items, a test form could be dynamically generated to custom fit the test taker's ability profile as it emerges. As a test taker's ability successfully meets the rigor of an item, the test selects a more difficult item. If the difficulty of an item overwhelms the ability of the test taker, an easier item is selected.

While a fully adaptive automated test powered by elicited imitation has been theorized for decades, the concept was robustly proofed internally before development began (Mayne, Hart, & Burdis, unpublished). This research showed that the available bank of calibrated items was broad and deep enough to power adaptive test forms that could discriminate across the full

[1]

range of the TrueNorth (TN) scale. It also showed that there would be remarkable efficiencies gained by reducing the number of items that would be presented.

While the total items that must be presented dropped from 30 to well below 10 in most simulations, modeling showed that the number of items providing the most information about a test taker's ability would actually increase in an adaptive test form. This allows the shortened test to retain and even gain reliability and precision to discriminate ability. This early research also provided compelling evidence that there would naturally be far more appropriate and desirable levels of item exposure and new barriers to test fraud behaviors with an adaptive version of the test.

The purpose of the current technical report was twofold. The first purpose was to develop a scoring algorithm via a supervised machine learning algorithm where scores on the adaptive version of the test were used to predict scores on the earlier fixed forms of the test. The second purpose was to validate scores derived from the adaptive test as being interchangeable with the scores derived from the fixed versions.

Method

Participants

Participant data were collected from multiple sources. These sources include an intensive English language learning program at a university in the United States (n = 136), a nonprofit organization dedicated to the support and development of refugees (n = 21), and administrators, teachers, and students in secondary schools in Tonga (n = 68) as well as other countries (n = 38). Participants whose time between the administration of the 2 instruments exceeded 30 days were excluded from the development of the scoring algorithm.

Instrument

We used the Version C of the test as the standard for validating the adaptive Version D. As previously mentioned, Version C comprised 30 elicited imitation items presented in order from least to most difficult. Each item was presented via an audio file comprising the recording of a voice actor or voice actress speaking a target sentence to be repeated verbatim by the examinees. The validity and reliability of this assessment had been established in previous validation studies (Emmersion, 2019; Habing, Grego, & Vessilinov, 2020).

We used a fully functional prototype of the adaptive test to examine the feasibility of its administration in a real-world setting. Its selection criterion algorithm was based on maximized

Fisher information (MFI), which computed the amount of information items provided given an examinee's provisional ability estimate. To reduce item exposure, 1 item was randomly selected from the 10 most informative items given examinees' provisional ability estimates. Provisional and final estimates of ability and their corresponding standard errors were computed via expected a posteriori (EAP; Bock & Mislevy's, 1982). To maintain face validity, a minimum of 12 items is administered to all examinees. To maintain efficiency relative to the fixed version, the maximum number of items administered to examinees was set to 20. All assessments ended when either at least 12 items were administered and the standard error was below .224 or when a total of 20 items had been administered.

Procedure

All examinees took both the fixed version and the adaptive version. The order in which the assessments were administered was counterbalanced. Before the start of each assessment, examinees responded to 3 items intended to test the computer microphone and ensure that the audio quality was sufficient for analysis via a third-party ASR application programming interface (API).

Data Analysis

Descriptive statistics were conducted to examine relevant behavioral characteristics exhibited by the examinees while taking the adaptive version (e.g., test duration and number of items administered). These characteristics were correlated with examinees' performance on the adaptive and fixed versions.

Next, we took participants' EAP ability estimates (thetas) and standard errors from the adaptive test and paired them with their corresponding scores from the 0-10 scale from the fixed test. We then randomly sampled 70% of these scores to generate a dataset for training a stochastic dual coordinate ascent (SDCA) model (Shalev-Schwartz & Zhang, 2013). We used the remaining 30% of the data as a testing dataset to evaluate the performance of the model. Because of a restriction in the range of ability, we simulated an additional 20 respondents representing true beginners. Ten of these simulated scores were appended to the training dataset, and the other tens simulated scores were appended to the testing dataset.

Finally, we examined the model metrics to evaluate the performance of the train SDCA model. This gave us an idea of how well the model performed. We also examined the Spearman rank-order correlation between the scores produced by the SDCA model and the scores derived from the fixed TNT. This provided evidence of criterion-related validity and construct validity.

Results

Overall descriptive statistics and group level statistics on testing behavior characteristics and outcomes are reported in the Table 1 below. These statistics include the mean, standard deviation, and the median length of time in days between taking the adaptive version and the fixed version, the reported score from the fixed version, the ability estimate from the adaptive version, the standard error of the ability estimate from the adaptive version, the number of adaptive items administered, and the number of audio errors.

As expected, speaking ability as measured by the adaptive test was positively correlated with the number of items, r_s = .46, p = .000. Because of the lack of more difficult items relative to easier items, more items were required to measure the language ability of the more capable examinees. Interestingly, speaking ability as measured by the adaptive version was negatively correlated with the number of audio errors that were detected in the adaptive TNT, r_s = -.16, p = .009. Although the relationship was modest, it was still unexpected because the more capable examinees responded to more items, giving them more opportunities to make mistakes. Indeed, there was no correlation between the number of items administered and the number of audio errors, r_s = .01, p > .05.

According to the trained SDCA model, 92% of the variance in the fixed version scores in the testing dataset was explained by the variance of the scores produced by the model. The root mean square error (RMSE), a measure of the discrepancy between actual and predicted results, indicated that the trained model did a good job of predicting scores, RMSE = .72. The mean absolute error, or the average distance from the actual score a predicted score deviates, was .58. This indicated that a predicted score was, on average, .58 points higher or lower than its corresponding actual score. Because the RMSE gives more weight to larger discrepancies, it is encouraging that these metrics were not as different as would be expected had there been larger discrepancies between the actual and predicted scores

Figure 1 shows the relationship between the fixed TNT scores and the adaptive scores derived from the trained SDCA model. Visually, this relationship between the scores appears to be rather strong. Statistically, the correlation indicates that there is formal alignment between the scores, r_s = .92, p < .001 (). This means that either score can theoretically be used interchangeably with the other score.

Table 1. Overall Descriptive Statistics

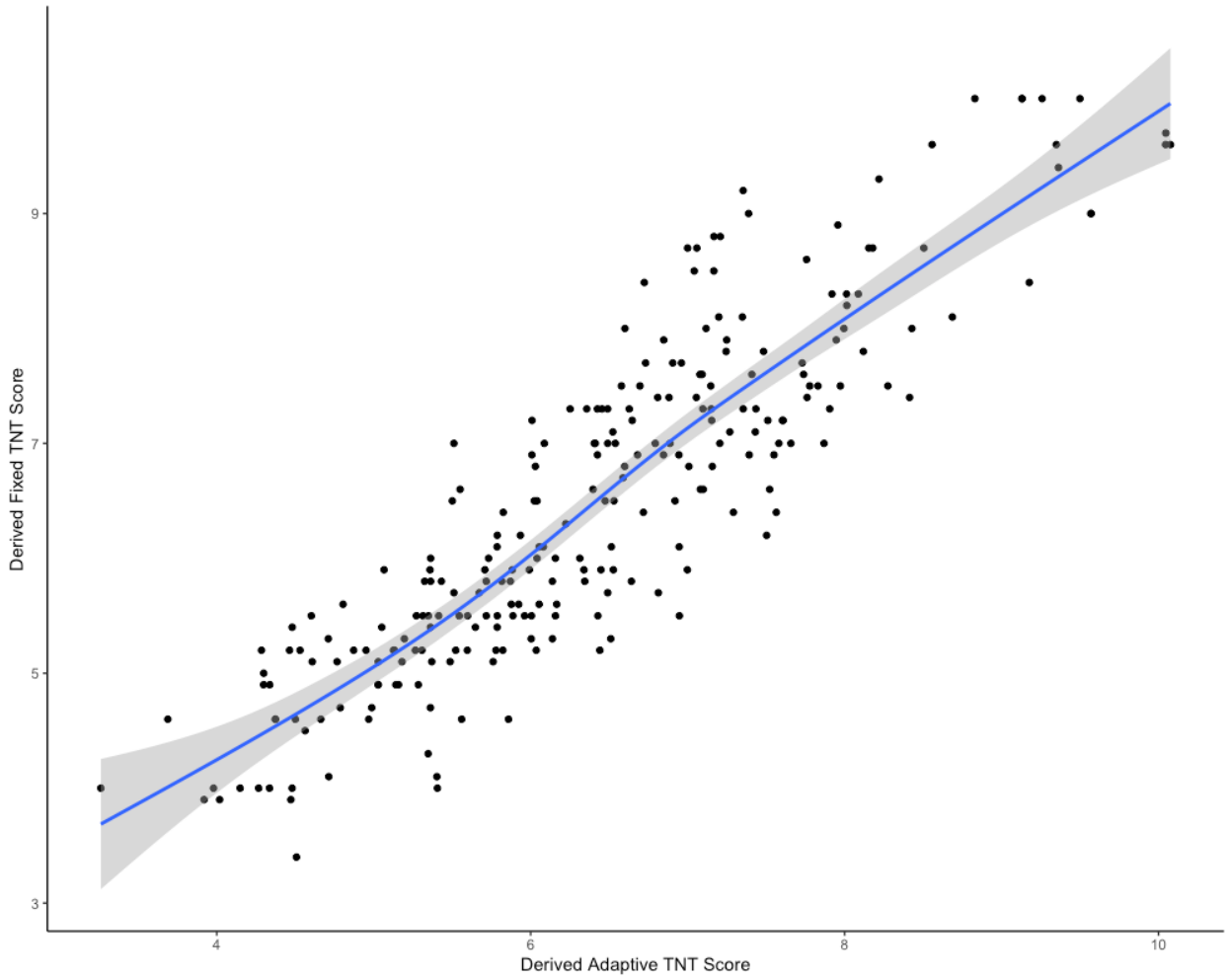
Overall			
	N	Mean (*SD*)	Median
Days between fixed TNT and adaptive TNT	260	1.00 (2.51)	.10
Fixed TNT Score	260	6.55 (1.55)	6.45
Adaptive TNT Ability Estimate	260	.35 (.99)	.29
Adaptive TNT Standard Error of Ability Estimate	260	.21 (.06)	.19
Number of Adaptive TNT Items Administered	260	12.47 (1.83)	12.00
Adaptive TNT Audio Errors	260	.15 (.60)	.00
University in United States			
	n	Mean (*SD*)	Median
Days between fixed TNT and adaptive TNT	133	.76 (.44)	.92
Fixed TNT Score	133	6.24 (1.30)	5.90
Adaptive TNT Ability Estimate	133	.21 (.84)	.04
Adaptive TNT Standard Error of Ability Estimate	133	.19 (.05)	.18
Number of Adaptive TNT Items Administered	133	12.26 (1.38)	12.00
Adaptive TNT Audio Errors	133	.15 (.58)	.00
Secondary Schools in New Zealand			
	n	Mean (*SD*)	Median
Days between fixed TNT and adaptive TNT	68	.82 (1.67)	.02
Fixed TNT Score	68	7.56 (1.55)	7.50
Adaptive TNT Ability Estimate	68	.91 (.96)	.83
Adaptive TNT Standard Error of Ability Estimate	68	.24 (.06)	.21
Number of Adaptive TNT Items Administered	68	13.04 (2.63)	12.00
Adaptive TNT Audio Errors	68	.12 (.56)	.00

Nonprofit Refugee Organization

	n	Mean (*SD*)	Median
Days between fixed TNT and adaptive TNT	21	.01 (.01)	.01
Fixed TNT Score	21	7.04 (1.44)	7.00
Adaptive TNT Ability Estimate	21	.40 (.91)	.34
Adaptive TNT Standard Error of Ability Estimate	21	.21 (.04)	.19
Number of Adaptive TNT Items Administered	21	12.38 (1.75)	12.00
Adaptive TNT Audio Errors	21	.05 (.22)	.00

Miscellaneous Examinees

	n	Mean (*SD*)	Median
Days between fixed TNT and adaptive TNT	38	2.70 (5.87)	.02
Fixed TNT Score	38	5.57 (1.37)	5.30
Adaptive TNT Ability Estimate	38	-.16 (1.17)	-.61
Adaptive TNT Standard Error of Ability Estimate	38	.20 (.06)	.17
Number of Adaptive TNT Items Administered	38	12.21 (1.30)	12.00
Adaptive TNT Audio Errors	38	.29 (.84)	.00



Summary

References

Bock, R. D., and Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444. doi: 10.1177/014662168200600405

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.

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

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.

Habing, Grego, & Vessilinov (2020).

Mayne, Z., Hart, J., & Burdis, J. (2020). The viability of using computerized adaptive testing to measure English-speaking ability. Manuscript submitted for publication.

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

Shalev-Shwartz, S., & Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research, 14*(Feb), 567-599.

Tran, M. (2015). Countries with High English Proficiency Are More Innovative. *Harvard Business Review.*

Vinther, T. (2002). Elicited imitation: A brief overview. *International journal of applied linguistics*, 12(1), 54-73.