A Common Testing Framework for Measuring Spoken Language Skills of Non-Native Speakers

Masanori SUZUKI[1], Yasunari HARADA[2]

[1]*Ordinate Corporation, 800 El Camino Real Suite 400, Menlo Park, CA 94025 U.S.A.*
msuzuki@ordinate.com

[2]*Waseda University, 1-6-1 Nishi Waseda, Shinjuku-ku, Tokyo, Japan 169-8050*
harada@waseda.jp

Abstract

A testing framework was developed to create automated spoken language tests in multiple languages. The tests built on this framework are delivered over a telephone and are automatically scored using a speech recognition engine and a computerized scoring system. With this testing system, a test can be administered in large volumes, completed in only 10-15 minutes, and scored very rapidly without sacrificing reliability or quality. The spoken English test and spoken Spanish test were built upon this common testing framework and are already in operation. Currently, a spoken Japanese test is under development. Data from the SET and SST will be presented to show how tests built on top of this common framework are both reliable and valid.

Keywords: *spoken language, speech recognition, test delivery, automated scoring*

1 Introduction

As global communication continues to advance, one of the major focuses of language instruction is to enhance learners' ability to communicate, that is, to enhance their oral communication skills. Therefore language assessment should emphasize the competent use of language in spoken communication.

Traditionally, Oral Proficiency Interviews (OPIs) are often viewed as assessments well-aligned with this goal. However, an intrinsic limitation of OPIs is that the number of tests that can be administered in a given language is constrained by the number of available trained interviewers.

Over the past decade, advances in speech recognition technology have enabled the development of an automatically administered and scored spoken language test in English (Bernstein, De Jong, Pisoni, Townshend, 2000). During the test, human-recorded prompts are played over a land-line telephone, and test-takers' responses are automatically scored using speech recognition and other computer technologies. Because the test is automated, large volumes of tests can be administered and scored very rapidly.

The English test, called SET (Spoken English Test), was first built on top of a common testing framework. The framework consists of three components: a test delivery system, a computerized scoring system, and a validation process. The two goals of interest when formulating the common testing framework were first, to enable the creation of high quality tests that take advantage of computer technologies to allow for precision and scalability not possible with human-scored tests, and second, to facilitate the creation of such tests in any language very efficiently while retaining the same level of quality. To date, the framework has been used to create spoken language tests for English and Spanish and is currently being employed for Dutch and Japanese test development.

In the subsequent sections, we first give an overview of how the tests are administered and are scored, then describe the common testing framework and present data from the two production tests (English and Spanish) to show how the tests built on top of the framework are reliable and valid.

2 Test Administration Overview

The tests built on top of the testing framework are automatically administered over a telephone and are automatically scored by the computerized scoring system. Prior to taking a test, a test-taker receives a test paper. One side of it has general test instructions and the other side has a unique Test Identification Number, a telephone number, the verbatim spoken instructions, and examples of tasks and items. When a test-taker is ready to take the test, the test-taker calls a telephone number on the test paper. Then, to begin the test, the test-taker is asked to enter the Test Identification Number printed on the test paper using the telephone keypad. The tests take approximately 10-15 minutes. Ordinate's test delivery system presents a test-taker with a series of spoken prompts in the target language (e.g. English), and the test-taker responds by speaking.

A score report becomes available on Ordinate's website usually within a few minutes after a test has been completed. The score report consists of one Overall score and four subscores: Sentence Mastery, Vocabulary, Fluency, and Pronunciation. In other words, the tests measure two aspects of the spoken skills: *what* the test taker said and *how* the test taker said it. The *content (what)* aspect of the spoken skills is reflected in the Sentence Mastery and Vocabulary subscores and the *manner (how)* aspect of the spoken skills is reflected in the Fluency and Pronunciation subscores. The scores are reported in the range of 20-80 and each aspect counts for 50% of the Overall score. These test administration procedures are schematized in Figure 1.

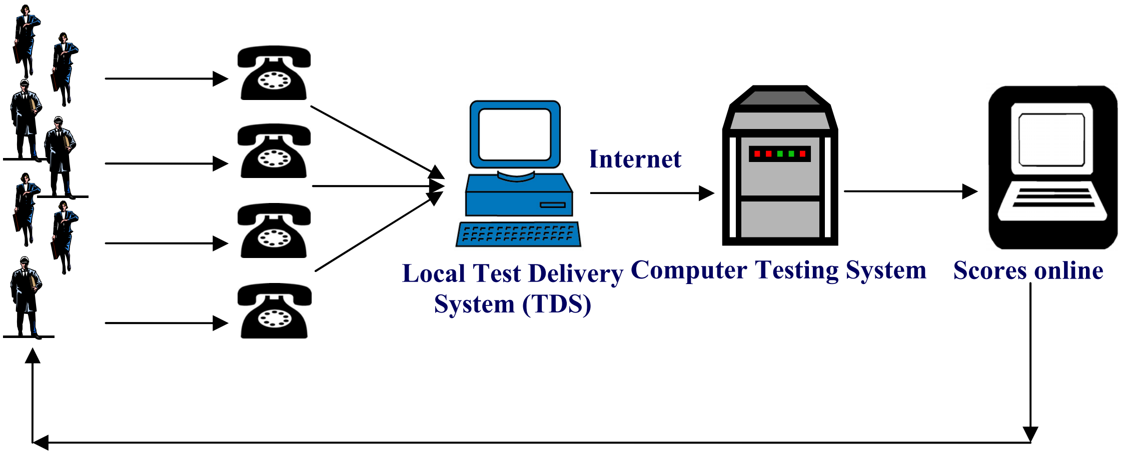


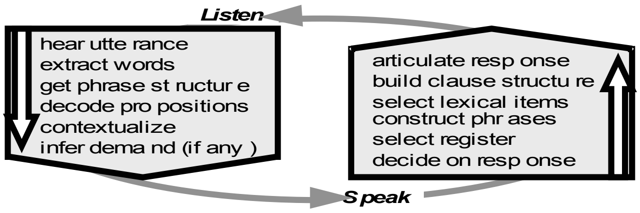
Figure 1. Test administration Scheme

3 Common Testing Framework

3.1 Test Construct

The basis for the common testing framework is the construct of its tests. Specifically, the computerized spoken language tests measure non-native speakers' *facility in a spoken language*; that is, the ability to understand the spoken language on everyday topics at a native-like conversational pace, and formulate appropriate and intelligible spoken responses in that language.

Spoken language facility is essential to successful oral communication – if language users cannot track what is being said, extract meaning as speech continues, and then formulate and produce a relevant and intelligible response in real time, they will not be able to interact in effective communication. These components are schematized in Figure 2, adopted from Levelt (1989).



Adapted from Levelt, 1989

Figure 2: Conversational Processing Components in Listening and Speaking

In the tests built on top of the common testing framework, all the test items are presented orally. Test takers need to understand them and answer them intelligibly. Each of these *listen-then-speak* items requires real time receptive and productive spoken language forms. In other words, the tests are intended to measure the degree of automaticity in basic decoding and encoding of oral language. Automaticity is the ability to access and retrieve lexical items, to build phrases and clause structures, and to articulate these without conscious attention to the linguistic code (Cutler, 2003; Jescheniak, Hahne, and Schriefers, 2003; Levelt, 2001).

Automaticity is required in order for the speaker/listener to be able to pay attention to what needs to be said rather than to how the encoded message should be structured. As a result, performance on the language tests predicts a more general spoken language ability, which is essential in successful oral communication.

3.2 Test Delivery System

Test administration is performed by the first key component of the common testing framework, the *Test Delivery System*. The test delivery is done over the telephone and via the Internet. As described above, each test-taker calls into the Ordinate testing system, listens to spoken prompts and answers them appropriately over the telephone. The test-taker's responses are stored in Ordinate's database system. In some countries such as Japan, Korea, China and some European countries, a local TDS (Test Delivery System) is set up and test-takers in those countries take tests using a local toll-free number. Test-takers' responses first go to the TDS and then are sent via the Internet to the Ordinate testing system for scoring.

Seven tasks have been developed to measure *facility in a spoken language*, as shown in Table 1. The seven tasks are Reading, Repeat Sentences, Opposites, Short Answer Questions, Sentence Builds, Open Questions, and Story-Retellings. Although the tests built on top of the common testing framework share these seven tasks or share some of these tasks, the individual items in each test are written specifically for that test by native speakers of the target language.

In Part A, test-takers are asked to read sentences at random from among the printed sentences on the test paper. In Part B, test-takers repeat sentences verbatim as they hear them. In Part C, test-takers are presented with a word (orally) and are asked to respond with a word that represents an opposite meaning. In Part D, test-takers are presented with a series of questions and they answer each question with a single word or a short phrase. Part E requires test-takers to make a reasonable sentence out of three short phrases that they hear. In Part F, test-takers hear a spoken prompt in the target language asking for an opinion, and they provide an answer with an explanation in the target language. In Part G, test-takers listen to a very short narrative and then are asked to re-tell what happened in their own words.

Table 1. Seven Tasks

Part A:	Reading
Part B:	Repeat Sentences
Part C:	Opposites
Part D:	Short Answer Questions
Part E:	Sentence Builds
Part F:	Open Questions
Part G:	Story-Retellings

When test items are presented, they are presented in a stratified random order so that the item difficulty generally increases over the sequence of items presented. The difficulty of each item is calculated using IRT (Item Response Theory) after the data collections are conducted from native speakers and non-native speakers of the target language. These items are assembled into tests from a larger item pool, so the likelihood of one particular test-taker seeing the same items over different test administrations is low. Each assembled test covers about the same range of item difficulty as measured by IRT.

3.2 Computerized Scoring System

The second key component of the common testing framework is *the computerized scoring system.* The same system is used for all tests and consists of many integrated components necessary for automatic scoring. Ordinate uses an HMM-based ASR (Automated Speech Recognition), speech to text alignment, and non-linear models to perform automatic scoring. The speech recognition itself has several components including an acoustic model, a dictionary, and a language model. Although these models have the same basic structure across languages, all the models are specific to the tested language. These models are trained on data collected from native and non-native of the tested language during data collection, so that the speech recognizer is optimized for various types of non-native speech patterns in each language.

Each incoming response is recognized automatically and the words, the pauses, the syllables, the phones, and even some subphonemic events are identified automatically and extracted from the recorded signal for measurement.

These recognition results are fed into the computer scoring system. The computerized scoring system examines the two aspects of the speech: the *content* aspect and *manner-of-speaking* aspect. The content of the response is scored according to whether the test taker used expected words in the correct sequence. The manner-of-speaking aspect is calculated by measuring the latency of the response, the rate of speaking, the position and length of pauses, the stress and segmental forms of the words, and the pronunciation of the segments in the words within their lexical and phrasal context. These measures are scaled according to native and non-native distributions and then combined so that they optimally predict human judgments.

3.3 Validation Studies

The third component of the common testing framework is validation studies. The general approach to validation in the common testing framework highlights three metrics as evidence of the test's quality: high reliability, ability to show effective separation between samples of native and non-native test takers, and strong correlations with other established measures of oral language proficiency. Successfully achieving these metrics relies not only on the integrity of the test, but also on a rigorous methodology.

Data from both native and non-native test takers are used for the validation process. Often data collected during the development process is set aside for this purpose. Native speakers are usually literate adults who live in a variety of regions and countries and represent a range of age groups. Non-native speakers represent a broad range of proficiency levels and native languages.

3.3.1 Reliability

From the validation data collection, split-half reliability is computed for all subscores and for the Overall score of the test. For the English and Spanish production-level tests built on top of the framework, Overall score reliabilities are .97 and .96 respectively.

3.3.2 Native and Non-native Group Performance

Next, statistics regarding the native and non-native samples are calculated. Of particular interest is whether the test shows a separation between the two groups.

For the SET-10, native speakers of English consistently obtain high scores. Fewer than 5% of the native sample scored below 68. Learners of English as a second or foreign language, on the other hand, are distributed over a wide range of SET-10 scores. Only 5% of the non-natives scored above 68. The Overall scores show effective separation between native and non-native test takers.

A similar analysis was done for the Spanish test. Figure 3 presents cumulative distribution functions which show the percentage of test takers in each group who received a given score on the test or lower for native and non-native speakers of Spanish.

Note that the range of scores displayed in Figure 3 is from 10 through 90, whereas the SST scores are reported on a scale from 20 to 80. Scores outside the 20 to 80 range are deemed to have saturated the intended measurement range of the test and are reported as 20 or 80.

The distribution of the native speakers clearly distinguishes the natives from the non-native sample. Fewer than 5% of the native speakers received a score below 75, while only 10% of the non-native speakers received a score above 75. The results from this analysis suggest that the SST has high discriminatory power among learners of Spanish as a second or foreign language, whereas native speakers obtain near-maximum scores.

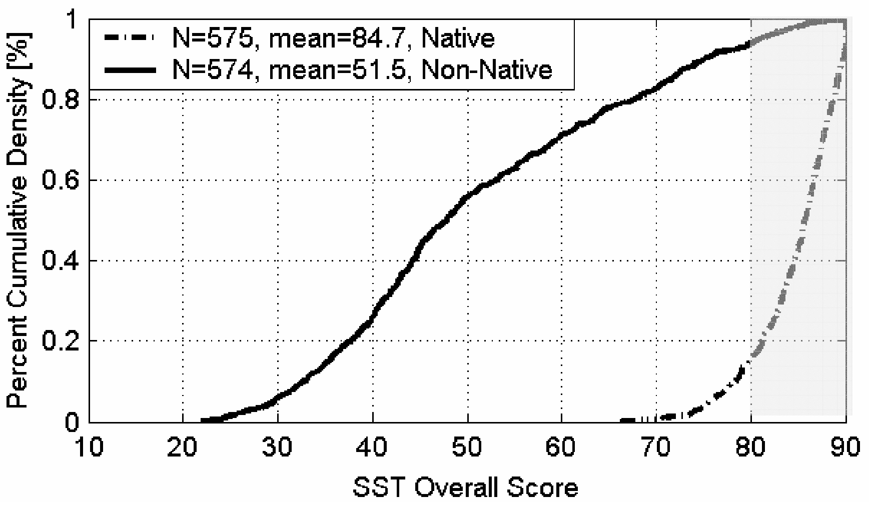


Figure 3. Cumulative distribution functions of SST Overall scores for native and non-native speakers.

3.3.3. Correlations between Test Scores and Human Ratings

The third validity metric is the correlation between the test scores and other well-established measures of oral language proficiency. Two variations for these concurrent validity studies are (1) estimating each test taker's proficiency level by having human experts rate responses to open questions and story retellings and then correlating these ratings to test scores, and (2) collecting ratings from independent Oral Proficiency Interviews and correlating these ratings to test scores.

For the Spoken English Test, proficiency estimates were collected for 268 non-native speakers and 35 native speakers on an oral interaction scale based on the CEF. Responses to Open Questions were assigned randomly to six raters who together produced 7,266 independent ratings in an overlapping design. The ratings from the two raters with the largest amount of overlapping data related to 397 responses. These raters showed perfect agreement in assigning a CEF level to 63% of the cases and differed by only one level in a further 30% of the cases. Rater agreement overall was 0.89.

Figure 4 shows the relationship between the SET-10 score and the CEF levels. The correlation was 0.88. The graph also shows how both instruments (the SET-10 and CEF) clearly separate the native and non-native groups.

Similar validation studies with human estimates of proficiency were conducted for the Spoken Spanish Test. In addition, results from the SST were correlated with OPI scores by ACTFL (the American Council on the Teaching of Foreign Language). The standard ACTFL interview, which was human-conducted and human-rated, was administered with at least two official ACTFL ratings per interview. ACTFL submitted 52 scores, one for each of the 52 participants. The test takers participated in the interview within a day of the SST administration.

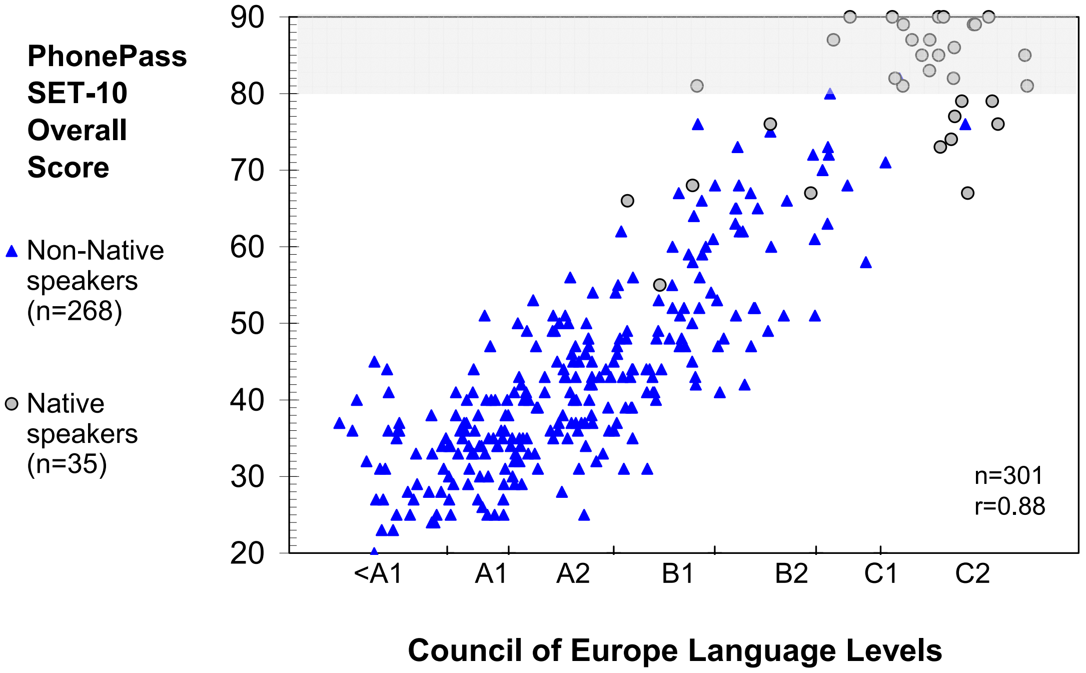


Figure 4. Correlation between SET-10 Overall scores and CEF-levels.

For data analysis, the computer program FACETS (Linacre, 2003) was used to estimate rater severity, subject ability, and item difficulty (Linacre, Wright, and Lunz, 1990) based on a one-parameter Rasch model. The model expresses scores in a mathematical unit called a Logit. Figure 5 is a scatter plot of the ACTFL OPI scores as a function of SST ratings for 52 Spanish learners.

The correlation for these two tests is 0.86, indicating a strong relation between the machine-generated scores and the human-rated interviews.

The validation data for both the Spoken English Test and the Spoken Spanish Test indicate that native speakers consistently obtain high scores on the tests, while learners are distributed over a wide rage of scores. This separation of the two sample populations illustrates the power of the tests as a measurement instrument of spoken language ability. In addition, the validation experiments show that scores of tests built on top of the common testing framework are reliable and correlate highly with other measures of spoken language ability. Although the same methodology for the validation process can be used for tests of any language developed in the framework, all the data collected from native and non-native speakers are unique to each language.

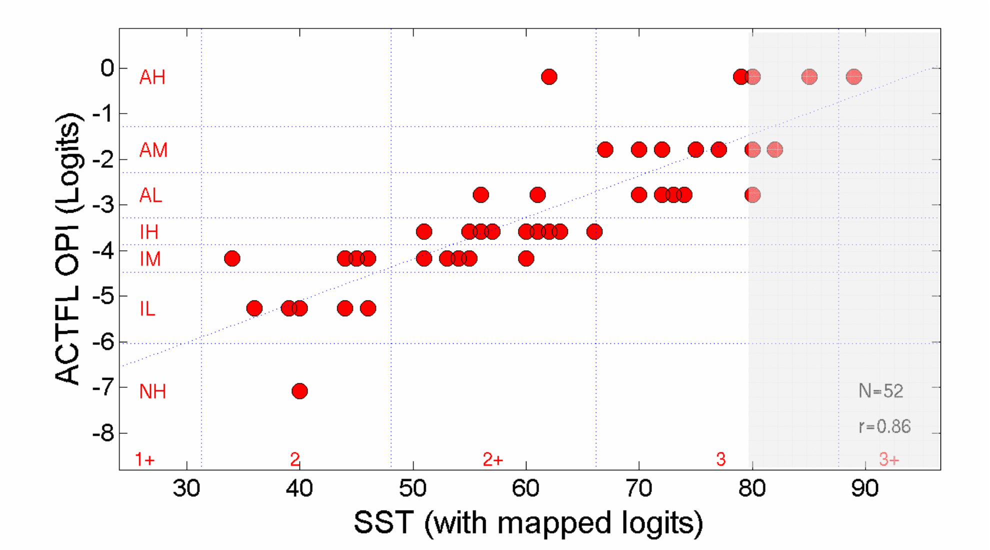


Figure 5. ACTFL OPI scores as a function of SST ratings.

4 Conclusion

The common testing framework was developed to create spoken language tests that are automatically administered and scored. The benefit of this type of test, compared to human conducted and scored interviews, is that they can be administered in large volumes and scored rapidly without sacrificing reliability or quality. The common testing framework facilitates the efficient development of these automatic spoken language tests in any language. The test architecture in which different tasks provide information about the content and the manner of speaking, generalizes to any language, although the individual items in each test are written specifically for that test. The computerized test delivery and scoring system provide a general means of automatically administering and scoring tests in any language, with the acoustic models, language models, pronunciation, and fluency models trained separately for each language. Finally, the validation process generalizes to tests for any language, although the collected data are unique to that test. Data from English and Spanish tests, which were built on top of the common testing framework, show that these tests are reliable, that they can distinguish native and non-native groups of test takers, and that they correlate highly with other measures of oral language ability.

Acknowledgements

The research reported here is partly supported by Grant-in-Aid for Exploratory Research #16652040, provided by the Japanese Ministry of Education, Culture, Sports, Science and Technology, and by funding from Ordinate Corporation.

References

Bernstein, J., De Jong, J.H.A.L., Pisoni, D., & Townshend, B. (2000). Two experiments on automatic scoring of spoken language proficiency. In P. Delcloque (Ed.) *Proceedings of InSTIL2000: Integrating Speech Technology in Learning*. University of Abertay Dundee, Scotland, 57-61.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Cutler, A. (2003). Lexical access. In L. Nadel (Ed.) *Encyclopedia of Cognitive Science, Vol 2, Epilepsy – Mental imagery, philosophical issues about*. London: Nature Publishing Group, 858-864.

Jescheniak, J.D., Hahne, A., & Schriefers, H.J. (2003). Interformation flow in the mental lexicon during speech planning: evidence from event-related brain potentials. *Cognitive brain research*, 15(3), 261-276.

Levelt, W. (1989). *Speaking*: From Intention to Articulation. Cambridge, MA: MIT Press.

Levelt, W.J.M. (2001). Spoken word production: A theory of lexical access. *PNAS*, 98(23), 13464-13471.

Linacre, J.M (2003). *Facets Rash measurement computer program*. Chicago: Winsteps.com.

Linacre, J.M., Wright, B.D., & Lunz, M.E. (1990). A Facets model of judgmental scoring. *Memo 61*. MESA Psychometric Laboratory. University of Chicago. www.rasch.org/memo61.html.

Miller, G.A. & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, 2, 217-228.