TRUENORTH

# Technical Report

English TrueNorth Test (TNT) Form D

September 24, 2020

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

## Introduction

The purpose of this white paper is to highlight key aspects of technical work behind the update of the TrueNorth English Speaking Test (TNT) from Version C.19.03 to Version D.20.09. This was a significant update in test form. Earlier forms of the TNT 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, and Superior (Mastery) levels. Thus, regardless of ability, each test taker would be presented with a battery of items across the full difficulty range in a set order.

These listen-and-repeat items are scored using an automated speech recognition (ASR) engine. The scoring of the items informs the final calculation of an ability estimate. This ability estimate drives the selection of tasks for Part 2 - Open Response. These early semi-adaptive versions of the TNT serve well their intended purpose of providing a fast, reliable, accurate, and scalable measure of speaking ability. This version has been used to measure the speaking ability of thousands of individuals around the world.

However, from the outset of the TNT's development, its creators have had a vision of creating an automated and fully adaptive version of the test. This will optimize the unique characteristics of elicited imitation (i.e., listen-and-repeat items) that allow for immediate automatic scoring. In an adaptive version of the test, 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. Thus, instead of inefficiently presenting a rigid fixed battery of items, a test form is dynamically generated to custom fit the test taker's ability profile as it emerges.

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theorized for decades, the concept was robustly proofed internally before development began (Mayne, Hart, & Burdis, *unpublished*).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 discriminates across the full range of the TNT scale. This early research also provided compelling evidence that the adaptive version of the test provided far more appropriate and desirable levels of item exposure and new barriers to test fraud behaviors.

Feasibility research also showed that there were significant efficiencies gained by the implementation of full adaptivity. The total items that must be presented dropped from the 30 in Version C.19.03 to well below 10 in most simulations. Modeling also showed that the number of items providing the most information about a test taker's ability increased in an adaptive test form. Thus, the shortened test retained and even gained reliability and precision in discriminating ability.

This technical report will present the evidence that despite the key changes in its form that result in an item by item fully adaptive experience, TrueNorth English Speaking test Version D.20.09 can be used with confidence in place of TrueNorth English Speaking test Version C.19.03. It will also show that TrueNorth English Speaking test Version D.20.09 achieves significant improvements in test efficiency.

## 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 in the United States 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 both instruments exceeded 30 days were excluded from the development of the scoring algorithm.

*Instruments*

The fixed TNT was used as the standard for validating the adaptive TNT. As previously mentioned, it comprised 30 elicited imitation items presented in order from least to most difficult. Each item presents the recording of a native English speaker speaking a target sentence; after listening, the test taker repeats and records what they heard as completely and accurately as their ability allows. The validity and reliability of this assessment had been established in previous validation studies (Emmersion, 2019; Habing, Grego, & Vessilinov, 2020).

A fully functional prototype of the adaptive TNT was used to examine the feasibility of Version D.20.09 being used in real-world settings. Key features of this prototype included:

- *Item bank:* An item bank of 153 EI items that had been previously calibrated. Item difficulty ranged from low difficulty to high difficulty.

- *Start condition:* The first item presented to a test taker was randomly selected from those 10 items that were 4 standard deviations from the mean in terms of difficulty.
- *Next item selection:* the prototypes selection criterion algorithm was based on maximized Fisher information (MFI), which computed the amount of information provided by the items 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.
- *Calculation of ability:* provisional and final estimates of ability and their corresponding standard errors were computed via expected a posteriori (EAP; Bock & Mislevy's, 1982).
- *Test length:* to maintain face validity, a minimum of 12 items was administered to all examinees. To maintain efficiency relative to the fixed TNT, the maximum number of items administered to examinees was set to 20.
- *Exit condition:* all assessments ended when either at least 12 items were administered and the standard error was below .316 or when a total of 20 items had been administered.

*Procedure*

All examinees took both the fixed form TNT (Version C.19.03) and the adaptive TNT prototype. 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). Following the listen and repeat section of the fixed form TNT (Version C.19.03), participants completed Part 2 Open Response. However, data from that section of the test were not included in analysis of this paper.

*Data Analyses*

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

Next, participants' EAP ability estimates (thetas) and standard errors from the adaptive TNT were paired with their corresponding scores from the 0-10 scale from the fixed TNT. Seventy percent of these scores were randomly sampled to generate a dataset for training a stochastic dual coordinate ascent (SDCA) model (Shalev-Schwartz & Zhang, 2013). The remaining 30% of the data were used as a testing dataset to

evaluate the performance of the model. Because of a restriction in the range of ability, data for 20 respondents representing true beginners were simulated. Ten of these simulated scores were appended to the training dataset, and the other ten simulated scores were appended to the testing dataset.

Finally, model metrics were examined to evaluate the performance of the train SDCA model. This presented insights into 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 TNT and the fixed TNT, the reported score from the fixed TNT, the ability estimate from the adaptive TNT, the standard error of the ability estimate from the adaptive TNT, the number of adaptive TNT items administered, and the number of audio errors.

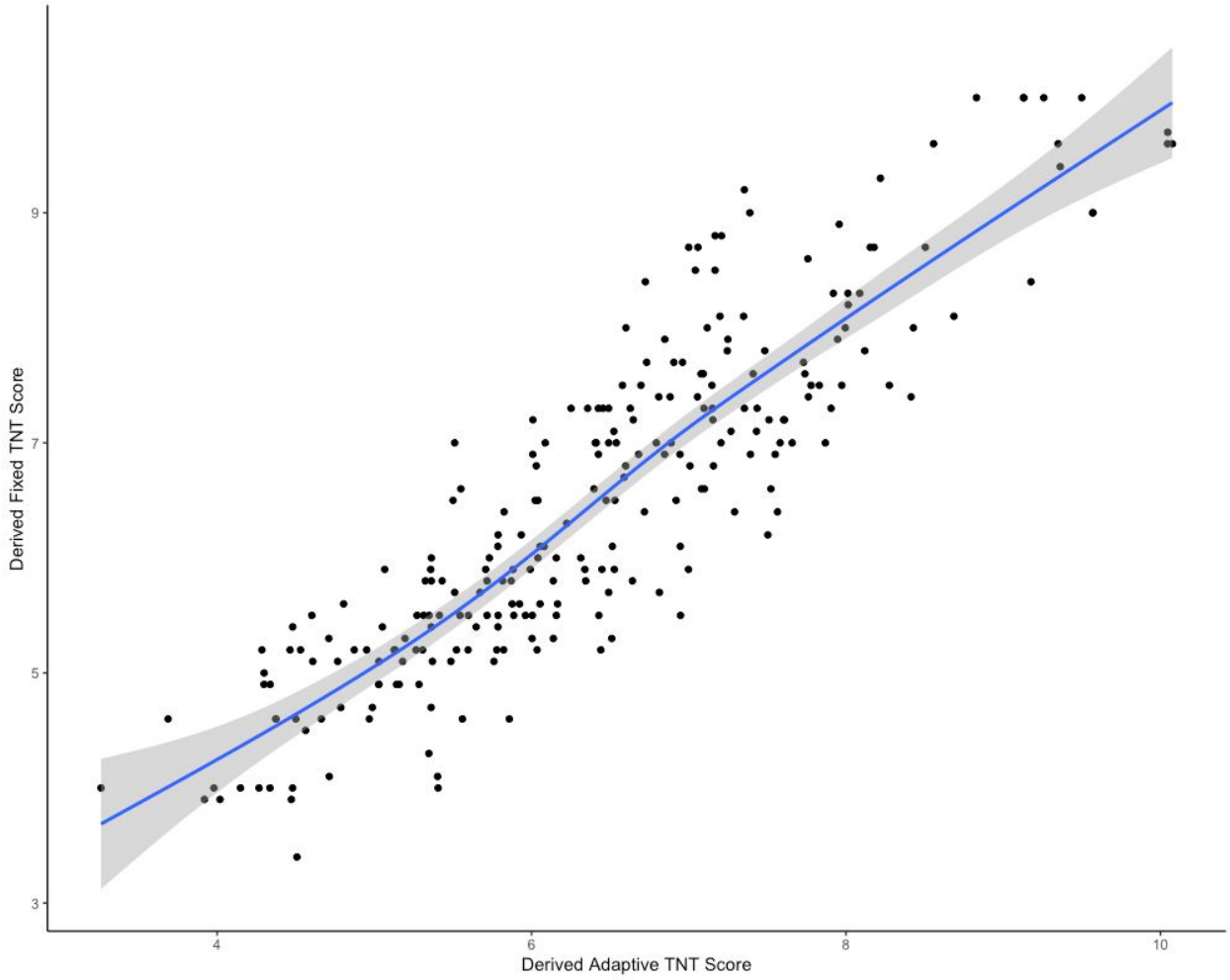
As expected, speaking ability as measured by the adaptive TNT was positively correlated with the number of items, $r_s = .46$, $p = .000$. Because of the lack of more difficult items relative to easier items, more items were required to measure the language ability of the more capable examinees. Interestingly, speaking ability as measured by the adaptive TNT 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 TNT 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

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 |

Figure 1. Scatterplot showing relationship between adaptive TNT scores and fixed TNT scores




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, $r_s = .92$, $p < .001$ (Dorans & Walker, 2007). This means that either score can theoretically be used interchangeably with the other score.

## Summary

This technical report provides evidence in support of the adaptive form of the English TNT as a valid replacement of the fixed form (Version C.19.03 ) of the English TNT. Most notably, the correlation between the scores derived from the two assessments reached a level to where a high degree of confidence can be ascribed to the adaptive

TNT. Stakeholders can be assured that the adaptive form of the English TNT will produce scores similar to the fixed form of the English TNT. Moreover, reliable scores can be generated in fewer than half of the number of items administered in the fixed form of the TNT for the vast majority of examinees.

## 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

Dorans, N. J., & Walker, M. E. (2007). Sizing Up Linkages. *Linking and Aligning Scores and Scales Statistics for Social and Behavioral Sciences*, 179–198. https://doi.org/10.1007/978-0-387-49771-6_10

Emmersion (2019). *The development of an automated online English-speaking proficiency assessment*.

Habing, Grego, & Vessilinov (2020). *Predictive and psychometric properties of the TrueNorth Test (TNT)*.

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

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