Predictive and Psychometric Properties of the TrueNorth Test (TNT)

FINAL REPORT

RESEARCH TEAM

BRIAN HABING, PhD

University of South Carolina

JOHN GREGO, PhD

University of South Carolina

ROUMEN VESSELINOV[1], PhD

City University of New York

This report represents the individual opinion of the authors and not necessarily of their institutions.

May 1, 2020

[1] Corresponding author: roumen.vesselinov@qc.cuny.edu .

EXECUTIVE SUMMARY

The study sample consists of 128 English language test takers at Brigham Young University in 2019/20. All participants were non-native speakers of English and they took two tests (in random order) of English oral proficiency: ACTFL OPIc and TrueNorth (TNT) test.

The participants were 50.8% female with an average age of 24 years and varying from 17 to 49 years old. Almost all participants had some college experience but did not have a college degree. The test takers were native speakers of: Spanish (n=55), Japanese (28), Chinese (16), Portuguese (11), with a small number of Arabic, Central Khmer, Creole, French, Korean, Russian, Thai and Turkmen native speakers.

This study evaluated the reliability and consistency of TNT and the relationship between ACTFL OPI and ACTFL TNT estimates.

MAIN RESULTS

- TNT is a very reliable, moderately strong scale.
- TNT is predominantly measuring one underlying latent trait.
- TNT can be used for language proficiency evaluations.
- The TNT estimate of ACTFL levels has a relatively good correlation with the OPI estimates.
- The TNT estimate of ACTFL should not be used as a substitute for ACTFL OPI in one test evaluations.
- TNT can be improved by reevaluating some less-than-perfect scale items.

Possible Extensions of the Main Results:

- This study evaluated TNT for English; we could hypothesize that this reliability and consistency would also apply for other languages.
- TNT score has 100 possible sublevels (0.0-10.0) and theoretically it can detect very small improvement in language proficiency. We could hypothesize that TNT can be used for language proficiency evaluations even for shorter than the 8-week period required for ACTFL.
- Since TNT is a very reliable scale, it can be hypothesized that ACTFL TNT estimates could be suitable for test-retest studies to measure the gain in language proficiency after a study period.

For more definitive answers to those possible extensions, more research is needed.

Table of Contents

Introduction

The Research Team received the testing results of 128 test takers at Brigham Young University in 2019/20, provided by Emmersion Learning, Inc[2]. All participants were non-native speakers of English and they took two tests (in random order) of English oral proficiency: ACTFL OPIc[3] and TrueNorth (TNT) test[4]. TNT is a newly developed oral proficiency test based on elicited imitation as a testing method in which participants hear an utterance in the target language and are prompted to repeat the utterance as accurately as possible. TNT provides an oral proficiency score between 0.0 and 10.0, with 10 being the highest. In addition, TNT provides ACTFL TNT estimates, and TNT estimates of CEFR[5] and TOEFL[6]. The TNT CEFR and TOEFL estimates were not available for this study.

The sample of 128 people was 50.8% female with an average age of 24 years (std=6.3) and varying from 17 to 49 years old. Almost all participants had some college experience but did not have a college degree. The native language of the test-takers was: Spanish (n=55), Japanese (28), Chinese (16), Portuguese (11), with a small number of Arabic, Central Khmer, Creole, French, Korean, Russian, Thai and Turkmen native speakers.

This study evaluates the reliability and consistency of TNT and the relationship between ACTFL OPI and ACTFL TNT estimates.

1. Reliability and Consistency of TNT

The high reliability and inter-rater consistency of ACTFL OPI is well-documented in the literature (e.g. Surface & Dierdorff, 2003).

We evaluated the reliability of TNT with data from this study. The TNT scale has 30 items/elements which can be represented as raw item scores (0-100) and as polytomous scores (0-3). The scores are rated automatically by TNT software but for this study they were also rated by hand by linguistic experts.

[2] Emmersion Learning, Inc. provided the data and funding for this study, but the analysis and the work of the Research Team were conducted independently.

[3] https://www.languagetesting.com/oral-proficiency-interview-by-computer-opic

[4] https://truenorthtest.com/

[5] https://www.coe.int/en/web/common-european-framework-reference-languages

[6] https://www.ets.org/toefl

The TNT scale reliability was very high with Cronbach’s Alpha of 0.932 for automatic item rating and 0.944 for hand-rating. Both were calculated using the median of ten runs of imputation to fill missing data using simulation from a fitted Generalized Partial Credit (GPC) model. There was one outlier case #120, with TNT automatic score of 5.9 and a TNT hand-rated score of 9.4. We repeated the analysis with this case excluded and the results were unchanged. This outlier is not an influential observation and it was retained in the sample.

Mokken Scaling for Measuring the Quality of TNT scale

While coefficient alpha gives a measure of reliability (in the sense of measurement error), it does not necessarily provide information about whether a set of items is unidimensional (Hattie, 1985). Mokken scaling coefficients (e.g. Sijtsma and Molennar, 2002) provide a measure of whether the set of items can be meaningfully thought of as measuring a single trait. (Many other widely used item response theory methods for assessing multidimensionality require a larger number of subjects than are available here). Item and test level scalability coefficients were estimated using the median of the ten imputed data sets. The TNT (Automatic) scale is a very reliable (Alpha=0.932), moderately strong (H=0.403) scale.

Table 1. Mokken Scaling for TNT (Automatic)

Item	H
1	0.413
2	0.244
3	0.436
4	0.403
5	0.378
6	0.327
7	0.436
8	0.350
9	0.341
10	0.459
11	0.339
12	0.432
13	0.471
14	0.316
15	0.455
16	0.431
17	0.368
18	0.415
19	0.400
20	0.371
21	0.391
22	0.402
23	0.504
24	0.463
25	0.452
26	0.447
27	0.378
28	0.443
29	0.412
30	0.413

Item 2 had an Hi below the threshold of 0.3, and items 5, 6, 8, 9, 11, 14, 17, 20, 21, 27 were below 0.4. If we remove those items, the resulting 19-item scale has approximately the same reliability (Alpha=0.932) and higher H (0.478) with the lowest item-H being 0.400.

Table 2. Mokken Scaling for TNT (Automatic) with Reduced Number of Items (n=19)

Item	H		Item	H
1	0.414		19	0.416
3	0.466		20	0.408
4	0.464		22	0.405
7	0.401		23	0.444
10	0.469		24	0.520
12	0.493		25	0.474
13	0.464		26	0.461
15	0.496		28	0.459
16	0.498		29	0.467
18	0.476		30	0.435

The TNT (Hand-rated) scale is a very reliable (Alpha=0.944), moderately strong scale (H=0.462).

Table 3. Mokken Scaling for TNT (Hand-rated)

Item	H	Item	H
1	0.405	16	0.523
2	0.378	17	0.454
3	0.417	18	0.509
4	0.488	19	0.457
5	0.361	20	0.410
6	0.360	21	0.373
7	0.424	22	0.437
8	0.508	23	0.519
9	0.513	24	0.536
10	0.475	25	0.538
11	0.501	26	0.470
12	0.491	27	0.362
13	0.496	28	0.451
14	0.471	29	0.499
15	0.519	30	0.425

Items 2, 5, 6, 21, and 27 were below the threshold of 0.4. If we remove those weaker items, the resulting 25-item scale gives a high reliability (Alpha=0.944) and higher H (0.502) with the lowest item-H being 0.426.

2. Comparison of ACTFL TNT and ACTFL OPI

ACTFL OPI has 10 proficiency levels:

Table 4. ACTFL OPI Levels

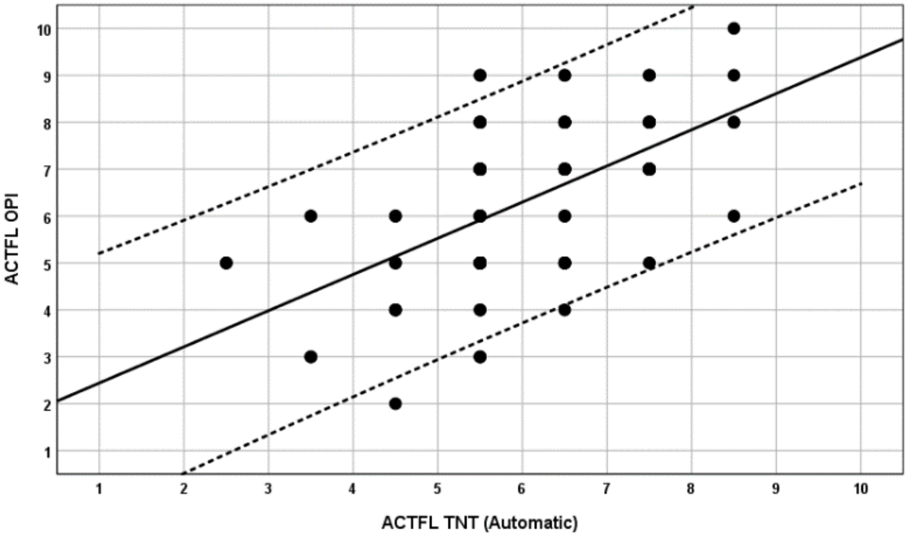
ACTFL	Definition
1	Novice Low
2	Novice Mid
3	Novice High
4	Intermediate Low
5	Intermediate Mid
6	Intermediate High
7	Advanced Low
8	Advanced Mid
9	Advanced High
10	Superior

ACTFL TNT provides interval estimates like Novice Low – Novice Mid (1-2), Novice Mid – Novice High (2-3), etc.

Because ACTFL TNT has interval ratings, it cannot be compared directly with the discrete ratings of ACTFL OPI (1-10). It is worth noting that technically, the ACTFL TNT automatic interval estimates contained the ACTFL OPI rating in 57.8% of the cases (74 out of 128). In 10.9% of the cases the TNT estimate was above the OPI rating and in 31.3% it was below.

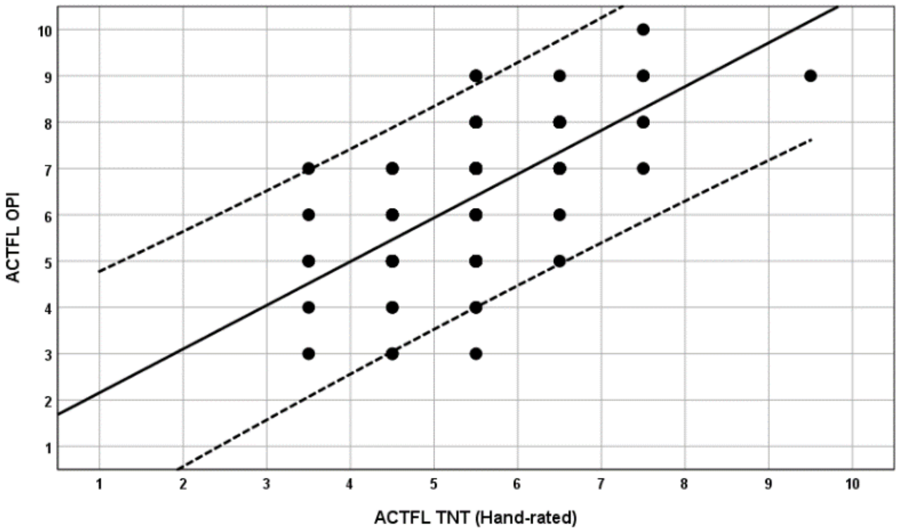
We recoded the interval ACTFL TNT into a discrete rating representing the middle of the interval. For example, ACTFL TNT rating (1-2) was represented with rating of 1.5, rating of (2-3) with 2.5, etc.

Figure 1. Descriptive Graphical Representation (Scatterplot) of ACTFL OPI and ACTFL TNT (Automatic score)



Note. In all scatterplot figures one dot can represent more than one case.
Solid line: OLS regression line.
Dashed lines: 95% Confidence Interval (CI).

Figure 2. Descriptive Graphical Representation (Scatterplot) of ACTFL OPI and ACTFL TNT (Hand-rated)



ACTFL TNT has relatively good correlation with ACTFL OPI of about 0.6 (Spearman) to 0.7 (Gamma). While correlations in this range are considered good in some cases, Dorans and Walker (Ch.10 in Dorans et al., 2007) recommend a minimum correlation of 0.866 before considering the formal aligning of scores. This corresponds to a reduction in uncertainty of 50%. One main difference between the OPI and TNT estimate is that the latter put 50% of the cases in one group (5-6) compared to 27% (level 5) and 12.5% (level 6) for the former.

Figure 3. Distribution of ACTFL TNT (Automatic)

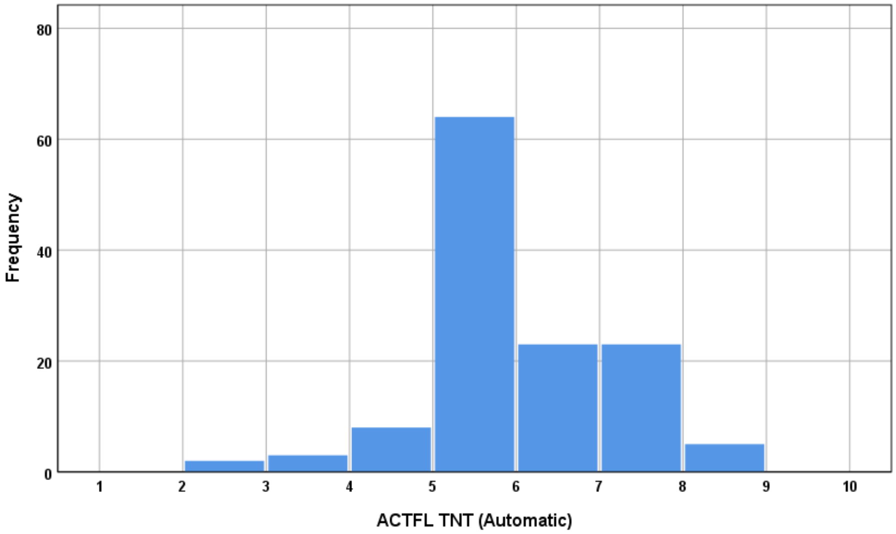


Figure 4. Distribution of ACTFL OPI

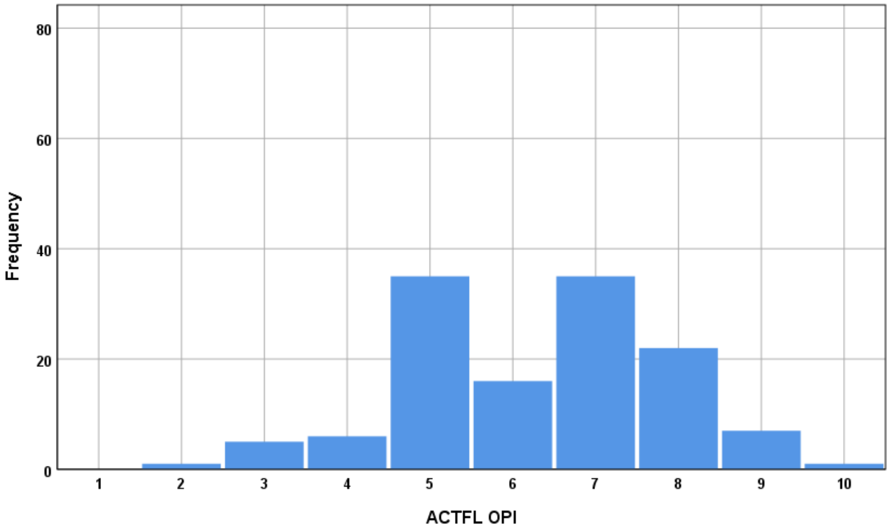
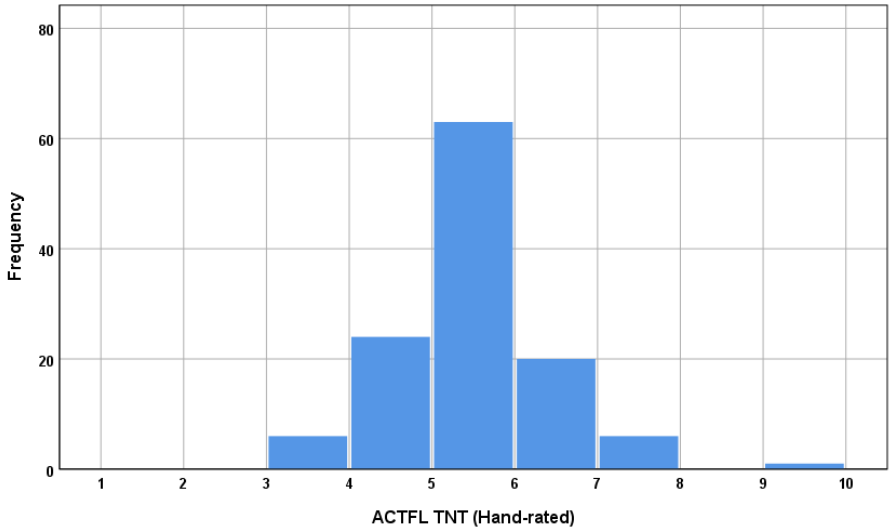


Figure 5. Distribution of ACTFL TNT (hand-rated)



The lower variability of ACTFL TNT leads to low estimates of the measures for agreement (Weighted Kappa=0.3). The hand-rated TNT estimates have slightly better correlation and agreement with ACTFL OPI ratings.

Figure 6. Agreement between ACTFL OPI (OPI) and ACTFL TNT

TNT Automatic (TOPI_ave)

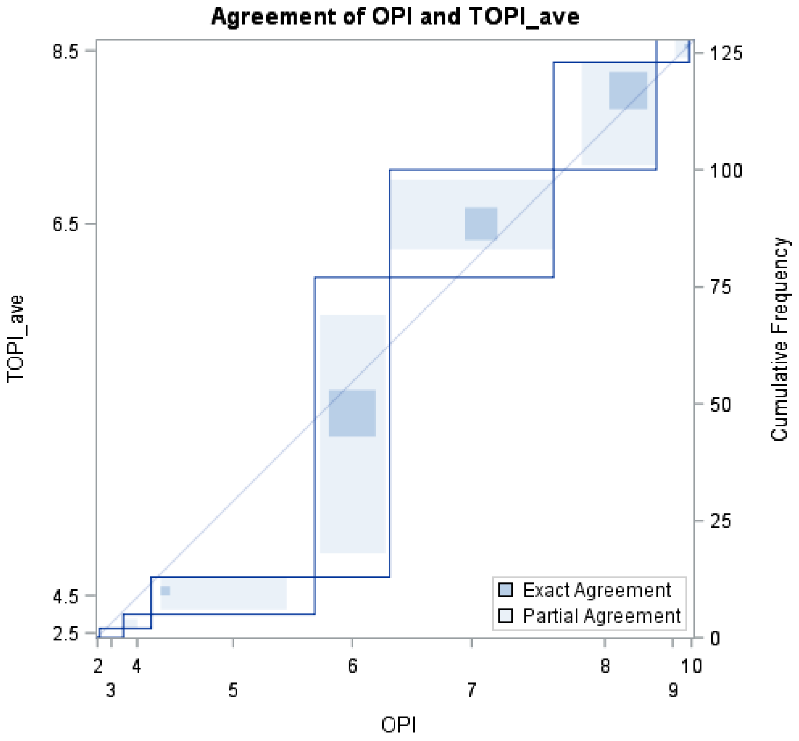
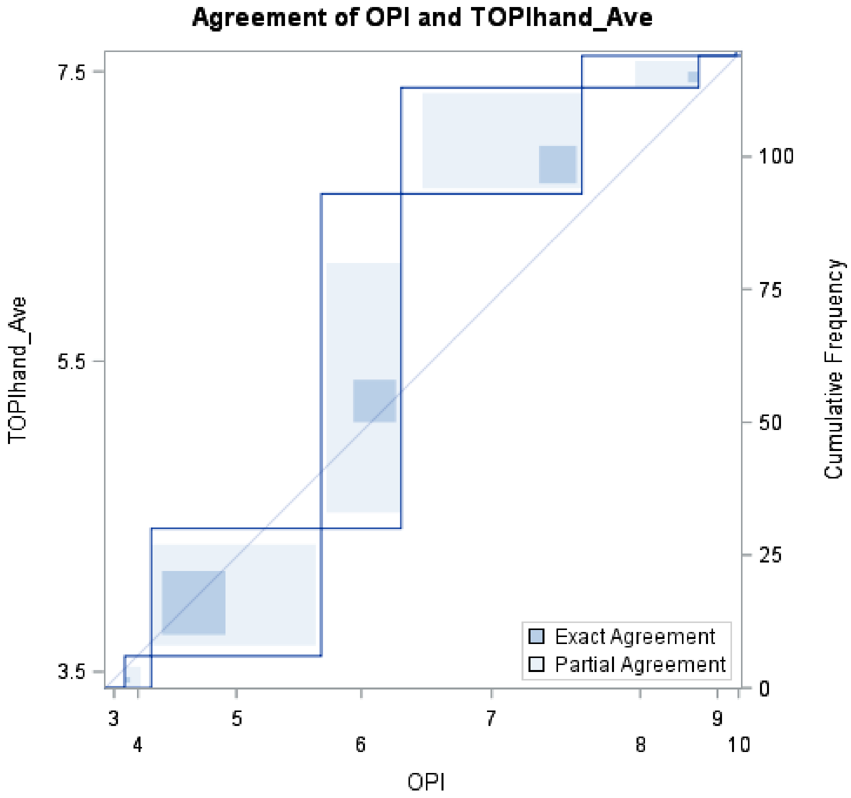


Figure 7. Agreement between ACTFL OPI (OPI) and ACTFL TNT

TNT Hand-rated (TOPIhand_Ave)



3. Comparison of TNT Numerical Score and ACTFL OPI Score.

Figure 8. Descriptive Graphical Representation (Scatterplot) of ACTFL OPI and TNT score (Automatic)

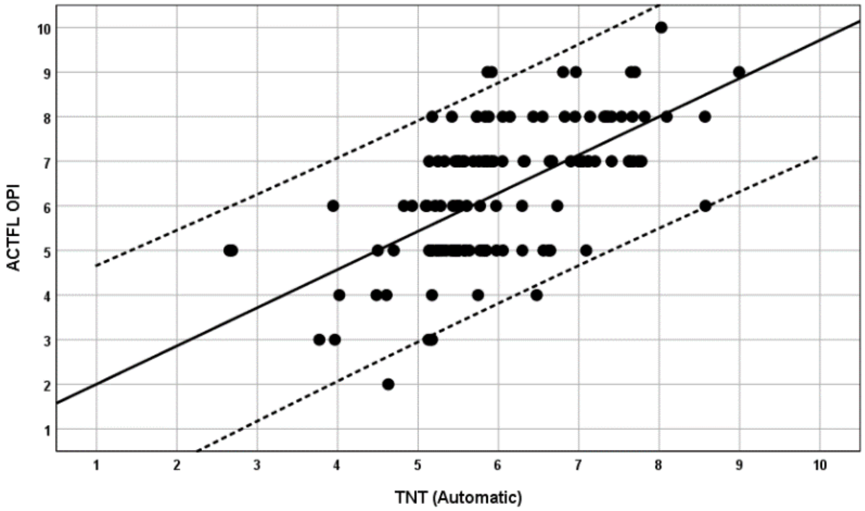
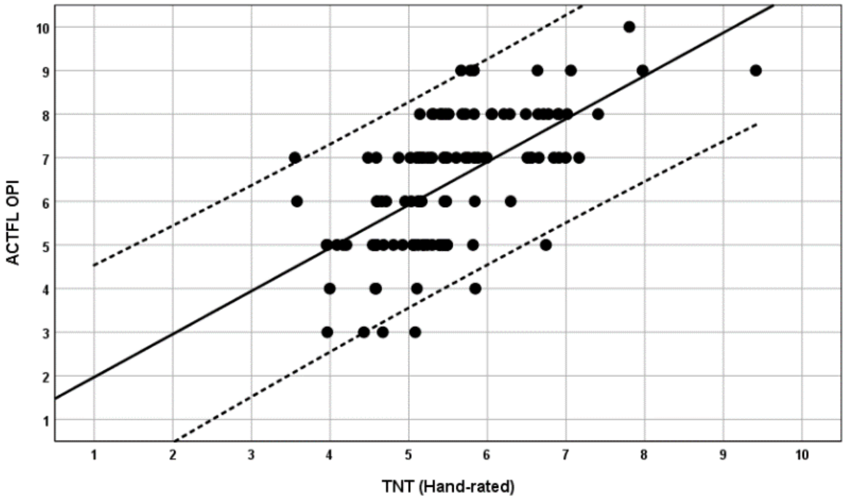


Figure 9. Descriptive Graphical Representation (Scatterplot) of ACTFL OPI and TNT score (Hand-rated)



There is a relatively good correlation of 0.63 (Spearman) between ACTFL OPI rating and TNT score (automatic or hand-rated). The correlation between the automatic and hand-rated scores of TNT is very strong (Spearman=0.82). Both correlations are below the suggested threshold of 0.866.

4. IRT Model for TNT and Comparison of the Theta Levels with ACTFL OPI

Both a generalized partial credit model (Muraki, 1992) and Masters (Rasch) partial credit model (Masters, 1982) were fit to the TNT (both automatic scoring and hand scoring). As we discuss later (Section 9) the various fit statistics were very similar between the two models and the correlation between the estimated latent trait values was greater than 0.99. There is some evidence a few items may differ enough from the Rasch assumption to merit further investigation.

The correlation between the IRT Thetas (automatic score) and ACTFL OPI was relatively good (Spearman=0.63) and the correlation for Theta (Hand-rated score) was about the same (Spearman=0.62). Both correlations are below the suggested 0.866 threshold.

When we adjust for attenuation (using Pearson correlation), assuming ACTFL OPI is as reliable as ACTFL TNT (Hand-rated), we can get an estimate for the correlation on the underlying latent traits. The new correlation is slightly higher (0.66). This estimate may be biased downward due to the discrete nature of OPI, but not by much. The maximum correlation between true ACTFL OPI and IRT Theta could be as high as 0.955. Scaling the two will bring the correlation to 0.7. Even this would not be high enough to treat the two traits being measured as the same. The true subject ACTFL TNT would be estimated to only explain about 50% (=0.7^2) of the variability in the true subject OPI.

If the actual reliability of ACTFL OPI is much lower, the underlying correlation could be higher. In this study we do not have the data to measure the reliability of ACTFL OPI.

Figure 10. Descriptive Graphical Representation (Scatterplot) of ACTFL OPI and IRT Theta (TNT Automatic)

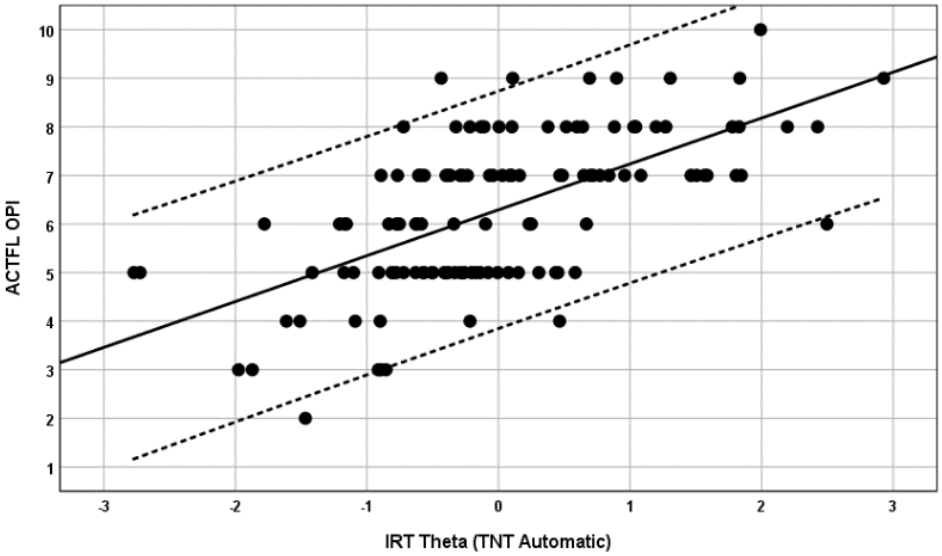
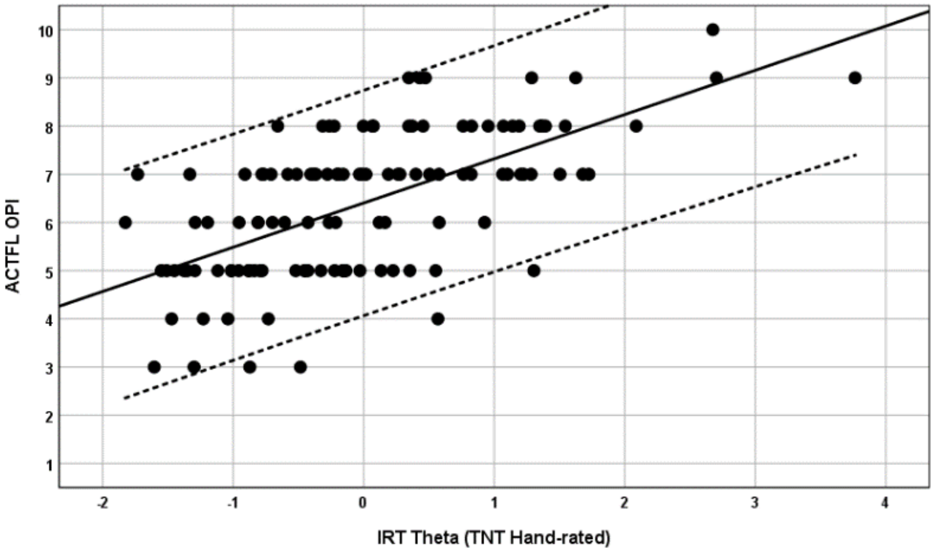


Figure 11. Descriptive Graphical Representation (Scatterplot) of ACTFL OPI and IRT Theta (TNT Hand-rated)



5. Comparison of TNT Hand-ratings and Automatic Ratings

We performed a multiple imputation (SAS PROC MI, n=5) procedure for the data for raw hand-rated (human) scores (H1-H30) and raw automatic (speech recognition) scores (X1-X30) because of missing data. For each set of scores, maximum likelihood estimates of a normal multivariate mean vector and covariance matrix were computed in PROC MI in SAS. Data were then imputed from a truncated multivariate normal distribution in R to ensure scores were bounded between 0 and 100.

Table 5. Correlation between the raw automatic and hand-rated scores

Spearman

Item	Correlation	SE
1	**0.307**	0.086
2	**0.285**	0.088
3	**0.106**	0.089
4	0.543	0.078
5	0.490	0.081
6	0.482	0.078
7	0.667	0.069
8	**0.160**	0.121
9	0.459	0.080
10	0.658	0.069
11	0.552	0.076
12	0.713	0.066
13	0.614	0.071
14	0.582	0.073
15	0.706	0.066
16	0.715	0.063
17	0.651	0.072
18	0.745	0.062
19	0.618	0.081
20	0.675	0.085
21	0.610	0.076
22	0.625	0.091
23	0.656	0.071
24	0.585	0.078
25	0.734	0.069
26	0.510	0.090
27	0.679	0.067
28	0.671	0.079
29	0.678	0.076
30	0.667	0.080

SE=Standard Error

All correlations are below 0.75. Items #1, 2, 3 and 8 have very low correlation between the automatic and hand-rated raw scores.

We also computed canonical correlations using one of the 5 imputed data sets, but the results were inconclusive. The first canonical variable suggested that a contrast of TNT 1 versus (TNT 6, 13, and 22) was correlated with a linear combination of (TNT Hand-rated 9, 12, and 25). The second canonical variable places emphasis on item 4, with a contrast of TNT 4 versus (TNT 16 and 29) strongly correlated with a contrast of TNT Hand-rated 4 versus (TNT Hand-rated 19 and 29).

Figure 12. Canonical Correlation for TNT (Automatic)

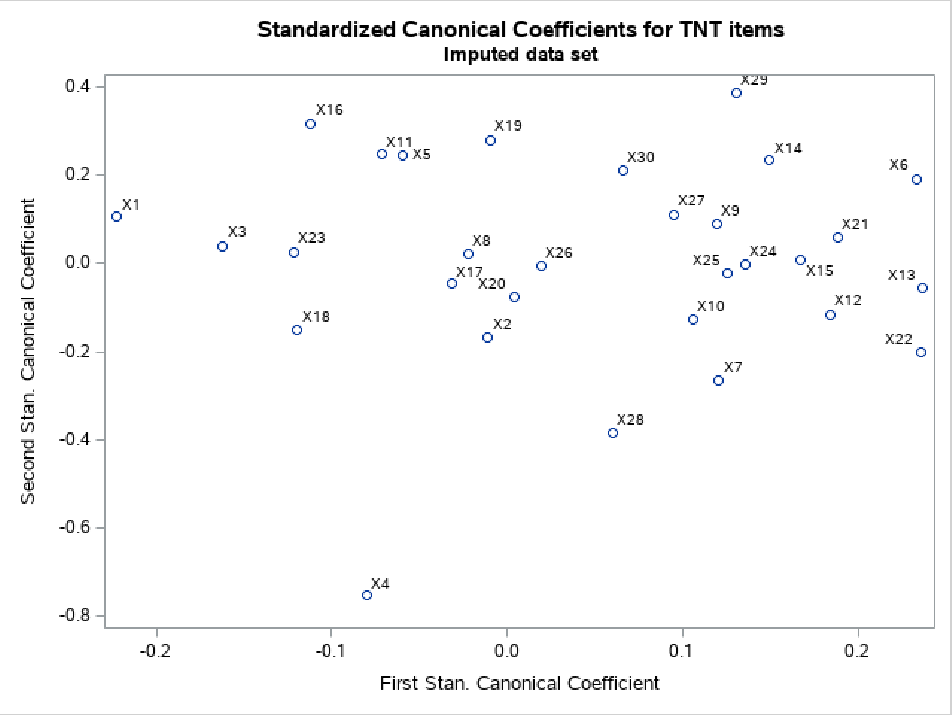
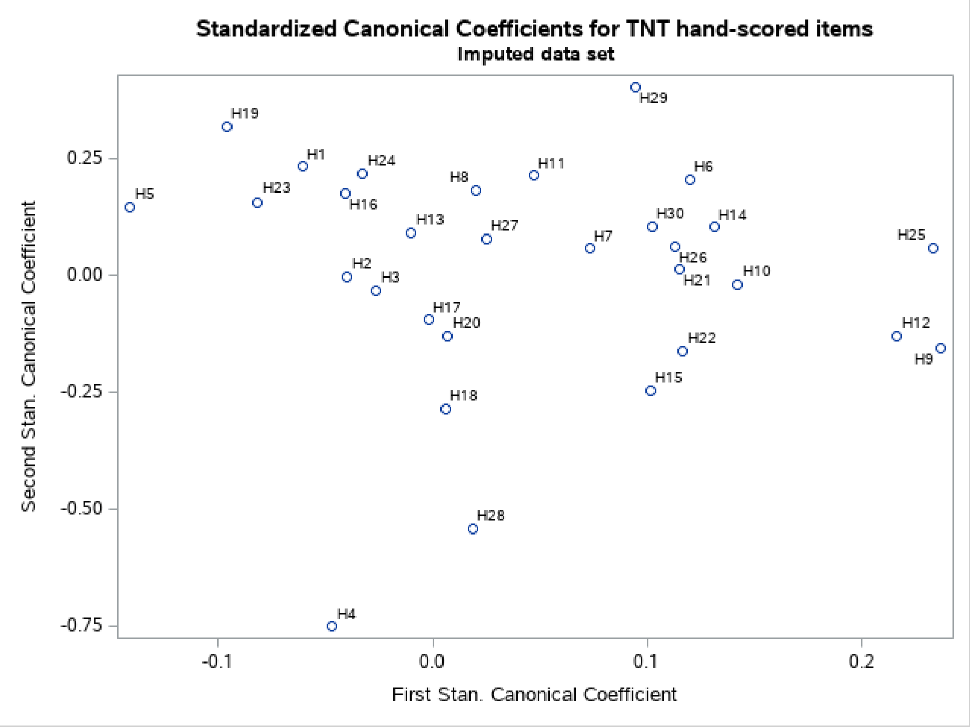
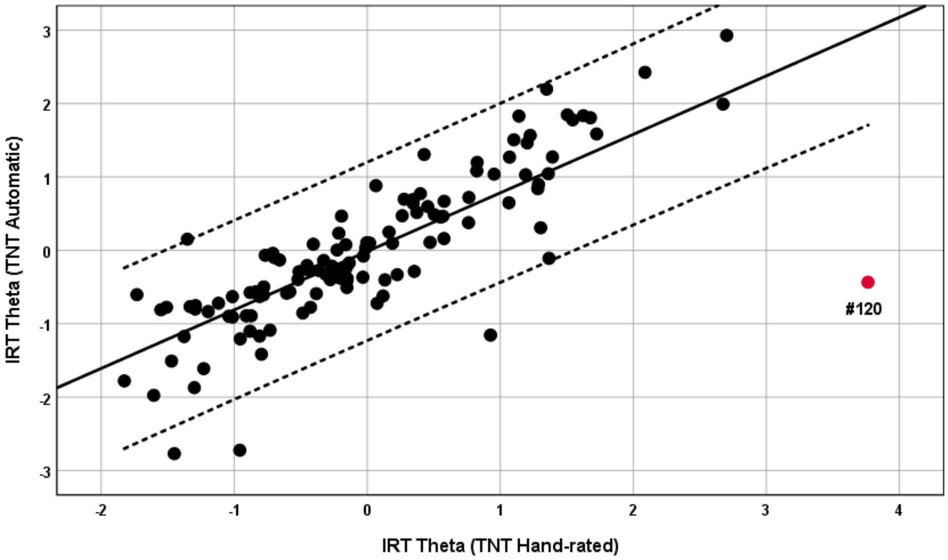


Figure 13. Canonical Correlation for TNT (Hand-rated)



6. Comparison of IRT Thetas for Automatic and Hand-Rating

Figure 14. Descriptive Graphical Representation (Scatterplot) of IRT Theta Automatic and Theta Hand-Rated



The biggest outlier (Theta Hand-Rated=3.76, Theta Automatic=-0.44) is for a participant with ACTFL OPI rating of 9 (Advanced High), ACTFL TNT (Automatic) rating of (5-6), and ACTFL Hand-Rated score of (9-10). This was a case of failure of the automatic TNT scoring. We will call it case #120 for reference purposes.

The correlation between the estimated Thetas (Automatic with Hand-Rated) is very high (0.8). But this is still below the recommended threshold of 0.866 for treating them for aligning the scores for any interchangeable uses.

Application of Mokken Scaling

We examined the item responses of the two scoring methods (Automatic and Hand-Rated) by calculating a version of the scalability coefficient for the item pairs across the two methods of scoring (the observed covariance between item i on each form divided by the maximum possible covariance based on the distribution of responses for that item). This is what the item pair scalability coefficient would be if the two forms were treated as a single exam. We also extracted the a-parameter (discrimination) from the estimated generalized partial credit model fit separately to the two versions.

Table 6. Mokken Scale for Comparison

Item	Hii	Discrimination Automatic	Hi Automatic	Discrimination Hand-Rated	Hi Hand-Rated
1	**0.36**	1.06	0.410	1.26	0.407
2	**0.42**	0.48	0.242	1.08	0.377
3	**0.21**	1.01	0.436	1.48	0.418
4	0.70	1.17	0.404	1.86	0.488
5	0.53	0.95	0.378	0.99	0.362
6	0.78	0.92	0.334	0.91	0.361
7	0.69	1.31	0.434	1.14	0.425
8	**0.35**	0.84	0.348	1.54	0.517
9	0.56	0.95	0.340	2.06	0.512
10	0.74	1.82	0.462	1.76	0.474
11	0.56	0.71	0.345	1.74	0.503
12	0.78	1.28	0.436	1.73	0.492
13	0.69	1.64	0.472	1.59	0.498
14	0.74	0.68	0.316	1.62	0.471
15	0.78	1.24	0.457	1.76	0.519
16	0.73	1.37	0.435	1.89	0.522
17	0.70	0.79	0.366	1.26	0.455
18	0.74	1.13	0.418	1.81	0.510
19	0.77	1.03	0.400	1.35	0.456
20	0.68	0.93	0.374	1.01	0.412
21	0.75	0.88	0.390	0.94	0.376
22	0.57	1.08	0.413	1.30	0.436
23	0.72	1.84	0.508	1.76	0.518
24	0.67	1.28	0.462	2.09	0.541
25	0.84	1.47	0.446	1.72	0.540
26	0.56	1.50	0.446	1.67	0.468
27	0.74	0.90	0.380	0.86	0.363
28	0.72	1.21	0.443	1.60	0.452
29	0.89	1.24	0.412	1.59	0.501
30	0.68	1.28	0.416	1.32	0.426

Items 1, 2, 3, and 8 have particularly low Hii values (keeping in mind that they should be the same item, just with different scorers) if treated as a single exam. Items 2 and 8 have very different item scalability coefficients and estimated discrimination parameters when comparing between the two versions (Automatic and Hand-rated). Examining the score distributions for the items makes it clear that some of the items have very different response patterns and uses of the rating scale across the two versions. The difference in the use of the 0-3 scale of the automatic and hand-rated scoring likely needs additional study.

Table 7. TNT Scale Item Distribution

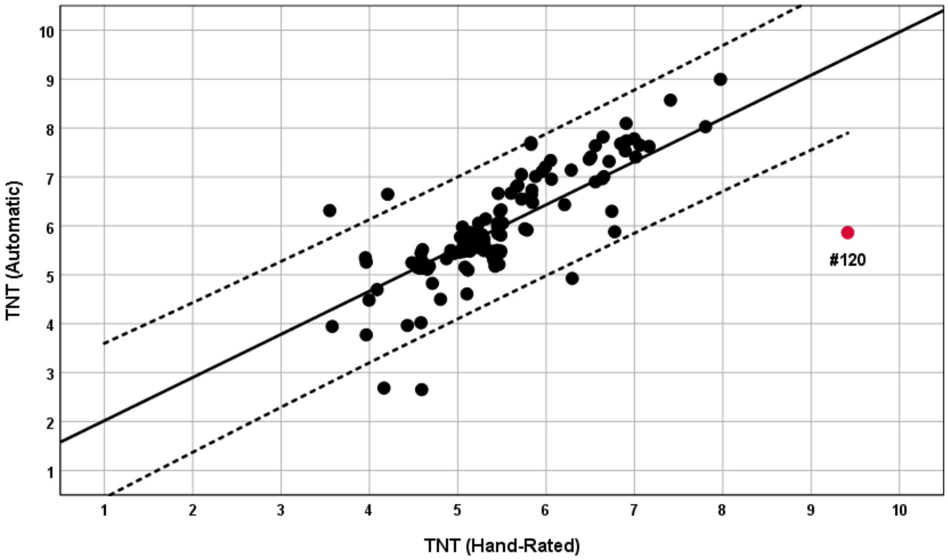
Item	TNT Automatic				TNT Hand-rated			
	A0	A1	A2	A3	H0	H1	H2	H3
1	4	5	18	99	1	1	50	68
2	5	2	**17**	**104**	0	7	**88**	**24**
3	2	2	6	118	0	0	35	85
4	6	10	24	88	4	6	52	58
5	3	13	22	89	2	7	51	60
6	1	7	25	94	3	12	63	42
7	8	26	35	58	2	21	43	54
8	6	11	**35**	**73**	4	6	**13**	**3**
9	1	5	27	94	0	4	89	27
10	3	18	45	62	1	9	61	49
11	12	24	26	62	7	18	52	43
12	9	25	37	55	5	13	73	29
13	4	31	41	49	5	36	56	23
14	9	11	36	70	3	23	69	24
15	16	35	12	58	11	29	45	35
16	8	28	45	44	12	28	58	22
17	12	22	27	63	11	37	52	19
18	12	38	47	28	8	43	54	15
19	18	59	32	8	25	62	28	5
20	16	38	49	18	18	59	35	7
21	16	55	30	20	13	77	25	5
22	17	55	29	8	18	64	27	3
23	23	51	35	15	23	44	43	10
24	10	56	37	21	14	48	41	4
25	18	45	43	11	42	48	27	3
26	17	75	22	2	20	82	17	1
27	11	38	41	32	20	51	42	7
28	10	77	31	6	14	84	21	0
29	45	60	11	0	59	55	6	0
30	5	39	68	13	10	78	28	3

Note. Hand-rated item 8 has only 26 observations.

7. Comparison of TNT Automatic and Hand-rated Scores

The correlation between the Automatic and Hand-rated TNT score is very strong (Spearman=0.82) but still below the suggested threshold of 0.866.

Figure 15. Descriptive Graphical Representation (Scatterplot) of TNT (Automatic) and TNT (Hand-Rated)



After removing the outlier (Case #120) the correlation remains about the same (Spearman=0.83).

8. Comparison of IRT Models for Automatic and Hand-Rated TNT Scores

For this task we computed several of the available fit statistics for the generalized partial credit and Masters (Rasch) partial credit model.

Table 8. Model Comparison

Goodness-of-fit	**IRT Model Automatic Score**		**IRT Model Hand-Rated Score**	
	GPC	**Rasch**	**GPC**	**Rasch**
M2* Statistics	369. 5	404.7	398.1	431.1
M2* p-value	0.18	0.14	0.041	0.035
RMSEA	0.023	0.024	0.033	0.033
LogL	-3339.5	-3367.0	-2897.8	-2922.5

Where,

M2* is asymptotically Chi-square overall Goodness-of-fit (Cai and Hansen, 2013).

RMSEA is the Root Mean Square Error of Approximation.

LogL is the Log Likelihood.

Overall, the two models, GPC and Rasch, have very similar Goodness-of-fit statistics. There is no significant evidence against the model fit in the case of automatic scoring. The M2* test rejects at the α=0.05 level for both models in the hand-scoring case. As Cai and Hansen note, it is necessary to consider the RMSEA before deciding if the misfit is large enough to be concerning.

In addition to the above overall goodness-of-fit measures we applied Orlando and Thissen's generalized S-X^2 item-fit index for polytomous IRT models (in Kang & Chen, 2007).

The low, median, and upper p-values for applying Orlando and Thissen's S-X^2 statistic to the ten multiply imputed data sets for each of the hand automatically scored and hand scored data, for each of the generalized partial credit (GPC) model and Masters partial credit (MPC) model are reported below. There were no major discrepancies in the p-values between the GPC and MPC. At the α=0.05 level on the median p-value with no adjustment for multiple comparisons, item 9 failed to fit for both models with automatic scoring. Item 11 failed to fit for the MPC and was not reassuring (low p-value < 0.05, median < 0.10) for the GPC. Item 19 was not reassuring for the MPC (and was just over those thresholds for the GPC). There were no concerning lacks of fit for the hand scored items. As usual, with any statistical goodness of fit test, there is likely a lack of power for small sample sizes (and the test would be over-powered in terms of practically unimportant misfit if the sample sizes were large).

Table 9. Orlando and Thissen's S-X^2 for TNT (Automatic)

Item	GPC			MPC		
	Low	Median	Upper	Low	Median	Upper
1	0.22	0.53	0.87	0.21	0.53	0.68
2	0.004	0.36	0.74	0.01	0.24	0.60
3	0.13	0.27	0.48	0.12	0.24	0.54
4	0.12	0.31	0.85	0.14	0.34	0.87
5	0.06	0.39	0.63	0.06	0.31	0.46
6	0.38	0.57	0.79	0.50	0.64	0.75
7	0.43	0.73	0.93	0.33	0.63	0.94
8	0.18	0.53	0.72	0.16	0.34	0.55
9	0.003	0.02	0.07	0.00	0.02	0.05
10	0.32	0.88	0.96	0.37	0.65	0.89
11	0.02	0.08	0.30	0.02	0.05	0.38
12	0.04	0.49	0.85	0.07	0.53	0.88
13	0.14	0.48	0.70	0.21	0.64	0.76
14	0.51	0.85	0.95	0.33	0.59	0.70
15	0.23	0.69	0.87	0.23	0.66	0.91
16	0.12	0.36	0.67	0.24	0.52	0.83
17	0.14	0.37	0.72	0.13	0.36	0.49
18	0.21	0.57	0.79	0.20	0.55	0.78
19	0.03	0.11	0.32	0.03	0.10	0.18
20	0.23	0.56	0.72	0.15	0.47	0.73
21	0.12	0.40	0.68	0.12	0.39	0.64
22	0.34	0.54	0.91	0.30	0.54	0.88
23	0.02	0.15	0.42	0.03	0.17	0.44
24	0.28	0.54	0.73	0.32	0.52	0.83
25	0.02	0.17	0.38	0.04	0.39	0.64
26	0.38	0.73	0.97	0.55	0.70	0.98
27	0.11	0.14	0.52	0.04	0.12	0.41
28	0.83	0.93	0.97	0.72	0.92	0.96
29	0.31	0.70	0.98	0.32	0.70	0.99
30	0.61	0.80	0.89	0.47	0.81	0.89

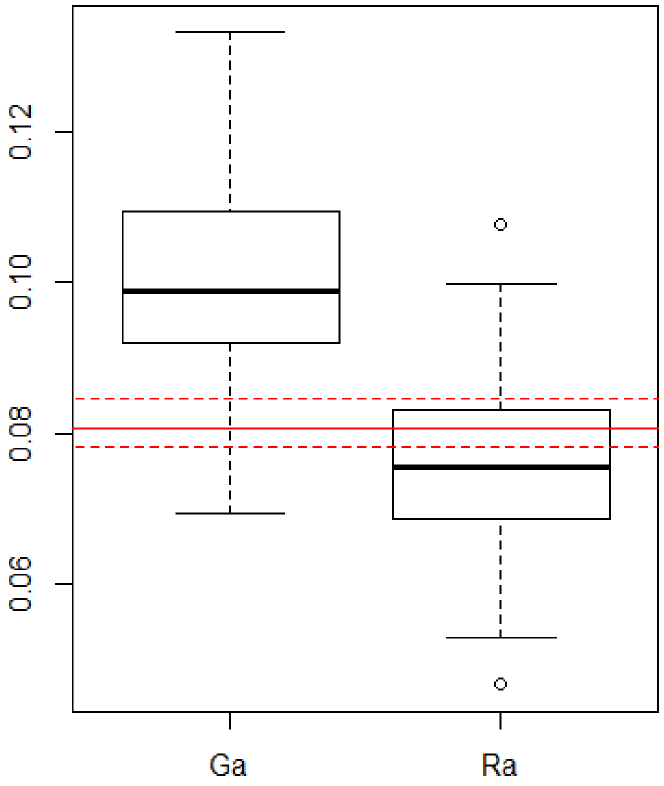
Table 10. Orlando and Thissen's S-X^2 for TNT (Hand-rated)

Item	GPC			MPC		
	Low	Median	Upper	Low	Median	Upper
1	0.08	0.23	0.71	0.09	0.26	0.84
2	0.06	0.30	0.74	0.01	0.13	0.52
3	0.65	0.91	0.98	0.66	0.93	0.98
4	0.05	0.40	0.89	0.03	0.41	0.93
5	0.19	0.37	0.66	0.05	0.33	0.88
6	0.09	0.42	0.57	0.04	0.14	0.30
7	0.22	0.52	0.87	0.15	0.55	0.84
8	0.04	0.33	0.80	0.04	0.39	0.78
9	0.09	0.24	0.47	0.07	0.20	0.38
10	0.13	0.23	0.43	0.14	0.29	0.40
11	0.55	0.78	0.95	0.44	0.64	0.89
12	0.13	0.34	0.47	0.04	0.43	0.91
13	0.03	0.39	0.74	0.05	0.43	0.77
14	0.11	0.43	0.76	0.04	0.30	0.78
15	0.02	0.17	0.62	0.03	0.23	0.70
16	0.10	0.49	0.72	0.08	0.37	0.76
17	0.18	0.53	0.83	0.16	0.47	0.84
18	0.29	0.36	0.52	0.17	0.25	0.53
19	0.08	0.59	0.70	0.16	0.51	0.78
20	0.30	0.47	0.79	0.19	0.41	0.83
21	0.04	0.13	0.37	0.03	0.11	0.45
22	0.21	0.51	0.79	0.23	0.53	0.79
23	0.25	0.48	0.64	0.33	0.46	0.72
24	0.24	0.44	0.74	0.25	0.58	0.73
25	0.05	0.26	0.74	0.07	0.34	0.73
26	0.25	0.73	0.89	0.22	0.76	0.89
27	0.08	0.30	0.54	0.14	0.24	0.49
28	0.09	0.37	0.74	0.09	0.36	0.55
29	0.01	0.37	0.94	0.00	0.49	0.93
30	0.15	0.25	0.50	0.08	0.22	0.52

Standard Deviation of the Biserials

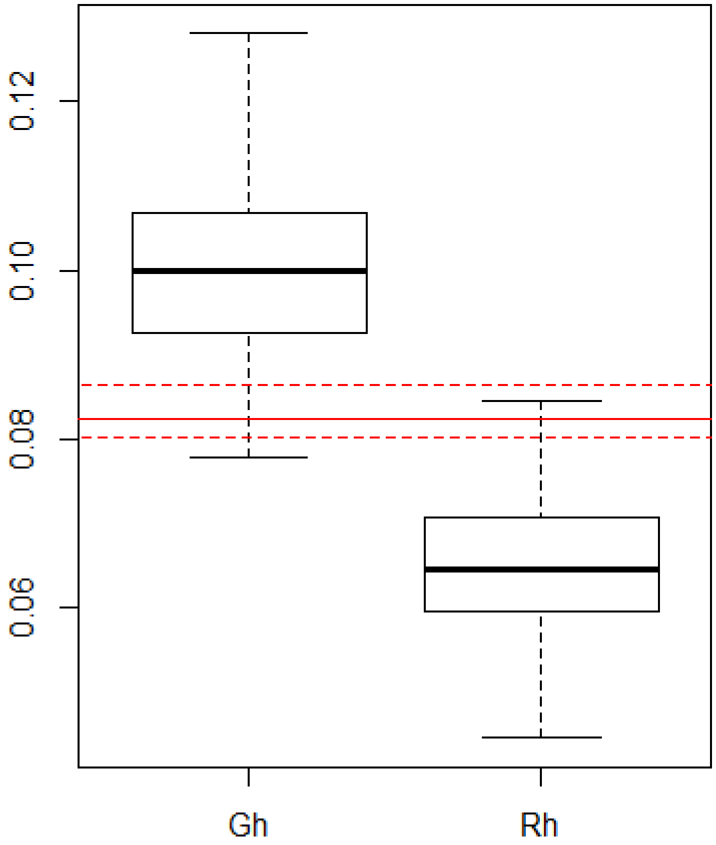
We applied a parametric bootstrap (instead of posterior predictive model checking) to Sinharay, Johnson, and Stern's (2006) idea of using standard deviation of the biserials to check the Rasch “equal slopes” assumption. The standard deviation of the biserial correlations of the dichotomous titem responses to the total score (excluding the item in question) should be related to the variability of the discrimination of test items (and thus the equal slope assumption). Comparing the standard deviation of the item biserials of the true data set, to that from simulated data sets fit by the Rasch and two-parameter logistic (2PL) IRT model could thus be used as a check on the equal discrimination assumption. Below similar graphs are made using the polyserial and simulated data sets from the generalized partial credit model and Masters partial credit model. The boxplots are the standard deviation of the biserials from 100 data sets simulated from the estimated GPC and MPC models. The red lines are the low, median, and high standard deviation of biserials from the 10 imputed data sets (that is, the actually observed data).

Figure 16. Box-plot for TNT (Automatic)



Ga: GPC Model; Ra: Rasch Model

Figure 17. Box-plot for TNT (Hand-rated)



Gh: GPC Model Rh: Rasch Model

The overlap in the boxplots between the two models suggests that the observable difference in fit for this sample size is not large based on this criterion, and that there is no significant evidence for lack of fit of the equal slopes assumption in the automated scoring case. For the hand scoring case it is suggestive that a few of the items might be misfitting, and additional examination may lead to their removal to ensure desired Rasch model properties.

Conclusion

TNT has very high reliability (Cronbach's Alpha=0.932) and it is a moderately strong scale (Mokken H=0.406). The hand-rated version has slightly better but similar results.
Overall, TNT is a good instrument for language proficiency evaluation for English as a Second Language (ESL). We could hypothesize that this reliability and scalability are also valid for other languages. Similar to the official ACTFL OPIc which offers evaluation in 13 languages[7] but not all of them have full psychometric evaluations in the literature. For a more definitive answer, a re-evaluation of the reliability for at least one of the popular languages (e.g. Spanish) could be helpful.

Language Testing International (LTI), the exclusive licensee of ACTFL, recommends 8 weeks[8] as the minimum time between test and retest. The reason obviously is related to the fact that it is difficult to increase your ACTFL level with only 10 levels available. TNT gives a continuous score (with one digit after the decimal point) between 0 and 10.0, or basically 100 micro-levels, and it can theoretically detect even a small incremental improvement. Given its high reliability and scalability, we could hypothesize that the TNT may be useful for language proficiency evaluations even for shorter than the 8-week period required for the ACTFL.

Of course, this option will depend of the number of study hours. For example, TNT (Spanish) was used in another language app efficacy study (Vesselinov & Grego, 2019). In this study, on average, for one hour of study the participants gained 0.13 TNT points (95% CI 0.08-0.17). In other words, it would take on average about 8 hours of study to increase the TNT level with 10 micro-levels, or one full level (e.g. from 1.x to 2.x). This efficacy has not been tested for TNT (ESL).

ACTFL estimate is an important feature of TNT. Because ACTFL TNT has interval ratings, it cannot be compared directly with the discrete ratings of ACTFL OPI (Levels 1-10).
Descriptively speaking, the TNT interval estimates in this study contained about 58% the OPI discrete levels. Working with the mid-interval points we can conclude that the TNT estimates

[7] https://www.actfl.org/professional-development/assessments-the-actfl-testing-office/oral-proficiency-assessments-including-opi-opic
[8] https://www.languagetesting.com/how-long-does-it-take

have a relatively good correlation (0.6-0.7) with the OPI determined levels. The correlation and other psychometric measures are not high enough for the TNT estimates to be equated to the OPI generated levels. For one-time language proficiency evaluation, the ACTFL TNT estimates cannot replace the ACTFL OPI levels.

But, since TNT is a very reliable scale, it can be hypothesized that ACTFL TNT estimates could be suitable for test-retest studies to measure the gain in language proficiency after a study period.

In general, the correlation between the automatic and hand-rated scores is strong (Spearman=0.82) but below the suggested threshold of 0.866, and there are some big differences present for some scale items. Examining the difference in use of the 0-3 response scale between the automatic and hand-rated scoring may indicate targets for further study. Both the automatic and hand-rated exams had items with lower than desired scalability (they did not measure the underlying latent trait measured by the scale as a whole as well as would be desired), and there was some evidence that there may be hand-scored items that do not satisfy the Rasch requirements adequately (although the majority of the scale might).

The correlation between the underlying latent traits (IRT theta) for TNT and the OPI is relatively strong (up to 0.7), but it is not above the suggested threshold of 0.866. This suggests that additional validity work is needed to determine how the trait measured by the TNT differs from that measured by the OPI, and the practical import of those differences.

Overall, the conclusion is that TNT is a reliable scale measuring one underlying latent trait and can be used for language proficiency testing and research. TNT has good psychometric characteristics but there is a room for improvement in terms of reviewing some of the included items, and in terms of the validity work in regard to the differences between the precise traits TNT and OPI measure.

References

Cai, L. & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, 66, 245-276.

Dorans, N., Pommerich, M., Holland, P. (Editors). (2007). Linking and Aligning Scores and Scales, Springer.

Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement*, 9, 139-164.

Kang, T., Chen, T. (2007). An Investigation of the Performance of the Generalized S-X^2 Item-Fit Index for Polytomous IRT Models, ACT Research Report Series, 2007-1. https://files.eric.ed.gov/fulltext/ED510479.pdf

Kolen, M., Brennan, R. (2004). Test Equating, Scaling, and Linking. Methods and Practices, 2nd Ed., Springer.

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

Muraki, E. (1992). A Generalize Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, 16, 159-176.

Sijtsma, K. Molenaar, I.W. (2002). Introduction to Nonparametirc Item Response Theory. New York: Sage.

Sinharay, S., Johnson, M., Stern, H. (2006). Posterior Predictive Assessment of Item Response Theory Models. *Applied Psychological Measurement*, 30(4), 298–321.

Surface, E., Dierdorff, E. (2003). Reliability and the ACTFL Oral Proficiency Interview: Reporting Indices of Interrater Consistency and Agreement for 19 Languages, *Foreign Language Annals*, Vo, 36, No. 4.

Vesselinov, R., Grego, J. (2019). Mango Languages Efficacy Study, http://comparelanguageapps.com/documentation/Mango_Languages_FinalReport2019.pdf