# The Development of an Automated Online German Speaking Proficiency Assessment

## Introduction

In the ever-increasingly globalized community, organizations need to be able to validly and reliably assess the speaking ability of their emissaries in communicating with multiple partners who may speak different languages. Institutions invested in the creation of various language measures have developed sophisticated assessments of speaking ability in multiple languages. These assessments, however, are usually expensive and not very scalable.

An alternative to these traditional assessment solutions is Emmersion's Speaking tests with their TrueNorth (TN) Score. Unlike most measures of language speaking ability, the Emmersion's speaking tests are relatively inexpensive and highly scalable. Also, unlike most measures of language speaking ability, Emmersion speaking tests do not require administration by highly trained proctors or rating of responses by highly-trained language experts. Automated speech recognition (ASR) services that utilize natural language procession (NLP) are used to rate the percentage of each auditory prompt that each examinee correctly reproduces. Further, Emmersion speaking tests can be administered and speaking ability estimated in as little as 15 minutes.

Emmersion speaking tests use elicited imitation as its methodological framework for assessment development and administration, where examinees are asked to repeat verbatim sentences presented to them auditorily (Vinther, 2002; Erlam, 2006; Burdis, 2014). Prior to implementing an ASR's NLP, however, language experts rate all of the examinee responses collected during the piloting phase. This is done to 1) help train an ASR's NLP to recognize and rate auditory responses relative to their corresponding auditory prompts and 2) to ensure that the ASR's NLP is a valid and reliable automated rater of speaking ability.

The purpose of this technical paper is to document the development and reliability of the Emmersion German Speaking (EGS) Form A as a valid and reliable measure of German speaking ability. Including the current section (i.e., introduction), it includes a method section, a results section, and a summary. The next section describes the method used to develop and validate the Emmersion German Speaking (EGS) Form A.

## Method

A pilot version of EGS Form A comprising 60 elicited imitation items was administered in collaboration with a stakeholder who regularly trains and sends its emissaries to and work toward its goals. Further, the Oral Proficiency Interview-computer (OPIc), developed by the American Council on the Teaching of Foreign Languages (ACTFL), was administered to collect evidence of criterion and convergent validity and to develop a scoring algorithm. Pilot test data were gleaned from 74 examinees, 62 of which also provided OPIc scores. Although the dataset was small, it was empirically sufficient for developing a valid and reliable assessment of German speaking ability.

Prior to calibration, each item's response score was transformed from a value derived from a scale ranging from 0-100 raters to a value derived from a polytomous scale ranging from 0-3. These initial scoring and calibration efforts were used to develop a beta version of the assessment. A final version of the assessment was developed after the ASR's NLP was confirmed as a valid and reliable measure of each item's response percentage accuracy. Scores between 0-10 were transformed to 0, scores between 10-50 were transformed to 1, scores between 50-90 were transformed to 2, and scores greater than or equal to 90 were transformed to 3.

After the scores were transformed to a polytomous scale and checked for equivalent response option distributions, the data were fitted to a graded response model (Samejima, 1969). Results from this analysis were used to inform the assembly of an assessment representative of the full range of speaking ability and to remove aberrantly fitting items and response patterns. Aberrantly fitting items were identified and removed via Orlando's and Thissen's (2000; 2003) and Kang's and Chen's (2007) $S\text{-}\cdot^2$ statistic; and aberrantly fitting response patterns were identified and removed via Drasgow's, Levine's, and Williams's (1985) Zh statistic. After aberrantly fitting items and response patterns were removed, the final dataset comprised 67 response patterns and 52 elicited imitation items.

To determine if the sample size was sufficient for confirming the range of item difficulties, if the items were sensitive enough to distinguish between range of speaking abilities of the examinees, and if the thresholds for transforming scores resulted in approximately equivalent response option distributions, the data were also fitted to a partial credit model to perform a Rasch reliability analysis and to examine the Andrich thresholds (Masters, 1982; Wright & Masters, 1982).

Spearman rank correlation was calculated to examine the reliability between EGS Form A's speaking ability estimate and the OPIc speaking proficiency score as well as between EGS Form A's  predicted OPIc sores and the actual OPIc proficiency scores. Agreement between predicted and actual OPIc scores was checked by examining the percentage of predicted OPIc scores that matched the actual OPIc scores and by examining the percentage of predicted OPIc scores that were within one of the actual OPIc scores.

The current technical paper is comprised of four sections: an introduction (this section), a method section, a results section, and a summary. The following section reports on the method used to develop and validate EGS Form A as a valid reliable measure of German speaking ability.

## Results

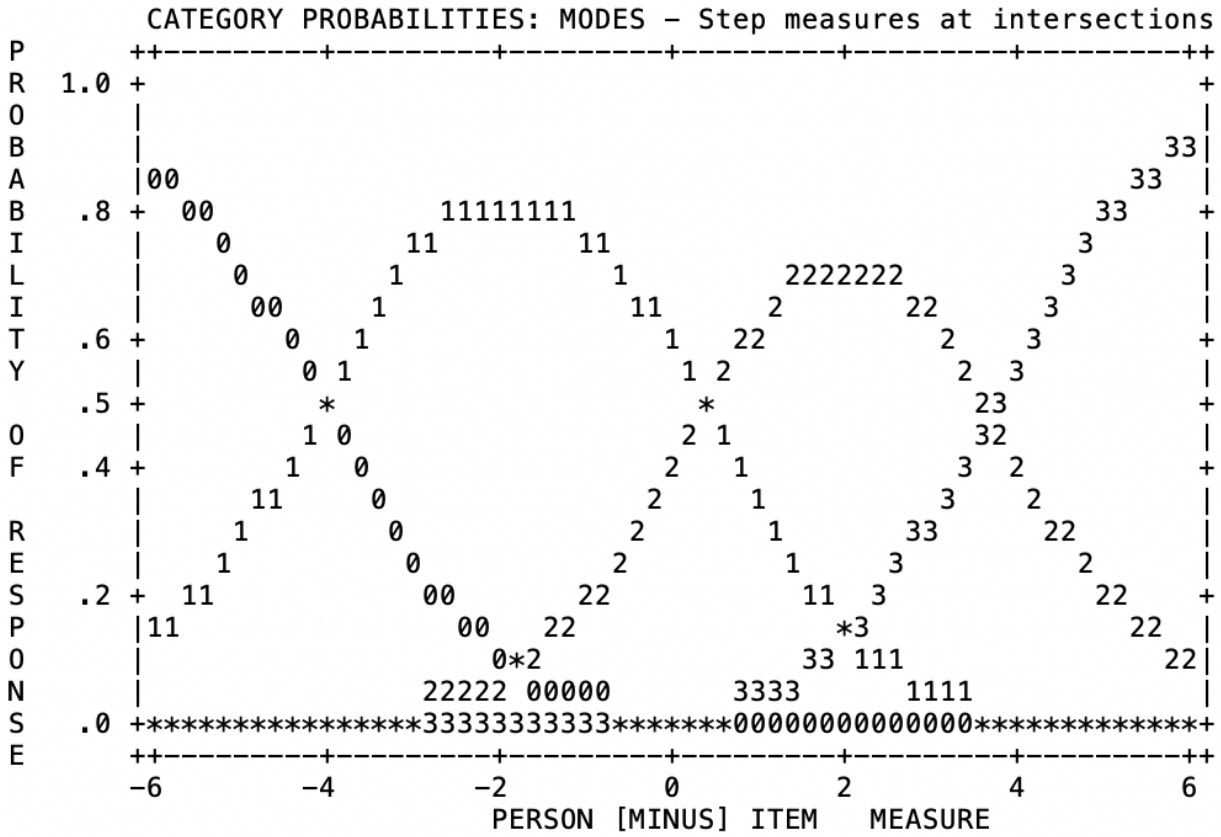
Andrich thresholds representing the distribution of response categories showed that the thresholds at which the human-rated responses were converted to polytomous values resulted in approximately equivalent distributions and that the distances between adjacent categories were between the recommended 1.4 to 5 logits (see Figure 1; Linacre, 2002). This implied that each item could be treated as up to three dichotomous items and that any non-informative spacing between the adjacent categories, indicated by large Andrich threshold advances (greater than 5 logits), was minimized. It also implied that the polytomous values could be treated as if they were continuous.

Person reliability, which ranges from 0-1 and indicates how well an assessment's items can distinguish between examinees' ability, was .99. Thus, well-fitting items comprising the pilot assessment could reliability distinguish examinees across the ability range. The ratio between the true ability range and measurement error, called the person separation index, was 9.23 (Wright & Masters, 2002). Using this index to estimate the number of available strata to categorize the examinees' ability indicated that there were approximately 13 ability strata that were statistically distinctive.

Item reliability, which ranges from 0-1 and indicates the reproducibility of item difficulty hierarchy relative to the sample size, was 98. Thus, there were enough test records to reproduce the item hierarchy. The ratio between the true range of item difficulty and measurement error, called the item separation index, was 8.08. Using this index to estimate the number of available strata to categorize the items by difficulty indicated that there were approximately 11 difficulty strata that were statistically distinctive.

The Wright Map in Figure 2 illustrates how examinees are distributed by their range of abilities range relative to how items are distributed by their range of difficulties, which use the same scale (Wright & Stone, 1979). This shows that the examinees and items well distributed across the scale.

## Figure 1. Russian TNT rating category distribution



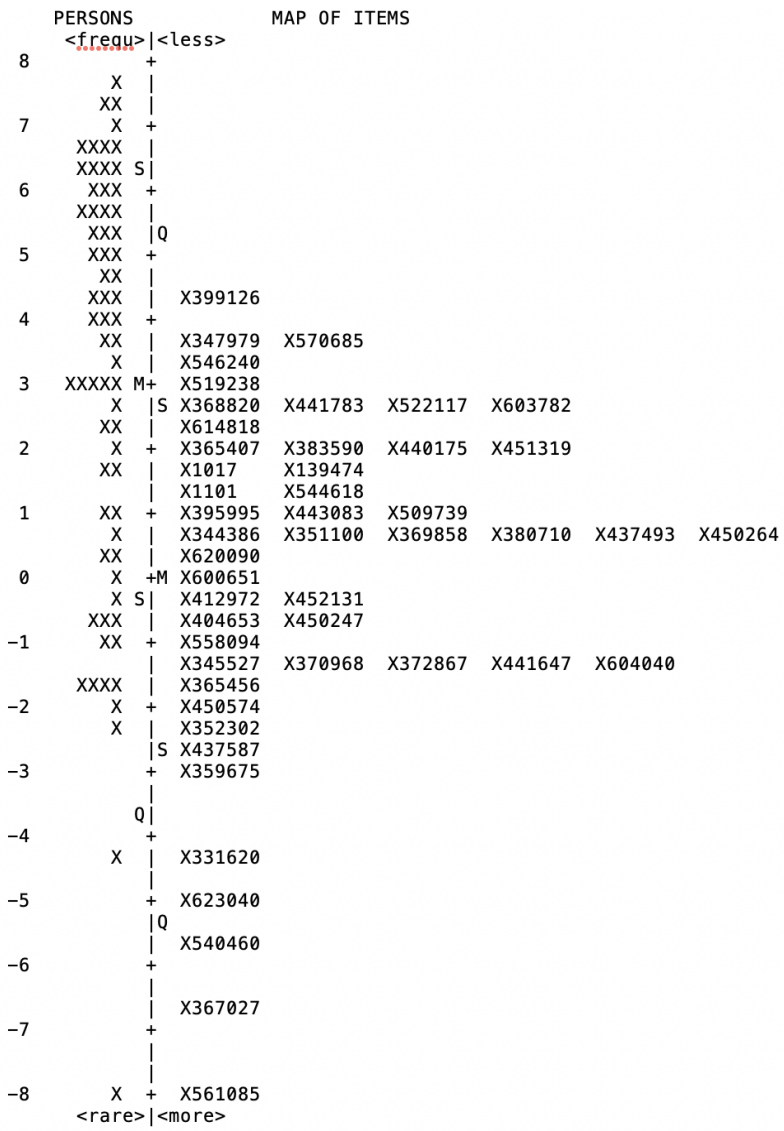

Treating the measures as alternate assessments of German speaking ability, the Spearman rank correlation between examinees' estimated speaking ability derived from EGS Form A and their speaking proficiency derived from the OPIc was used to validate EGS Form A as a valid and reliable measure of German speaking ability. According to this statistic, there was excellent reliability between the two measures, $r_s = .900$, $p < .001$.

For the predictive scoring algorithm, a generalized additive model was fitted by expressing OPIc scores as the sum of the smooth functions of estimated speaking ability and its standard error as predictors:

$$ACTFL\ Level = s(\theta) + s(\theta_{SE})$$

This predictive model fit the data excellently, explaining 95.5% of the deviance, which is a generalization of the $R^2$ statistic that is used in conventional regression as a measure of how much variance is explained to generalized additive modeling (see Figure 3). The Spearman rank correlation coefficient indicated that that there was excellent reliability between the predicted OPIc scores and the actual OPIc scores, $r_s = .905$, $p < .001$. Finally, there was sufficent agreement between predicted OPIc scores and actual OPIc scores at a 76% perfect match and a 94% match within 1 in either direction.

## Figure 2. Wright Map

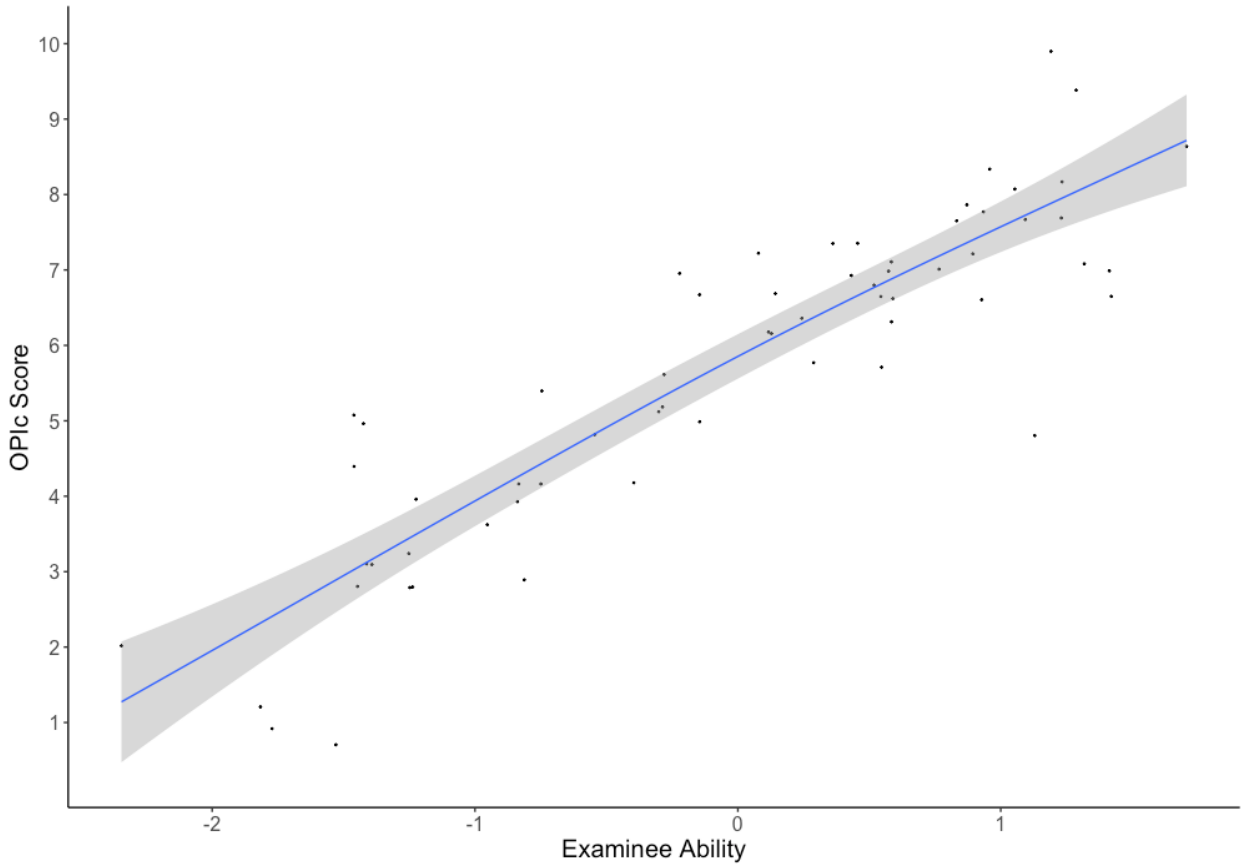


## Summary

Evidence in this technical report was presented in support of the German TNT Form A as a valid and reliable measure of speaking ability: 1) Rasch reliability and separation statistics and 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 speaking proficiency indicated that reliability was excellent; 3) 95.5% of the deviance was explained by the predictive model; 4) the Spearman rank correlation between predicted OPIc scores and actual OPIc scores indicated that reliability was excellent; and 5) the scoring algorithm met internal guidelines for acceptable agreement between predicted OPIc scores and actual OPIc scores. Thus, we believe that sufficient evidence was presented in support of the German TNT Form A as a valid and reliable assessment of German speaking ability.

### Figure 3

*Scatterplot illustrating the relationship between OPIc scores and German TNT Form A ability estimates*



## Scoring Version Update

In early 2023,  the speech recognition engine that had been used for scoring of the EIS items was deprecated. The update to the speech recognition engine showed some discrepancies between the old version. As a result, we performed a handrater study where we compared the SRE output of the new model and tests scored by an expert German rater. The updated model for the previous speech recognition did not perform as well as a different speech recognition engine. This new SRE, however, showed sufficient agreement with the handrater for us to have confidence in replacing the previous SRE.

However to further mitigate any potential disruption we did a full psychometric version update including confirming scoring model, updating item difficulty parameters and retraining the machine learning prediction of TrueNorth Score. Following this work, rescoring thousands of previous assessments showed agreement between the previous version and the updated version of .967.

## 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. (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.