LANGUAGE
TESTING

# Validating automated speaking tests

Language Testing
27(3) 355–377

DOI: 10.1177/0265532210364404
http://ltj.sagepub.com
SAGE

**Jared Bernstein, Alistair Van Moere and Jian Cheng**
Pearson Knowledge Technologies, Palo Alto, California

## Abstract

This paper presents evidence that supports the valid use of scores from fully automatic tests of spoken language ability to indicate a person's effectiveness in spoken communication. The paper reviews the constructs, scoring, and the concurrent validity evidence of 'facility-in-L2' tests, a family of automated spoken language tests in Spanish, Dutch, Arabic, and English. The facility-in-L2 tests are designed to measure receptive and productive language ability as test-takers engage in a succession of tasks with meaningful language. Concurrent validity studies indicate that scores from the automated tests are strongly correlated with the scores from oral proficiency interviews. In separate studies with learners from each of the four languages the automated tests predict scores from the live interview tests as well as those tests predict themselves in a test-retest protocol ($r = 0.77$ to 0.92). Although it might be assumed that the interactive nature of the oral interview elicits performances that manifest a distinct construct, the closeness of the results suggests that the constructs underlying the two approaches to oral assessment have a stable relationship across languages.

## Keywords

automated scoring, language testing, speech recognition, test validity, Versant

This paper evaluates the validity of spoken language tests in which the delivery and scoring is fully automated. Currently there are three main methods of testing spoken language: oral proficiency interviews, semi-direct tests, and automated tests. The first method, the oral proficiency interview (OPI), is a live interaction between test-taker (or pair of test-takers) and human examiner(s). The examiners both deliver the test material and score the spoken responses. Examples include the Interagency Language Roundtable (ILR) OPI, the American Council on the Teaching of Foreign Languages (ACTFL) OPI and the International English Language Testing System (IELTS) interview. Interview tests are often perceived to be valid because they simulate conversation, and because the test format allows examiners to probe the ceiling of the test-taker's ability (ETS, 1982, p. 11).

**Corresponding author:**
Jared Bernstein, 299 S. California Avenue, Suite 300, Palo Alto, CA 94306, USA
Email: jared.bernstein@pearson.com

The semi-direct method of spoken language testing uses computers to present tasks and to capture spoken responses, which are evaluated by human raters following OPI-style criteria. These tests are different from real OPIs (Shohamy, 1994), but the scores are very comparable to, yet more reliable than, live tests (Stansfield and Kenyon, 1992).

The third method of spoken language testing is fully automated, as exemplified by the Versant tests (Pearson, 2009b), the speaking tasks within the Pearson Test of English (Pearson, 2009a), and the speaking section of the TOEFL iBT practice test (Zechner et al., 2009). In the Versant tests, a series of recorded spoken prompts are presented to which the test-taker gives a spoken response in real time. The test-takers perform on various item-types (sentence repeats, short questions, sentence builds, passage retells), and their responses are analyzed by algorithms which establish the content of the utterances and the manner in which they were spoken. Scores are produced on the dimensions of sentence mastery, vocabulary, fluency and pronunciation. A perceived disadvantage of this approach is that the language samples elicited are relatively short and constrained in comparison to more open tasks such as describing a memory or explaining an opinion. However, the elicited responses do manifest the construct claimed for the Versant tests, which differs from the more communicative construct of the oral proficiency interview.

There is consensus (Cronbach, 1988; Kane, 1992) that validity is a feature of the inferences or decisions made on the basis of test scores or score use. Validity is supported by an overall interpretive argument, where evidence justifies intended interpretations and uses of scores, interpretations are conceptually and empirically supported, and rebuttals should be made against reasonable counter-interpretations. Validity arguments are based on empirical data which addresses test reliability and content, and the concurrent, predictive, and consequential aspects of the test scores. Since tests vary in format and score use, it is reasonable that different tests prioritize different aspects of validity as more or less important.

We present several kinds of evidence to support the validity of the Versant automated tests. We define the language ability that is tested (construct definition), explain how this ability contributes directly to test scores (construct representation, or linking test scores to a theoretical score interpretation), and explain how Versant scores relate to abilities that support language use in a broader domain by establishing their relationship with OPI scores (concurrent evidence). Definition, representation and relation are crucial, though not sufficient, for a validity argument. Interpretive arguments also involve extrapolation from observed test behavior to non-test behavior and links from score-based interpretation to intended test use. The facility construct, however, is not situated within a target context: *facility* provides a measure of performance with the language without reference to any specific domain of use. Therefore, the basic claim for the automated test is: if person 1 scores higher than person 2, then person 1 has greater core spoken language skills than person 2. Inferences made on the basis of test scores reflect core spoken language ability, independent of social skill or traits such as intelligence, academic aptitude, or charisma. Thus, an unimaginative test-taker, who is socially awkward and uninformed, can be viewed through the facility construct as having full control of core speaking skills (see Hulstijn, 2006, quoted below, for a definition of core skills). Thus, extrapolation from a facility score to a communicative domain of use could only be established by mediated linking as schematized in Figure 1.

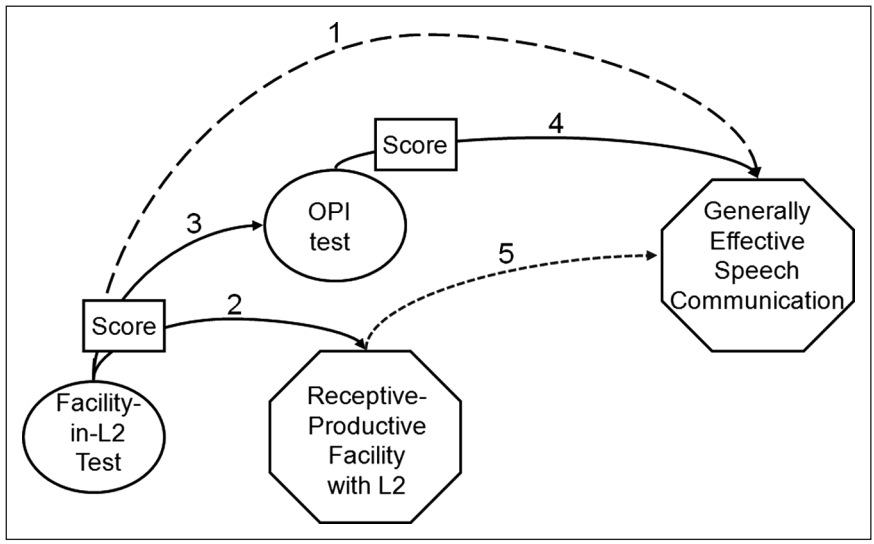


**Figure 1.** Direct and mediated links from facility-in-L2 tests to non-test language use domain

Establishing a direct link (arc 1 in Figure 1) is difficult because the target domain of score application (generally effective speech communication, or oral proficiency) is not specified in a well agreed upon model. Thus, in this paper, the representation of the facility construct (arc 2) is explained and supported with psychometric evidence. Then arc 3, linking facility-in-L2 test scores to OPI scores is established empirically. If one assumes that OPI scores are a good indicator of general effectiveness in speech communication, then there may be grounds for a mediated link (3→4 ~ 1) between a Versant test and effective speech communication. Similarly, if a warranted link can be established between decontextualized psycholinguistic task performance and the language use domain (see, for example, Hulstijn, 2007, and Schoonen, 2009), there would be a second mediated link (2→5 ~ 1). Evaluation of the 2→5 path is beyond the scope of this paper.

The following three sections set out the facility construct, present evidence for the interpretation of Versant scores as valid indicators of *facility* with a spoken language, and present a body of evidence that Versant scores are strong predictors of OPI scores. If the assumption holds that test-taker performance in a carefully administered OPI is predictive of behavior in non-test situations, then high correlation between the Versant and OPI scores would indicate that Versant scores are also predictive for similar non-test situations. If so, facility test scores may be useful across many target domains as part of a battery of tests for specific-domain speaking skills. These claims are of course open to counter-claims, which are also addressed in the final section.

## Spoken Language Constructs

A family of facility-in-L2 tests has been developed for English, Spanish, Dutch, and Arabic. The Dutch test is administered by the Dutch government under the name 'TGN',

**Table 1.** Test structure comparison shows the number of items presented by item type; unscored task-item types are shown in parentheses (n)

| Task-item type | Versant Spanish | TGN | Versant Arabic | Versant English |
|---|---|---|---|---|
| Read sentences aloud | 6 | 0 | 6 | 8 |
| Repeat sentences aloud | 16 | 24 | 30 | 16 |
| Say opposite words | 8 | 10 | 0 | 0 |
| Give short answer to questions | 16 | 14 | 20 | 24 |
| Build a sentence from phrases | 8 | 0 | 10 | 10 |
| Answer opinion questions | (2) | 0 | 0 | (2) |
| Retell spoken passages | 2 | (2) | (3) | 3 |
| **Total items presented** | **58** | **50** | **69** | **63** |
| *Average test duration* | *15 minutes* | *12 minutes* | *17 minutes* | *14 minutes* |

while the others are published by Pearson under the brand name *Versant*™. These automated tests share many aspects of task design and scoring technology, and all purport to measure the construct *facility* in the spoken language. *Facility* can be defined as *the ability to understand the spoken language on everyday topics and to speak appropriately in response at a native-like conversational pace in an intelligible form of the language*. 'Appropriate' in this context means the response is one of the likely forms observed from natives and/or high-proficiency non-natives that also is judged 'correct' by experts.

The facility construct focuses on a person's ability to perform accurately with the language, receptively and productively, and to do so consistent with the pacing of a usual conversation. If a person can extract the basic meanings from words in reasonably authentic spoken sentences and produce appropriate intelligible spoken language in the time-frame that is observed in native-paced conversation, then that person can be said to have *facility* with the language. The facility construct is not directly related to the context or scoring of functional communication or extra-linguistic goals. Facility-in-L2 tests measure performance outside the setting of language in use for communicative tasks. The theoretical viewpoint compatible with the facility-in-L2 tests is that spoken language can be fruitfully conceived as a channel of communication or a tool for the transmission of information (e.g. Cherry, 1966), including the language forms that convey social and rhetorical information. That is, even though spoken skills develop in the context of functional communication, the skills develop in individuals and find expression in activities that are not communicative (for example, anticipating the next word or phrase, reciting poetry, or understanding word play).

## *Structure of facility-with-L2 tests*

All four tests described here (of English, Spanish, Dutch and Arabic) are similar in structure, function and scoring logic. They all rely on sentence repetition and short-answer questions as main item types, with other item types that vary from test to test. Table 1 presents the test structures.

During test administration, an automated system presents a series of recorded spoken prompts in the L2 at a conversational pace and elicits spoken responses in the L2.

Administration of a test over the telephone or via a computer generally takes about 12–18 minutes depending on the tasks included. Spoken responses from the test-taker are analyzed automatically by computer. The scored responses provide multiple, independent measures that underlie facility with the spoken L2, including phonological fluency, sentence construction and comprehension, passive and active vocabulary use, listening skill, and pronunciation of rhythmic and segmental units.

We now describe the basic tasks, sentence repetition and short answer questions, that are shared by all four facility-in-L2 tests.

In the sentence repetition task, recorded sentences are presented to test-takers in an approximate order of increasing difficulty, and test-takers repeat the sentence aloud verbatim. Sentences range in length from three words in two syllables (e.g. English 'I'm here.') to twelve or more words in more than 20 syllables (e.g. Spanish 'Parece mentira que todavía le tengas confianza, después de lo que pasó.' [English: It's amazing that you still have confidence in it, after what happened.]). As a rule of thumb, if fewer than 90% of a native sample can repeat the item verbatim, then the item is not used in the test. As much as feasible, the sentence texts are adapted from ambient conversation or transcriptions of spontaneous native conversation in L2.

In the short answer question task, the test-taker is required to answer questions with one word or a simple phrase. The questions generally include at least three or four content words. Each question asks for basic information, or asks for a simple inference based on time, sequence, number, lexical content, or logic. The questions are designed not to presume any special knowledge of L2 culture, or other specific topic. Examples in English include 'What is frozen water called?' and 'Which season comes between summer and winter?' and 'Either Larry or Susan had to go, but Susan couldn't. Who went?'

Descriptions of the other item types (readings, sentence builds, opposites, passage retellings, and open questions) are available (Pearson, 2009b). Items are restricted to the core vocabulary of the language (usually the most frequent 4000–8000 word stems in L2), and the general requirement that items refer to common objects and events and should be easy for almost all native-speaker test-takers. The voices that present the item prompts belong to native speakers or to high proficiency non-native speakers (depending on the test), providing a range of accents, speeds, and speaking styles. This has several advantages. First, the test-taker may be presented with 8 or 10 different voices in the item prompts, offering a sample of accents and speaking styles. Second, clarity of enunciation and speaking rate in prompts is varied and contribute to item difficulty. Last, because statistical properties of observed responses from native speakers and L2 learners are known, they inform a more precise scoring.

In common with Fulcher's (2000) characteristics of communicative tests, the two basic facility tasks use unpredictable and fairly authentic input to elicit speaking performances; however, unlike communicative tests, there are no extra-linguistic functional goals to score and no multi-turn sequences. A well-conducted OPI simulates social and interactive aspects of natural speech communication, but the specific L2 forms and styles of speaking used by the examiner(s) are not strictly controlled. In contrast, facility-in-L2 tests do not simulate natural speech communication, but rather exercise the core skills that are the building block of any conversation.

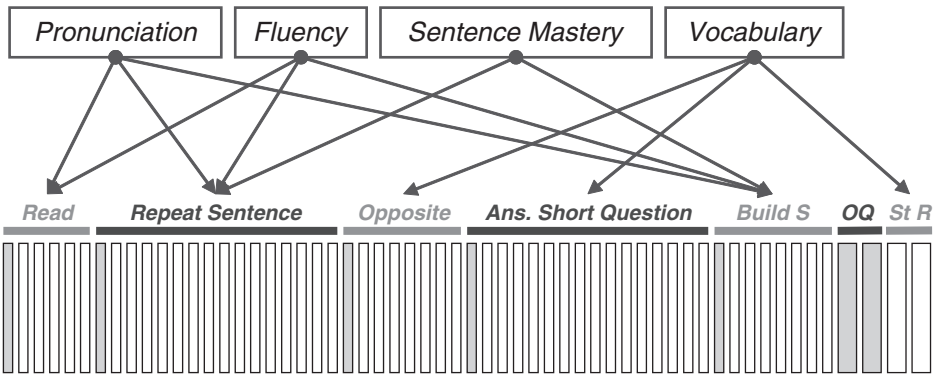


**Figure 2.** Mapping of subscores to item types in the Versant Spanish Test (adapted from Pearson, 2009b). (*Build S* is Sentence Build, *OQ* is Open Question, *St R* is Story Retelling.) Vertical rectangles represent recorded responses; unscored responses are grey. Summed duration of scored speech samples is about 5.2 minutes for the average VST test-taker.

## *Scoring method and construct representation*

In this section, we explain how *facility* with L2 contributes directly to test scores, that is, how the construct is represented in the tasks and how it is measured in the scoring of test-takers. The Versant-branded facility-in-L2 tests report four subscores, Sentence Mastery, Vocabulary, Fluency, and Pronunciation, averaged into an Overall score with weights of 30, 20, 30, and 20 respectively. Their subscores and overall score are reported on a scale 20–80. There is no listening score, as such. However, as every item requires listening skill for successful completion, Versant scores are conceived to be listening-speaking scores, consistent with the facility construct. We describe here the scoring of the Sentence Repetitions and Short Answer Questions in the Versant Spanish Test (VST). Figure 2 illustrates which sections of the VST test contribute to which subscores. The VST manual (Pearson, 2009b) defines the subscores in more detail.[1]

*Scoring.* Of the four subscores, Sentence Mastery and Vocabulary measure the linguistic content of the response, and Fluency and Pronunciation measure the manner in which the response was said. Accurate content indicates how well the test-taker understood the prompt and could respond with appropriate linguistic content. Manner scores indicate how faithfully the test-taker approximates the articulation and rhythm of native speakers (or favorably judged non-natives) according to a statistical model based on field test data (see Bernstein and Cheng, 2007). All four tests discussed here use an augmented automatic speech recognition (ASR) system based on HTK (Young et al., 2000) that has been optimized for accuracy with non-native speech. This augmented ASR system provides lexical units and spectral measures used in scoring as well as a time-aligned transcription of the response, locating structures down to sub-phonemic details like consonant closures and aspiration that can be further analyzed.

*Sentence Mastery scoring.* A scoring engine calculates the Sentence Mastery subscore from the word strings that are produced by the augmented speech recognizer. For sentence repetition items, the recognized string is compared to the word string recited in the prompt, and the number of word errors is calculated as the minimum number of substitutions, deletions, and/or insertions required to find a best string match in the response. This matching algorithm ignores hesitations and filled or unfilled pauses, as well as any leading or trailing material in the response. An example of the error count calculation follows:

[Prompt] Larry took down five, but one at a time.

[Response] *uh Larry took now five one time I think yes.*

The leading 'uh' and the trailing 'I think, yes.' are ignored, and the remaining underlined part of the response can then be optimally modified to match the prompt string by substituting 'down' for 'now' (1), then inserting 'but' (2), 'at' (3), and 'a' (4) in the correct positions. This response therefore is assigned four word errors. A verbatim repetition would have zero word errors. Item responses with different numbers of word errors are scored according to a partial credit Rasch model (Linacre, 2003).

The sentence-build items are scored the same way as the repetition items and located on a common IRT-based item-with-errors difficulty scale with the repeat items. A test-taker's Sentence Mastery subscore is derived from the words (mean 135) that are heard and repeated successfully (or not) in the 22 scored repeat or sentence build items in a particular test administration. The estimate of Sentence Mastery is then scaled to subtend the 20–80 reporting scale. Henning (1983) reports that a simpler form of sentence repetition scoring, with human item presentation, was an excellent predictor of ability in spoken English. Vinther (2002) reviews similar results relating sentence repetition accuracy to oral proficiency tests.

*Vocabulary scoring.* In the VST, Vocabulary scores are extracted from the test-takers' performance on the short answer, opposites, and story retelling items. For the short answer and opposite item types, the scoring uses a dichotomous IRT model based on the occurrence of key words or sequences that constitute correct answers. Correct answers are determined by expert judgment, and extended to include reasonable answers observed in the responses from native speakers and high proficiency non-native speakers. For example, in the item that presents a recording of the word 'right' and requires the test-taker to say the opposite, the experts posited 'left' and 'wrong', but the response 'read' (i.e. the opposite of 'write') was also common among high proficiency non-natives and was added to the list of correct answers for that item.

Story retelling items are scored for vocabulary by scaling the weighted sum of the occurrence of a large set of expected words and word sequences that may be recognized in the spoken response. Weights are assigned to the expected words and word sequences according to their semantic relation to the story prompt text using a method similar to latent semantic analysis (Landauer et al., 1998). These scores also contribute to the VST vocabulary score.

*Fluency and pronunciation scoring.* The manner-of-speech scores are derived from a large set of base physical measures (e.g. durations of segments, syllables, and silences, and spectral properties of segments and subsegments) that are extracted during the recognition of the item responses. Words in responses are identified using acoustic models that are trained exclusively on non-native speech; then the identified word string is re-aligned to the response using native acoustic models. For all four languages discussed here, the native models are trained on a wide sample of speakers of different ages and a range of native accents, approximately balanced for gender.

The base physical measures are then scaled and transformed into fluency and pronunciation scores by a two-step process. First, the physical measures are scaled according to their likelihood in a native-speaker model of the relevant context-dependent linguistic unit. For example, if the test-taker pauses for 500 milliseconds in a position where it is not unlikely that a native might pause that long, the algorithm scales that pause as relatively fluent. Similarly for spectral properties of segmental units, if a spectrum sampled late in the first [t] of 'potato' is very unlikely in a model of the distribution of late spectra in native productions of [t] before stressed front vowels, then that spectrum is scaled as a relatively poor pronunciation. Thus, the first step effectively defines both the rhythmic and segmental aspects of the performance to be native likelihoods of producing the observed base physical measures. Then, in the second step, the scaled values derived from a training set of test-taker responses are optimized with parameter weights in a non-linear combination formula to match a set of human judgments of the fluency and pronunciation abilities of these test-takers.

Both the Fluency and Pronunciation scores are derived from the same responses to repetition items (and sentence build and reading items), but they are derived from different sets of physical parameters. The Fluency score is developed from durations of events, (e.g. response latency, words per time, segments per articulation time, inter-word times) and combinations of theses durations. Rosenfeld et al. (2003) present more detail on these parameters. The Pronunciation score is developed from segmental properties including the context-independent and context-specific spectral likelihoods according to native and learner segment models applied to the recognition alignment. Pronunciation scoring in the facility-in-L2 tests is similar in logic and result to that reported by Bernstein et al. (1990), and Franco et al. (2010).

*Representation of the facility construct in the scoring methods.* As arc 2 in Figure 1 suggests, it is necessary to provide a theoretical or evidential claim which explains how the construct measured is reflected in test scores. As the facility construct is narrow and explicit, the evidence for its representation in the test scoring can be quite specific. The construct can be analyzed into four assertions about a person's abilities to process and produce spoken language *in real time*.

1. A person extracts words and phrases from the incoming speech stream.
2. A person demonstrates understanding by producing appropriate and timely spoken responses.
3. A person speaks with a latency and at a rate (local and global) that is similar to that of natives and high-proficiency non-natives.
4. A person pronounces words and clauses in a manner that is similar to that of natives and high-proficiency non-natives.

Abilities 1 and 2 are straightforward inferences from the facility construct as stated (in slightly varying forms) in the documentation of the facility-in-L2 tests. Abilities 3 and 4 are related to the construct definition by the assumption that measurable similarity to likely response forms (in latency, timing, and spectral properties) produced by known high-proficiency speakers is a strong predictor of ability to speak intelligibly in a way that is suitable for 'native-pace' conversational interaction.

Simply put, these four abilities are what test-takers demonstrate in order to score highly on the facility-in-L2 tests. In all tasks, the algorithms that measure the appropriateness of the answers have been scaled with reference to a set of field test responses, and then validated with unseen data. That is, scoring algorithms are developed using a large set of spoken test responses (the training data), then the algorithms are validated on a separate set of test responses from a distinct group of speakers that has been set aside for that purpose.

From a design point of view, the scoring algorithms used to estimate facility constitute an empirically derived and verifiable model of listening and speaking performance. Further, the model embodies a hypothesis about the development of real-time speaking and listening skills at different levels of L2 proficiency: it makes definite predictions about new speech samples from L2 speakers in relation to their judged proficiency and therefore has the advantage of being falsifiable. It may be seen as a data-driven hypothesis that aligns with the theories propounded by Hulstijn (2006) discussed below.

## *Concurrent validation data*

The comparison of automated test scores with scores from concurrent administrations of well-established tests with a similar purpose or related construct can be an important element in the validation of the automated tests. Because the purpose of the new test is to estimate the ability of the test-taker rather than to predict one score on one administration of a concurrent test, it is prudent to administer both tests more than once and in circumstances conducive to score independence. Such procedures also allow us to estimate the reliability of the scores on the concurrent test and the relative size of the variance in all the test scores associated with different sources.

In this section, we summarize the results of five experiments that compared scores from facility-in-L2 tests with scores from concurrent tests that were human-administered and human-scored. The five experiments include one each with Spanish, Dutch, and Arabic, and two experiments with English. In three of the experiments, both the automated tests and the human tests were administered twice. In one experiment, there was only one administration of the concurrent human test but the two interviewers submitted independent ratings of the test-taker's performance. In another experiment, the automated test was administered once, returning one score, while the human test was administered twice with double ratings for each administration. In the automated tests, items presented to the test-takers were randomly generated in each administration from relatively large item pools. For all experiments, when two raters were reporting scores based on the same performance, raters were instructed to report a score without consultation with the other rater. Table 2 summarizes the five data sets; more specific descriptions of the experiments are provided below and in the citations listed in the table.

**Table 2.** Summary characteristics of the data sets 1 through 5

| Data Set | Number of test-takers | Test-taker sample | Automated test | Automated test administrations | Human test | Human Test administrations (total human scores) | Observed human test score range | Correlation | Citation |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 37 | Convenience sample | Versant Spanish | 1 | ILR-SPT interviews | 1 (2) | ILR 0+ to 4 | 0.92 | Balogh & Bernstein (2007) |
| 2 | 228 | Netherlands immigrants, near CEFR A1 | Toets Gesproken Nederlands (TGN) | 2 | CEFR OPI and career interviews | 2 (3) | CEFR A1− and A1 | 81% classification consistency | De Jong et alia. (2009) |
| 3 | 118 | Convenience sample | Versant Arabic | 2 | ILR-OPI interviews | 2 (4) | ILR 0 to 4 | 0.87 | Ordinate (2008) |
| 4 | 151 | Students in Adult Education ESL Classes | Versant English | 2 | Best+ tests | 2 (2) | BEST Plus 337 to 961 | 0.81–0.86 | Present-Thomas & Van Moere (2009) |
| 5 | 130 | TOEFL test-takers in Iran | Versant English | 1 | IELTS interviews | 2 (4) | IELTS 2.75 to 9 | 0.77 | Farhady (2008) |

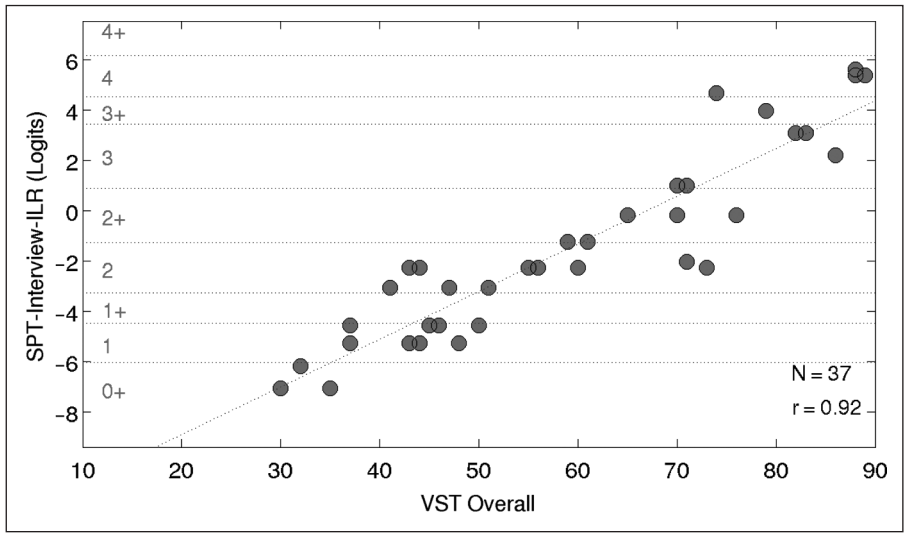


**Figure 3.** SPT interview scores versus Versant Spanish Test scores (r = 0.92; N = 37)

The goal was to see how well the test-takers' ability as estimated on the facility-in-L2 tests corresponds to their ability as estimated on human tests. For this reason, whenever possible, all available human judgments (either two or four judgments) were combined to yield the best estimate of the test-takers' ability level as measured in the human test. The following paragraphs review the concurrent experiments.

*Data Set 1.* In this study (Balogh and Bernstein, 2007), a sample of 37 test-takers took one telephone administered Versant Spanish Test (VST) and one telephone administered interview with government-certified raters in accordance with the Spoken Proficiency Test (SPT) procedure, with scores reported on the Interagency Language Roundtable (ILR) scale. Each SPT interviewer-rater independently provided ILR-based ratings for each of the 37 candidates, for a total of 74 ratings. All test-takers participated in the interview within one day of the Versant Spanish Test administration. Figure 3 shows a scatter plot of ratings for 37 non-native Spanish speakers for the automated and human measures. The y-axis shows how Rasch logits map to ILR levels, revealing that the sample of test-takers distributed across ILR levels 0 through 4. The data in Figure 3 is for an early form of the VST that did not score story retelling responses. The split-half reliability of the VST scores for this sample was r = 0.97. The inter-rater reliability of the SPT was r = 0.93. The correlation between the Versant Spanish Test scores and the ratings from the SPT Interview was r = 0.92, with no obvious outliers.

*Data Set 2.* In this study (De Jong et al., 2009), a sample of 228 test-takers took the TGN automated test of spoken Dutch twice in succession and participated in two independent interviews. The test-takers were all learners of Dutch thought to be at or below the CEFR A1 level. The comparison was conducted to understand the relative accuracy of the

**Table 3.** Agreement on A1-minus cut-off: human–human, human–machine, machine–machine (adapted from Table 8, p. 58, in de Jong et al., 2009)

| Decision 1 | Decision 2 | 1–1 | 1–0 | 0–1 | 0–0 | **Agreement** |
|---|---|---|---|---|---|---|
| CEFR Interviewer (OPI performance) | CEFR Rater (Career interview) | 64% | 10% | 9% | 17% | **81%** |
| CEFR Observer (OPI performance) | CEFR Rater (Career interview) | 64% | 9% | 11% | 16% | **80%** |
| Human–Max. (of 3 decisions) | TGN–Max. (of 2 decisions) | 70% | 14% | 7% | 8% | **78%** |
| TGN-1 | TGN-2 | 58% | 7% | 12% | 23% | **81%** |

automated TGN test and human interviews in making a pass-fail decision with early-stage language learners. One interview was an OPI-style language interview and one was an independently administered career interview, but both were scored on a pass-fail basis according to a cut-off point at 'A1-minus' on a downward extension of the standard CEFR scale. The OPI-style test yielded two independent ratings – one from the interviewer and one from an observer, and then another rater independently rated the career interview as either above or below A1-minus. This procedure resulted in three independent human ratings of each test-taker's Dutch ability: two from the OPI and one from the career interview. The test-retest reliability of the TGN was r = 0.78, which is reasonably high considering that the sample was intentionally selected to be near one classification boundary. Considering the high-stakes use of the TGN, this reliability is not as high as one would like. Note, however, that the classification consistency between the two human interviews for this narrow sample was 81% and was also 81% between two administrations of the TGN. Table 3 presents the percentage agreement for A1-minus decisions for pairs of decisions wherein the decisions in a pair are based on different performances.

Noting the pattern of agreements in Table 3 (their Table 8), De Jong et al. (2009, p. 59) conclude, in part, that these 'two very different' spoken language tests (one fully automatic and the other based on human ratings of a human interaction) 'show the same degree of overlapping decisions as two sets of human ratings do.'

Although the above concurrent study sampled only low-level learners to establish pass-fail consistency, a separate benchmarking study used a different sample of Dutch speakers exhibiting performances over the full range of CEFR levels. This study involved 14 raters assigning a total of 5,984 ratings to 1,009 test-takers who performed on open question and story-retell tasks. De Jong et al. (2009) report that the TGN scale covered the whole range (A1-minus to C2) permitting the establishment of cut-offs at each of the CEFR levels.

*Data Set 3.* In this study (Pearson, 2008) a sample of 118 test-takers (112 learners and 6 native speakers, representing a total of 14 different language backgrounds) took two forms of the automated Versant Arabic Test (VAT) and two Arabic OPI tests within a 15-day window. Both tests are designed to measure ability in spoken Modern Standard Arabic, a non-colloquial form of Arabic; the spoken form of which is often used in

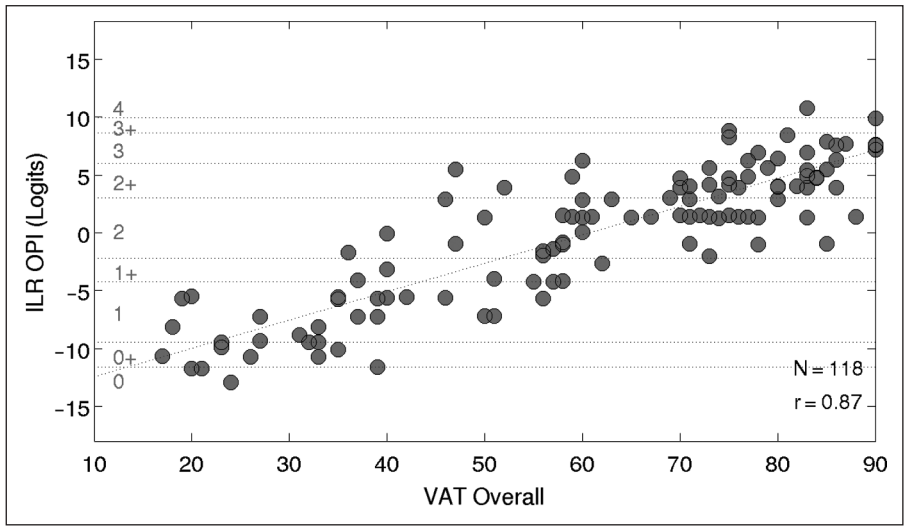


**Figure 4.** Test-takers' ILR OPI scores as a function of VAT scores (r = 0.87; N = 118)

international broadcasting. Seven active U.S. government-certified oral proficiency interviewer-raters conducted the ILR OPIs over the telephone, following the official ILR procedures. The Versant Arabic Tests were administered by computer or over the telephone. The test-retest reliability of the Versant Arabic Test for this sample of test-takers was r = 0.93. The test-retest reliability of the OPI was r = 0.91. When comparing each rater's individual scores with the average score from the other interview, correlation coefficients ranged from r = 0.86 to 0.93. The four independent OPI ratings were subject to Rasch-scaling using the computer program FACETS (Linacre, 2003), where the value of a '+' rating was modeled as 0.5 (thus, '1+' is entered as 1.5 in the Rasch modeling). For the human scores displayed in Figure 4, the boundaries of the different ILR levels were mapped onto a continuous logit scale. This is shown on the y-axis, revealing that the sample distributed across ILR levels 0 through 3+. Figure 4 is a scatter plot of the ILR OPI ability estimates as a function of VAT scores for the sample of 118 test-takers. (The wide range of logits can be explained by unusually high agreement among raters, which, together with four observations per candidate, has resulted in very little randomness in the data. The data have an almost deterministic-Guttman pattern.)

The correlation between these two sets of test scores is 0.87, indicating a reasonably close relation between machine-generated Arabic scores and scaled scores from the human-rated interviews. Note that this correlation of 0.87 is in the range of those observed between a single certified ILR OPI interviewer and the average of two others conducting another independent interview. Figure 5 plots randomly selected single-interviewer scores for interview 1 versus interview 2 for each of the 118 test-takers for whom two double-rater interviews were available. Scores are given in their OPI values, but displayed (with added jitter) at their nominal values on a logit scale.

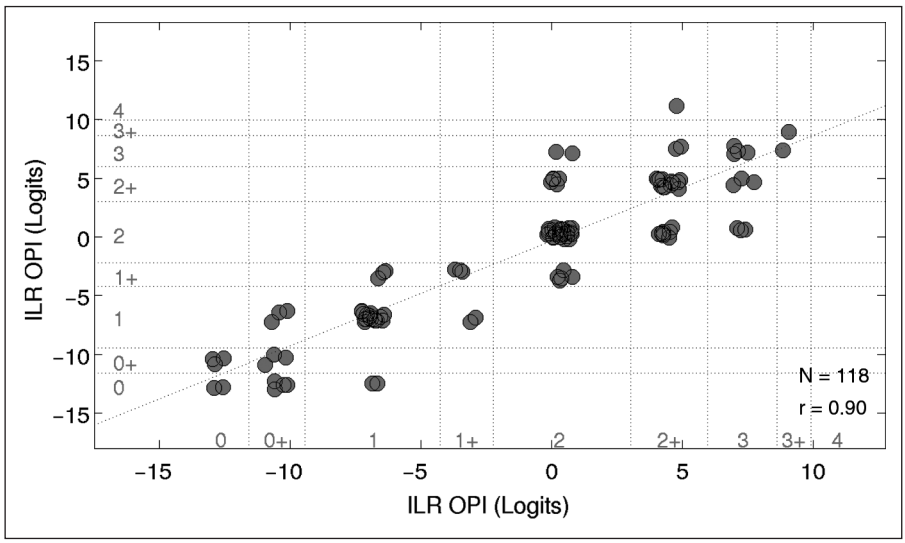


**Figure 5.** Single-rater ILR scores for test-takers in one OPI interview as a function of a single rater on a different interview. Scores are presented as ILR values (0, 0+, 1, 1+, …) positioned on a logit scale, with small jitter added to each score to reveal multiple points with the same values from the two raters (r = 0.90; N = 118)

*Data Set 4.* In this study (Present-Thomas and Van Moere, 2009), a sample of 151 test-takers completed four tests within 24 hours: two BEST-Plus tests and two administrations of the Versant English Test (VET). The BEST-Plus is a scripted oral interview (CAL, 2005) that lasts anywhere from 3 to 20 minutes, depending on the test-taker's proficiency level. The BEST-Plus interviews were conducted by qualified examiners, trained and certified by the Center for Applied Linguistics. Observed scores range from 337 to 961 on BEST-Plus and from 20 to 80 on the VET. The 151 test-takers were sampled from adult ESL classes with an expectation that most would perform at a relatively low level in spoken English. In fact, the sample of test-takers spanned the full score range on both BEST-Plus and Versant, but were over-represented at low-intermediate levels. The sample included 25 different language backgrounds with ages from 18 to 79.

Figure 6 shows two scatterplots of the average score on the two administrations of the BEST-Plus test – first as a function of the scores from the first administration of the automated test (VET1) and then as a function of the second administration (VET2). The VET1 test was a non-standard version that was modified not to present a portion of the difficult items. Scores were analyzed to gauge agreement between human and machine tests in relation to test reliability. The test-retest reliability for this sample of test-takers was r = 0.93 for the Versant English Test and r = 0.86 for the BEST-Plus. As Table 4 shows, the Versant test scores predict the average BEST-Plus score as accurately as one BEST-Plus score predicts a second BEST-Plus score. The data are not normally distributed; the Spearman's rho coefficients are higher (VET1-BEST-Plus: 0.87, and VET2-BEST-Plus: 0.88).

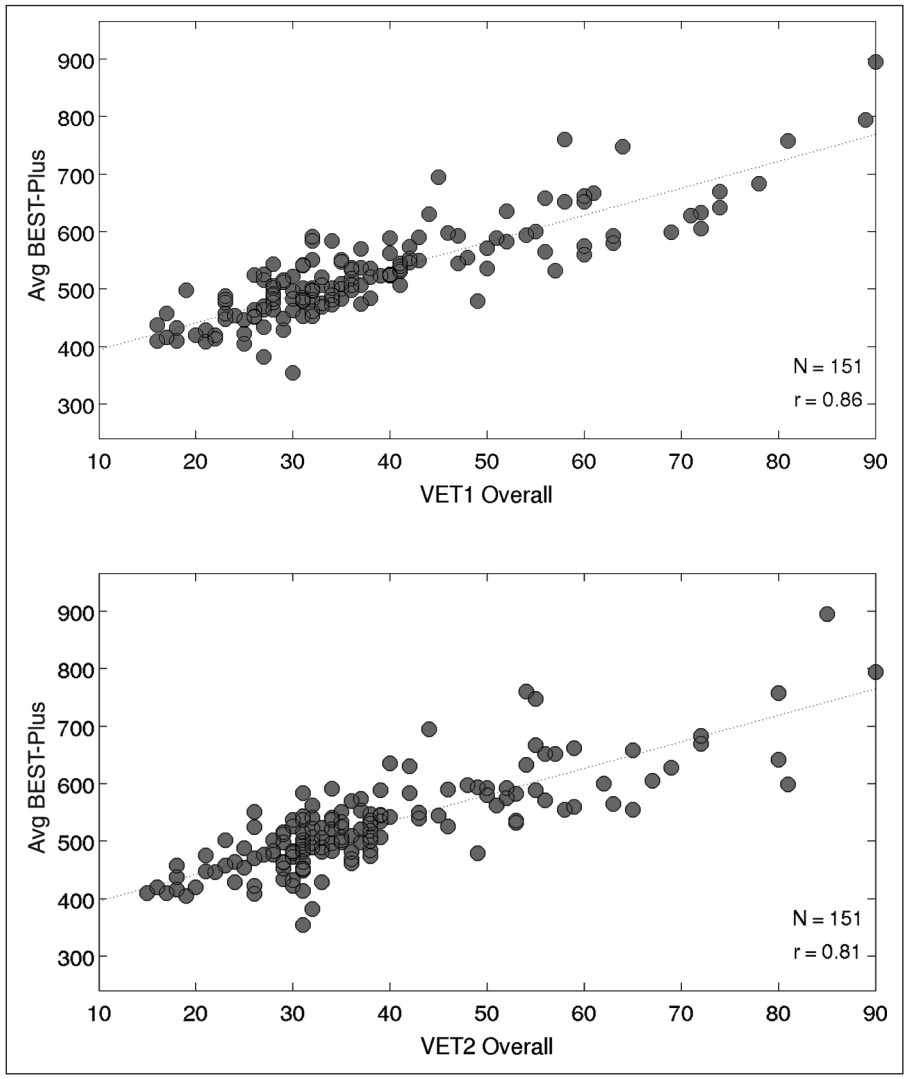


**Figure 6.** Average of two BEST-Plus scores as a function of Versant English Test (VET) score. VET1: r = 0.86, VET2: r = 0.81, N = 151

*Data Set 5.* In this study (Farhady, 2008), 130 English learners in Iran who had signed up to take the official TOEFL iBT test, were invited to take one Versant English Test and two interview speaking tests based closely on the IELTS procedure and administered by active IELTS examiners. The participants represented various language backgrounds including Persian, Azari, Kurdish, Balouchi, and Guilaki. For each of the 130 test-takers, there are four test scores of interest: one Versant English score, one official TOEFL iBT score, and two IELTS-like interview scores. In the speaking section of the

**Table 4.** Spearman rho correlations between administrations of the Versant English Test and the BEST Plus

| Test | Versant English 1 | Versant English 2 | BEST Plus 1 | BEST Plus 2 | BESTPlus Average |
|---|---|---|---|---|---|
| Versant English 1 | | | | | |
| Versant English 2 | 0.93 | | | | |
| BEST Plus 1 | 0.85 | 0.80 | | | |
| BEST Plus 2 | 0.83 | 0.81 | 0.86 | | |
| BEST Plus Average | 0.87 | 0.88 | – | – | |

iBT, six speaking items are presented by computer and the spoken response recordings are scored by human listeners. In the IELTS-style interviews, test-takers answered questions, took a long speaking turn, and participated in a discussion. Interviews were conducted by a total of four raters; in each of the two interviews, one rater acted as the interlocutor and the other just rated. Observed average scores range from 2.75 to 9.00 on the IELTS scale and from 30 to 80 on the VET. IELTS speaking scores reported here are the average of the two ratings, one from the interviewer and one from an observer/rater.

For the Iranian sample, the split-half reliability of the Versant English Test was r = 0.93. The inter-rater reliability for ratings of the same performance in the IELTS-style interviews was between r = 0.77 and r = 0.79. The reliability of the TOEFL iBT speaking scores for this sample is unknown, but the score distribution was normal: on a scale of 0–30 the scores ranged from 10 through 30, with a mean of 19.8 and standard deviation of 3.8. The correlations among the four test scores for this set of 130 Iranian test-takers are presented in Table 5. The highest coefficient in the table is the first and second IELTS-style interviews (r = 0.88). The next highest correlations are the Versant English Test with the IELTS-style interviews (r = 0.77) and the Versant English Test with the TOEFL iBT speaking scores (r = 0.75).

*General observations on these data sets.* All the automatic facility-in-L2 tests under study in these experiments have high reliability (r = 0.93 to 0.96, depending on the experiment). The human rated tests also have generally high reliability (r = 0.86 to 0.93). In the Dutch experiment, the automatic TGN test agreed with a careful human decision 78% of the time, while the human decisions agreed for a given test-taker 81% of the time. In all four

**Table 5.** Correlations between score-pairs in four sets of test scores

| | Versant English | TOEFL iBT Speaking | IELTS-style Interview 1 | IELTS-style Interview 2 |
|---|---|---|---|---|
| Versant English Test | | | | |
| TOEFL iBT Speaking | 0.75 | | | |
| IELTS-style Interview 1 | 0.77 | 0.71 | | |
| IELTS-style Interview 2 | 0.71 | 0.72 | 0.88 | |
| IELTS-style Average | 0.77 | 0.73 | – | – |

cases with continuous data, the automatic test score accounted for a major portion of the reliable variance in the human rated scores from the sample of test-takers under study.

Looking for counter-examples, visual inspection of the scatter plots reveals cases where, despite a high overall correlation coefficient, a single position on the automated test score axis describes test-takers of widely varying OPI score values. Data Set 3, for example, shows test-takers who score comparably in the automated test yet who have scores on the ILR scale varying from 1 to 2+ (Figure 4). This is likely due to two main factors. The first factor is measurement error. Part of it is exhibited in the test-retest scatter plots of the Arabic OPI (Figure 5), which shows that measurement of individuals varies considerably even across occasions of the same test. (Note, however, that the test occasions are assumed to tap the same construct because the test tasks are consistent and the scores exhibit sufficient mutual variance). The second factor is that facility-in-L2 and OPI tests genuinely tap somewhat different constructs due to test methods and task design. Nevertheless, it is argued that as the facility-in-L2 tests measure core skills that are building blocks of speaking proficiency, the relationship between Versant scores and interview test scores is attributable to a common construct underlying speaking proficiency. When concurrent studies are undertaken on other tests purporting the same or similar constructs, considerable cross-dimensional spread is often evident. The TOEFL and IELTS both purport to assess language skills within the academic domain and can be used for university admissions decision-making, yet they correlate at anywhere between 0.67 and 0.83 (Geranpayeh, 1994); lower than the coefficients reported here.

Despite differences in construct, interactivity, and complexity of spoken material, the automatic test scores generally cover abilities ranging from 0 to 3+ on the ILR scale (the full scale being 0–5), and account for most of the reliable variance that is found in the interview test scores. The scores from the two kinds of test are strongly related across five experiments in four languages with various populations and score-use contexts.

## *Discussion*

Several kinds of evidence support an interpretive argument for the valid use of scores from a fully automated spoken language test. As shown in Figure 1, these are: a construct definition for an automated test (facility-in-L2); a theoretical rationale for linking test-taker ability to test scores, and psychometric evidence for test score consistency (arc 2); concurrent data relating automated test scores to communicative tests (arc 3); and a hypothesized mediated link between a facility-in-L2 test and the target domain (arc 1). The relation between OPI tests and the target domain (arc 4) is posited, but not directly addressed.

This line of argument should not be construed as: ‘the test scores correlate reasonably highly with the OPI scores, therefore the test is valid.’ Rather, the argument asserts:

**if** other excellent tests of L2 oral skills are predictive of behavior in the real world and are valid for score-based inferences, and
**if** the facility-in-L2 test correlates with these tests as well as these different tests correlate with each other, and
**if** facility-in-L2 tests also meet reasonable theoretic and psychometric criteria,
**then** there are grounds for a mediated link between the facility-in-L2 test scores and real world spoken performance in L2.

In this section, we address four counter-claims to the interpretive argument (space limits us to four). The *first counter-claim* is that validity can only be established with reference to test score use and decisions made on the basis of test scores, and not merely on the basis of consistently measuring test-takers according to a defined construct. Test selection should not only be a case of 'Will this test tell me who has better skills?', but rather, 'Who is the test intended for? What decisions will be made on the basis of scores? What is the domain that test language and tasks should represent?'

While some extrapolation and predictive validity data has been gathered (Suzuki et al., 2008) for facility-in-L2 tests, more such evidence can usefully be collected. However, since the domain – 'generally effective speech communication' – is unspecified, an interpretive argument is difficult to make based on extrapolation to that non-test domain, which is a problem for OPIs and facility-in-L2 tests alike.

Nevertheless, there are two rebuttals to the counter-claim that hinge on domain specificity or test score use. First, although language test designers prefer to specify a target domain, research has failed to establish the link between domain-specific tests as an exclusive predictor of domain-specific performance. Domain-specific testing should not only show divergent relationships with tests that sample other domains, but should also show divergent relation to real-world behavior in other domains. However, to our knowledge, no such evidence has been found. As Fulcher and Reiter (2003) argue, researchers in EAP have failed to isolate specificity of test tasks (Clapham, 2000), therefore we are not yet able to link task-specific performance with domain-specific performance in a way that also demonstrates non-domain divergence. One therefore can posit a general target domain which can be sampled by non-domain-specific tests as implied by the work of Hulstijn (2006) and Schoonen et al. (2009), who posit a core of language proficiency shared by adult native speakers, drawing on knowledge and skill. In Hulstijn's theory:

> For speech perception and speech production, knowledge refers to:
>
> 1. Speech sounds, phonemes, stress and intonation patterns
> 2. Lexicon (frequent items)
> 3. Morphosyntax (frequent structures).
>
> Skill refers to the ability to accurately online process phonetic, lexical and grammatical information receptively and productively in utterances that occur in any communicative situation, common to all native speakers (cleaners, car mechanics, and journalists). (Hulstijn, 2006, p. 17)

A second response to the first counter-claim hinges on proper score use. We assert that, in many cases, the automated test scores alone should not be the sole basis of decision-making, but rather facility scores are one piece of evidence about a candidate that would contribute to decision-making. For example, when deciding on entry into an academic course of study, decision-makers might need to know the candidate's facility with the spoken language, *in addition to* the candidate's grades, work habits, their ability to read and assimilate difficult text, and whether or not they have the presentation skills to give a reasoned argument in an academic setting. But since the automated test is easy to administer and reliable, it might be used as the first test in a series of listening and

speaking proficiency tests, where a low score would demonstrate that the candidate was below the essential threshold for core facility with the spoken L2, and thereby could be precluded from further testing.

We acknowledge that this approach to validity may be at odds with current thinking on the justification of test scores for a specific use, nevertheless, a test of 'generally effective speech communication' is useful in many contexts as one part of a test battery.

A *second counter-claim* is that the facility-in-L2 test lacks the functional, strategic, and complex language content of the target domain. This is important as content is an important aspect of validity (O'Sullivan et al., 2002). In other words, we would normally expect to see certain elements of real-life communication represented in the test, including functions such as persuading or hypothesizing, strategies such as interrupting or turn-taking, and complex linguistic units such as strings of sentences in coherent long turns. This is a compelling criticism for linguists grounded in authentic and domain-representative test methods, since the facility-in-L2 test does not elicit functional content or conversation strategies.

The primary response to the second counter-claim is that facility-in-L2 tests measure essential linguistic skills that operate on the smaller core elements used in all language situations and underlie the ability to perform all of the communicative functions. The facility-in-L2 test elicits densely sampled linguistic units at the sentence level and below. Selecting words for phrases, building phrases into clauses, and clauses into sentences, in real time, is the core skill in spoken language communication. On the one hand, functional content would not be possible without these core linguistic skills, and on the other hand, functional abilities are strongly predicted by these core linguistic skills. It is possible that some functional applications of language are borrowed from L1. Concerning the measurement of communication strategies such as turn-taking, much evidence suggests that these are, in any case, very difficult to elicit and measure in a standardized and consistent way in interview tests, since task and interlocutor variables play an important part in the demonstration of these skills (Brown, 2005). Research results suggest that when the strategic competences are measured, there is contamination from extraneous traits such as 'talkativeness' (Van Moere and Kobayashi, 2004).

A *third counter-claim* could be that the eliciting materials and expected responses in the facility-in-L2 tests are too short and too simple to measure the full range of oral proficiency. More exactly, the claim might be that because the facility-in-L2 scores are overwhelmingly based on responses of less than 10 or 12 words in length, higher levels of skill cannot be distinguished. A rebuttal to this claim is that the Standard Error of Measurement of the automatic scores is roughly constant from the lowest to the highest score on the scale (SEM $\approx$ 3 points on the 20–80 scale of the Versant English, Spanish, and Arabic tests) and that where there is available data the facility scores continue to align closely to human-test scores up to ILR level 3 or even 3+ (see Figures 3 and 4). This data suggest that the materials may not be too short to discriminate L2 speakers up to the high-intermediate level.

For the Fluency scores, in particular, there is a more interesting rebuttal to the 'too short' claim. Rubin (2009) reports that the most potent suprasegmental parameter predicting oral proficiency is intra-run fluency. That is, when combining base measures like speaking rate and articulation rate to estimate fluency, it is the fluency of small,

inter-silence bursts of speech that yield the most information about a speaker's oral proficiency. As these runs are typically much less than 10 or 12 words long, it would seem that the facility-in-L2 item materials are of sufficient length. Furthermore, observed articulation rates are quite distinct for different linguistic content (texts) that are spoken. For this reason, the articulation rate can provide more useful information if the linguistic content is accurately known, and much more exact information if the distribution of rates for this content (over a range of candidates) is known ahead of time. These favorable conditions are met in scoring the sentence repetition and sentence build items in the facility-in-L2 tests.

The question still remains: is analysis of responses to informal item material on everyday topics too simple to differentiate performance at levels above basic? Schoonen et al. (2009) looked at elements of language skill and task to see which factors are useful in distinguishing candidates between the adjacent intermediate CEFR levels B1 and B2. Among Schoonen et al.'s variables, the best at distinguishing B1 and B2 speaking performance (at about 80% accuracy) are linguistic knowledge (such as vocabulary) and phonological fluency. After these, the next best distinguishing characteristics were sentence-construction latency and pronunciation – two aspects of speaking performance that are measured quite appropriately in sentence repetitions and sentence build tasks.

Over a range of task types, Schoonen et al. found that task characteristics such as language complexity, discourse type, and setting were not influential for distinguishing B1 and B2 speakers. In fact, these measurements were stable across tasks, and showed no systematic differences between task types. One of the best tasks for separating B1 and B2 candidates was a low-formality, low-complexity description, rated by experts as a CEFR level A2 task. Thus, empirical data suggests that the facility-in-L2 tasks may yield information at levels at least up to high-intermediate (CEFR B2 or ILR 3).

The *fourth counter-claim* relates to the washback on teaching and learning that might occur as a result of sentence-repetition tasks being used in high-stakes testing. It is possible that the automated tests are susceptible to off-construct coaching. As yet, there is no data to support or refute the fourth counterclaim, although the automated test providers welcome such research.

In conclusion, the scoring of the facility-in-L2 tests has properties that no practical human-scored test has. That is, the facility scoring implements an empirically derived quantitative model of listening and speaking performance at different levels of L2 proficiency. This performance model is by no means complete, and facility has yet to be reconciled within a model of communicative competence, but it is grounded in psycholinguistic research. The model operates upon linguistic units at the sentence level and below – the units that are best understood and most densely exhibited in real language use. The psychometric data suggest that facility is an important component of effective speech communication or oral proficiency. As such, these tests offer a starting point for a generation of more accurate measurement methods.

## Acknowledgements

The authors acknowledge that they work for the publisher of the assessment instruments described in this paper. We gratefully acknowledge the key role of Brent Townsend in the development of the instruments, and the role of John de Jong in framing some validity issues. We also thank the guest editor for her helpful comments on drafts of this manuscript.

## Note

1. *Sentence Mastery:* Sentence Mastery reflects the ability to understand, recall, and produce Spanish phrases and clauses in complete sentences. Performance depends on accurate syntactic processing and appropriate usage of words, phrases, and clauses in meaningful sentence structures.

   *Vocabulary:* Vocabulary reflects the ability to understand common everyday words spoken in sentence context and to produce such words as needed. Performance depends on familiarity with the form and meaning of everyday words and their use in connected speech.

   *Fluency:* Fluency reflects the rhythm, phrasing and timing evident in constructing, reading and repeating sentences.

   *Pronunciation:* Pronunciation reflects the ability to produce consonants, vowels, and stress in a native-like manner in sentence context. Performance depends on knowledge of the phonological structure of everyday words as they occur in phrasal context.

## References

Balogh J, Bernstein J (2007). Workable models of standard performance in English and Spanish. In Matsumoto Y, Oshima DY, Robinson OR, and Sells P (Eds.), *Diversity in language: Perspective and implications* (pp. 20–41). Stanford, CA: Center for the Study of Language and Information Publications.

Bernstein J, Cheng J (2007). Logic, operation, and validation of a spoken English test. In Holland VM, Fisher FP (Eds.), *The path of speech technologies in computer assisted language learning* (pp. 174–194). New York: Routledge.

Bernstein J, Franco H (1996). Speech recognition by computer. In Lass, N (Ed.), *Principles of experimental phonetics* (pp. 408–434). St. Louis, MO: Mosby.

Bernstein J, Cohen M, Murveit H, Rtischev D, and Weintraub M (1990). Automatic evaluation and training in English pronunciation. In *Proceedings of the ICSLP-90: 1990 International Conference on Spoken Language Processing* (pp. 1185–1188). Kobe, Japan.

Brown A (2005). Interviewer variability in oral proficiency interviews. *Language Testing and Evaluation 4*. Frankfurt: Peter Lang.

Center for Applied Linguistics (2005). *Technical report: Development of a computer-assisted assessment of oral proficiency for adult English language learners.* Washington, DC: Centre for Applied Linguistics.

Cherry C (1966). *On human communication* (2nd ed.). Cambridge, MA: MIT Press.

Clapham C (2000). Assessment for academic purposes: Where next? *System,* 28, 511–521.

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

Cronbach L (1988). Five perspectives on validation argument. In Wainer H, Braun H (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

De Jong JHAL, Lennig M, Kerkhoff A, and Poelmans P (2009). Development of a test of spoken Dutch for prospective immigrants. *Language Assessment Quarterly*, 6(1), 41–60.

Educational Testing Service (1982). *Oral proficiency testing manual*. Princeton, NJ: Educational Testing Service.

Farhady H (2008). Human operated, machine mediated, and automated tests of spoken English. Paper presented at the American Association of Applied Linguistics, Washington, DC.

Franco H, Bratt H, Rossier R, et al. (2010) EduSpeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing,* 27(3): 401–418.

Fulcher G (2000). The 'communicative' legacy in language testing. *System*, 28, 483–497.

Fulcher G, Reiter R (2003). Task difficulty in speaking tests. *Language Testing,* 20(3), 321–344.

Geranpayeh A (1994). Are score comparisons across language proficiency test batteries justified? A TOEFL-IELTS comparability study. *Edinburgh Working Papers in Applied Linguistics*, 5, 50–65.

Henning G (1983). Oral proficiency testing: Comparative validities of interview, imitation, and completion methods. *Language Learning,* 33(3), 315–332.

Hulstijn J (2006). Defining and measuring the construct of second/language proficiency. Plenary address at the American Association of Applied Linguistics (AAAL), Montreal.

Hulstijn JH (2007). Psycholinguistic perspectives on second language acquisition. In Cummins J, Davison C (Eds.), *The international handbook on English language teaching* (pp. 701–713). Norwell, MA: Springer.

Kane M (1992). An argument-based approach to validity. *Psychological Bulletin,* 112(3), 527–535.

Landauer TK, Foltz PW, and Laham D (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25, 259–284.

Linacre JM (2003). *Facets Rasch Measurement Computer Program*. Chicago, IL: Winsteps.com.

O'Sullivan B, Weir C, and Saville N (2002). Using observation checklists to validate speaking-test tasks. *Language Testing,* 19(1), 33–56.

Pearson (2008). *Versant Arabic Test: Test description and validation summary*. Pearson Knowledge Technologies, Palo Alto, California. Available online at www.ordinate.com/technology/VersantArabicTestValidation.pdf (accessed December 2009).

Pearson (2009a). *Official guide to Pearson Test of English Academic*. London: Longman.

Pearson (2009b). *Versant Spanish Test: Test description and validation summary*. Pearson Knowledge Technologies, Palo Alto, California. Available online at www.ordinate.com/technology/VersantSpanishTestValidation.pdf (accessed December 2009).

Present-Thomas R, Van Moere A (2009). NRS classification consistency of two spoken English Tests. Paper presented at the East Coast Organization of Language Testers Conference (ECOLT), Washington, DC.

Rosenfeld E, Massaro D and Bernstein J (2003). Automatic analysis of vocal manifestations of apparent mood or affect. *Proceedings of the 3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, Florence, Italy.

Rubin D, Kang O, and Pickering L (2009). Relative impact of rater characteristics versus speaker suprasegmental features on oral proficiency scores. Paper presented at the Language Research Testing Colloquium (LTRC), Denver, CO.

Schoonen R, De Jong N, Steinel M, Florijn M and Hulstijn J (2009). Profiles of Linguistic Ability at Different Levels of the European Framework: Can They Provide Transparency? Paper presented at the Language Research Testing Colloquium (LTRC), Denver, CO.

Shohamy E (1994). The validity of direct versus semi-direct oral tests. *Language Testing,* 11(2), 99–124.

Stansfield CW, Kenyon DM (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System,* 20(3), 347–364.

Suzuki M, Yokokawa H, and Van Moere A (2008). Effects of a short-term study abroad program on the development of L2 speaking skills. Paper presented at the American Association of Applied Linguistics (AAAL), Washington, DC.

Van Moere A, Kobayashi M (2004). Group oral testing: Does amount of output affect scores? Paper presented at Language Testing Forum (LTF), Lancaster University, UK.

Vinther T (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics,* 12(1), 54–73.

Young S (1996). Large vocabulary continuous speech recognition. *IEEE Signal Processing Magazine,* 13(5), 45–57.

Young S, Kershaw D, Odell J, Ollason D, Valtchev V, and Woodland P (2000). *The HTK Book Version 3.0*. Cambridge, UK: Cambridge University Press.

Zechner K, Higgins D, Xi X, and Williamson D (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication,* 51, 883–895.