Computerized assessment of second language speaking: A review of tests

Yujia ZHOU [周 育佳]
Tokyo University of Foreign Studies

Abstract
Testing second language speaking via computer is a fast-growing area in second language assessment. The purpose of this article is threefold: to familiarize readers with the state of computer technology as it can be applied to speaking tests by discussing three speaking tests: the speaking section of the Test of English as a Foreign Language Internet-based test (TOEFL iBT), the Versant English Test (VET), and the Basic English Skill Test Plus (BEST Plus); to summarize the evidence for these tests' validity for potential end-users to evaluate them; and to provide critical issues to consider for developing a computer-delivered speaking test and potential areas for future research.

1. Introduction
Computerized assessment of second language speaking is a fast-expanding area in second language assessment. Various attempts have been made to apply computer technology to the delivery of speaking tests.

In order to benefit readers who wish to advance their knowledge in testing speaking via computer, this article discusses three speaking tests that have successfully integrated computer technology into their test delivery, task selection and presentation, and scoring process, respectively. The selected tests are the speaking section of the Test of English as a Foreign Language Internet-based test (TOEFL iBT), the Versant English Test (VET), and the Basic English Skill Test Plus (BEST Plus).

The second purpose of this article is to summarize the validity evidence of the three tests. These tests may be used for various purposes, including placing learners in different levels, measuring learners' improvements, and collecting spoken data in research studies. It is thus critical for teachers and researchers to collect information in order to judge to what extent these tests are reliable and valid for their purposes.

Finally, with an increasing use of computer-delivered speaking tests, it is urgent to gather evidence to support the use of the tests. Based on the descriptive and evaluative review of the selected speaking tests, this article offers perspectives for future trends and future research avenues in testing speaking via computer.

In sum, this paper aims to answer the following questions:

(1) What are the recent developments of computer application in testing second language speaking as reflected in the TOEFL iBT, the VET, and the BEST Plus?
(2) What empirical evidence exists for the reliability and validity of the speaking section of the TOEFL iBT, the VET, and the BEST Plus?
(3) What are the critical issues to be considered for test development and future research in testing second language speaking via computer?

The rest of the article is presented in three major sections. The first introduces a validity framework to guide the discussion of validity issues on the speaking section of the TOEFL iBT, the VET, and the BEST Plus. The second focuses on the description of the key characteristics of the three tests and reviews available validity evidence. The last section is devoted to the future direction of test development and research.

2. Validity framework

Validity has been defined as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA/APA/NCME, 1999, p. 9). Although there has been agreement in considering validity to be a unified concept, various criteria for examining validity exist. Among the different frameworks that incorporate various taxonomies of validity evidence, this study adopts the one proposed by Messick (1995) and adapted by Miller and Linn (2000) to the performance test. This framework indicates that the validation of performance assessments should incorporate six aspects of validity: content, substantive, structural, generalizability, external, and consequential. We found this framework useful for organizing the validity issues concerning speaking assessment that belongs to a method of performance assessment. In the following, we briefly describe each of these aspects as well as their implications for evaluating the validity of inferences based on the scores of speaking tests.

2.1. Content

The content aspect of validity is related to the degree of the congruence of test tasks to the test specifications, and how well they represent the construct measured. According to Miller and Linn (2000), evidence for content validity could be gathered through content analysis in two phases. First, during initial task development phase, a blueprint with a theoretical base provides evidence of the content and skills being assessed. Second, after the tasks are constructed, empirical evidence is gathered from individuals not involved in the development process through interview or questionnaire.

2.2. Substantive

The substantive aspect is concerned with the process used by examinees when they respond to the test tasks and the consistency of those processes with the construct the assessment is intended to measure. Thus, two aspects of evidence are sought to demonstrate that tasks lead examinees to engage in the intended cognitive processes and that the influence of construct-irrelevant factors is minimized. For the first aspect, experts must review the tasks and scoring rubrics to judge whether scores are based on the successful completion of the intended process. Empirical evidence is also essential on the congruence of what examinees are actually engaged in when performing the test...
tasks with the intended process. This line of study could include analysis of the speech samples as well as the cognitive processes and strategies examinees use via think-aloud protocols. Second, to minimize construct-irrelevant variances, potential sources of these variances, particularly on the part of examinees, must be identified. For the computer-delivered speaking assessment, examinees’ reactions to the test and examinees’ characteristics, such as computer anxiety and computer familiarity, may be systematically related to test scores, and therefore should be investigated.

2.3. Structural

In general, structural validity is evaluated by investigating the internal structure of the test to ensure that the relationships among test items and sections correspond with the construct definition. In assessing speaking, this aspect mainly concerns about whether the scoring categories are consistent with the construct. The structural validity requires theoretical evidence that the scoring model is based on an existing knowledge base regarding the construct and empirical evidence that is highly dependent on scoring method. For analytic ratings, it is critical to ensure that rating categories measure related but distinct elements, and for holistic ratings, whether raters could cover all dimensions when issuing a holistic score is important. Further, since the scoring method is central to understanding the inferences that can be made from tests, empirical evidence that provides rationale for the selection of a certain scoring method is also required.

2.4. Generalizability

The generalizability aspect focuses on the replicability of test results across multiple levels of the assessment procedure. A major concern with performance assessment is that score interpretation not be limited to the sample of assessed tasks, but be broadly generalizable to the construct domain. This issue of generalizability of score inferences across tasks and contexts goes to the very heart of score meaning. Generalizability theory is often used to examine the degree of the generalizability of test scores. In addition, the limits of score meaning are also affected by the degree of generalizability across occasions and raters of task performance. Such sources of measurement error associated with the sampling of tasks, occasions, and raters underlie traditional reliability concerns.

2.5. External

The external aspect refers to the extent to which relationships between test scores and external assessments reflect those as constructs specify. Similar to what is known as criterion-related validity, this aspect usually examines correlations of test scores with relevant criteria, such as other measures of the targeted construct (convergent validity) and variables hypothesized to be unrelated to the construct (discriminant validity). Convergent evidence is based on high correlations due to shared constructs; discriminant evidence is based on low correlations due to different constructs.
2.6. Consequential

The consequential aspect examines the degree to which assessments have both intended positive effects and plausible unintended negative effects. Intended consequences can include changes in the instructional and curricular practices of teachers that lead to better learning environments for students. Unintended consequences can include bias in assessment leading to misinterpretation. The emphasis of the evaluation should be on the degree to which the positive outcomes of the test outweigh any negative consequences.

The foregoing validity framework provides a useful framework for organizing validity evidence for various types of speaking assessments, and it is desirable to have as many sources of evidence as possible that can significantly contribute to certain aspects of validity. Yet, depending on the purpose of a test and the types of interpretations of test scores that are made, some sources of validity evidence may be more credible than others. Since the three speaking tests to be reviewed could be used in different contexts for different purposes, it is not the intention of this article to argue for each test. Rather, readers must evaluate each in view of their own experience.

3. Test description and evaluation

This section reviews the speaking section of the TOEFL iBT, the VET, and the BEST Plus, for each of which of key characteristics of the tests are reported from three aspects: general description, test task characteristics, and scoring method. Table 1 and Table 2 present summaries of the characteristics for the three tests, respectively. Evidence for the evaluation of the tests is then presented in the previously introduced framework of validity, and summarized in Table 3 at the end of this section.

3.1. TOEFL iBT speaking section

3.1.1. General description

The TOEFL iBT was developed by Educational Testing Service (ETS). It debuted in September 2005 in North America and was later launched in other countries. It is a four-skill test in which the speaking section is mandatory. The speaking section of the test is used to evaluate examinees' oral communication skills in the context of their readiness for study in English-speaking universities. Specifically, it measures the ability to speak about everyday familiar topics, and to summarize, evaluate, compare, and synthesize information from multiple sources in a spontaneous manner.

The TOEFL iBT is delivered via an Internet-based delivery system but not completely Internet-based in that examinees need to take the test on particular computers at ETS-certified test centers around the world. During the test, examinees wear noise-canceling headphones and speak into a microphone. Responses are digitally recorded and sent to ETS's Online Scoring Network where they are scored by raters.

3.1.2. Test task characteristics

The speaking section of the TOEFL iBT consists of six tasks. Two are independent
tasks, which require examinees to express opinions on familiar topics related to campus life or lecture situations. The other four are integrated tasks. Two of the four are listening/speaking tasks, which ask examinees to listen to a short spoken recording and then respond to it. The final two are reading/listening/speaking tasks, which require examinees to read a short text, listen to a spoken recording that relates to what they read, and then respond about what they read and heard. Reading texts are 75–100 words long; recording lasts between 60 and 120 seconds. Examinees can take notes and use them when responding to the speaking tasks. Their comprehension is not separately assessed, as these texts are considered short and memorable.

The test is not computer-adaptive; each examinee must give response to all six tasks, and the tasks are not adapted to the examinees’ speaking level. For each task, examinees are given 15–30 seconds to prepare their response, and 45–60 seconds to respond. In total, the speaking section takes approximately 20 minutes.

3.1.3. Scoring

The TOEFL iBT speaking section uses holistic scoring, and three to six different certified raters issue a holistic score for each response on a scale of zero to four based on delivery, language use, and topic development. Delivery refers to the pace and clarity of speech, and raters consider speakers’ pronunciation, intonation, rate of speech, and degree of hesitancy. Language use includes the range, complexity, precision, and automaticity of vocabulary and grammar use. Raters evaluate examinees’ ability to select words and phrases and to produce structures that appropriately and effectively communicate their ideas. When assessing topic development, raters take into account the progression of ideas, the degree of elaboration, and completeness, and, in integrated tasks, the relevance of the content.

The average rating of all the raters is then converted to a scaled score from zero to 30. Examinees can view their scores online 15 business days after the test and choose to receive a copy of their score report by mail. Universities and agencies can also receive scores in paper or electronic formats on request. The ETS uses the Online Scoring Network, a secure Internet-supported system, to share data, train raters, and monitor scoring of examinees’ speaking samples. Raters are certified before they can begin scoring work and prior to each scoring session on a daily basis.

3.1.4. Evaluation

Empirical evidence for the validity of the TOEFL iBT speaking section is available for each of the six aspects as specified by the validity framework.

Content

In developing speaking tasks for the TOEFL iBT, a theoretical framework was first developed in which speaking construct measured by various task types was specified (Butler, Eignor, Jones, McNamara, & Suomi, 2000). Following this framework, task statements were drafted and then prototype tasks were developed.
<table>
<thead>
<tr>
<th>Test</th>
<th>Test delivery method</th>
<th>Task delivery</th>
<th>Task type (No. of items)</th>
<th>Topic</th>
<th>Planning time</th>
<th>Response time</th>
<th>Rater</th>
<th>Rating scale</th>
<th>Rating category</th>
<th>Score report</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL iBT</td>
<td>Internet-based through designated Computers</td>
<td>Non-adaptive</td>
<td>-Independent (1)</td>
<td>Familiar topic</td>
<td>15s</td>
<td>45s</td>
<td>Human</td>
<td>Holistic</td>
<td>-Delivery</td>
<td>Scale of 0–30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Independent (1)</td>
<td>Familiar topic</td>
<td>15s</td>
<td>45s</td>
<td></td>
<td></td>
<td>-Language use</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Integrated Reading/listening/speaking (1)</td>
<td>Campus life</td>
<td>30s</td>
<td>60s</td>
<td></td>
<td></td>
<td>-Topic development</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Integrated Reading/listening/speaking (1)</td>
<td>Academic course content</td>
<td>30s</td>
<td>60s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Integrated Listening/speaking (1)</td>
<td>Campus life</td>
<td>20s</td>
<td>60s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Integrated Listening/speaking (1)</td>
<td>Academic course content</td>
<td>20s</td>
<td>60s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note**: In the columns of planning time and response time, “s” refers to seconds.
Two studies provided positive empirical evidence for the content validity of the test. In Rosenfeld, Leung, and Oltman (2001), a large number of university faculty members and students in North America rated a series of task statements in relation to the importance of each task statement to the successful completion of coursework. Results support the claim that test content as indicated in task statements was relevant and representative of what examinees encounter in a university context. Cumming, Grant, Mulcahy-Ernt, and Powers (2005) evaluated seven speaking prototype tasks by asking seven experienced ESL instructors whether they thought the prototype tasks represented the domain of academic English required for studies at English-medium universities and fulfilled the purposes for which they had been designed. The instructors viewed the prototype tasks positively, particularly integrated tasks that required students to speak in reference to reading or listening to source texts.

**Substantive**

The substantive aspect of validity on the TOEFL iBT speaking section has been examined by analyzing speech samples, examinees' strategic behaviors, and their attitudes toward the test. While the former two types of analyses provided positive evidence, the latter yielded only negative evidence.

Brown, Iwashita, and McNamara (2005) analyzed speech samples of examinees on six prototype tasks to investigate whether the content of the rating scales was relevant to performance on particular tasks. They found that the greater fluency, more sophisticated vocabulary, better pronunciation, greater grammatical accuracy, and more relevant content were characteristic of speech samples receiving higher holistic scores from raters, which indicated that the construct that the tasks were designed to measure was indeed represented in the discourse.

Although use of strategy is considered integral to performing tasks, it is not included as a part of the speaking construct in the speaking section of the TOEFL iBT. To address the concern of construct-underepresentation, Swain, Huang, Barkaoui, Brooks, and Lapkin (2009) asked 30 Chinese ESL learners to report strategic behaviors when taking the test and examined how the reported strategies related to their test scores. No significant differences were found in reported strategy use across proficiency levels or between the total number of reported strategic behaviors and test scores. The results of the study did not change the construct of the TOEFL iBT speaking section.

Stricker and Attali (2010) administered a questionnaire concerning attitudes toward the test to examinees in four countries. Overall attitudes about the TOEFL iBT were moderately positive in most countries. Yet, compared with other skills, attitudes toward the speaking section were consistently less favorable in all countries, and were unfavorable in two countries. Although this study did not relate speaking test scores to examinees’ attitudes toward the speaking section, the results showed that attitude toward the test may be one potential source of construct-irrelevant variances.
Structural

Two investigations into the appropriateness of using a holistic rating scale have justified the validity of the structural aspects of the test. First, to ensure the relevance of the content of the holistic rating scales with features of speaking proficiency deemed by expert judges, Brown, Iwashita, and McNamara (2005) used verbal-report methodology to examine what aspects of examinees’ speech the raters paid most attention to when scoring tasks. Results showed that with no guidance, domain experts distinguished between performances using a common set of criteria similar to those included in rating scales developed for the tasks. Second, Xi and Mollaun (2006) justified using a holistic rating scale rather than analytic rating scales. The study found that the reliability of scores obtained from the two types of rating scales did not differ substantially. But since the analytic scores were highly correlated, they would not provide additional information beyond what the holistic scores could offer for most examinees.

Generalizability

Generalizability of the TOEFL iBT speaking test was verified across occasions, raters, and tasks. Zhang (2008) analyzed the scores of examinees who repeated the test once within 30 days. Result showed a moderate correlation of 0.84 between the scores of the two tests, suggesting that the difficulty of tasks on different occasions could be considered equivalent. With regard to rater generalizability, inter-rater reliability was not published by the ETS for operational settings, and instead, the ETS claimed that raters are constantly monitored each time they score a test. In addition, Xi and Mollaun (2009) found that Indian raters could rate the speech of examinees of various national backgrounds reliably, which provides positive evidence for recruiting nonnative speakers to score the test. Finally, as to the generalizability of the scores across tasks, the generalizability index published by the ETS is 0.88, higher than other sections. This may be attributable to the Lee’ study (2005), which justified the test design concerning the present number of tasks for each task type in order to achieve acceptable reliability.

External

Convergent evidence for external validity has been collected through comparison with the self-assessment of examinees and with local test. Two studies provided evidence for the distinctiveness of what is measured in the speaking section from other parts of the test. With a sample of over 2000 examinees, Wang, Eignor, and Enright (2008) found that the correlation between the responses to the statements of self-assessment on speaking skills and the summative scores for the test was a moderate 0.55. In addition, the speaking section is also used also for selection of ITA (International Teaching Assistant). Therefore, Xi (2008) examined the relationships between the speaking scores and those on several local ITA tests, and the regression results suggest that the speaking test scores could significantly predict scores on students’ TA assignments. In Sawaki, Stricker, and Oranje (2008), the factor loading patterns of the test indicate that both independent and integrated tasks define the
speaking constructs accurately, and are minimally involved in the reading and listening constructs. Stricker and Rock (2008) verified the result of Sawaki et al. for different groups in native language and period of exposure to the English language.

**Consequential**

As to washback, the addition of the speaking section into the TOEFL iBT has improved the impact of the test on teaching and learning. Several studies by Wall and Horak (e.g., 2008) reported on the practices of teachers a year after the launch of the TOEFL iBT in Europe. They concluded that the TOEFL iBT has indeed had the desired effect on the content of TOEFL preparation classes, in that much more emphasis has been placed on the teaching of speaking abilities and integrated tasks.

### 3.2. Versant English Test

#### 3.2.1. General description

Versant English Test (VET), developed by Pearson, intends to measure what the developers called "facility in a spoken language." It is defined as the ability to understand spoken English on everyday topics and to respond appropriately at a native-like conversational pace in intelligible English. This ability is assumed to underlie high performance in communicative settings, since learners must understand interlocutors correctly and efficiently in real time to be able to respond. Learners must also be able to formulate and articulate a comprehensible answer without undue delay.

Without specifying the uses of the test, developers claimed the test could be used for various purposes, including measuring proficiencies, placement, and achievement. The VET is Internet-based and can be taken wherever there are telephones or computers with access to the Internet. However, the end users of the test are responsible for administration: verifying the identity of the examinee, giving the examinee the required material, and monitoring the administration of the test. Examinees can retrieve their results from a secure website within minutes of test completion.

#### 3.2.2. Test task characteristics

The VET consists of 63 items grouped into six sections: reading aloud, repeating, short answer questions, sentence builds, story retelling, and open questions. This variety provides multiple fully independent measures that underlie facility with spoken English, including phonological fluency, sentence construction and comprehension, passive and active vocabulary use, and pronunciation of rhythmic and segmental units.

Developers designed the test so that each section consists of increasingly challenging items delivered by native speakers from various English-speaking regions. In addition, through the use of algorithms, the Versant Testing System selects items for each examinee from a pool of items graded for difficulty, thus ensuring that each administration event is unique. The printed text is available on a test paper given to the examinee a few minutes before the beginning of the test, and is also available for those using a computer. The entire test lasts approximately 15 minutes.
Table 2: Summary of major features of the VET and the Best Plus

<table>
<thead>
<tr>
<th>Test</th>
<th>Test delivery method</th>
<th>Task delivery</th>
<th>Task characteristics</th>
<th>Scoring method</th>
<th>Score report</th>
</tr>
</thead>
<tbody>
<tr>
<td>VET</td>
<td>Internet-based through designated computers</td>
<td>Non-adaptive</td>
<td>-Read aloud (8)</td>
<td>Pronunciation Fluency</td>
<td>Score of 20-80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Repeat sentence (16)</td>
<td>- Sentence mastery Fluency</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Short answer (24)</td>
<td>- Vocabulary</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Sentence builds (10)</td>
<td>- Pronunciation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Story telling (3)</td>
<td>Sentence mastery</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Open questions (2)</td>
<td>Fluency</td>
<td></td>
</tr>
<tr>
<td>BEST  Plus</td>
<td>Computer-based</td>
<td>Adaptive</td>
<td>Photo description</td>
<td>Listening comprehension</td>
<td>Score of 88-999</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Not applicable)</td>
<td>- Language</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Entry item (Not applicable)</td>
<td>complexity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Yes/no question (Not applicable)</td>
<td>- Communication</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Choice question (Not applicable)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Personal expansion (Not applicable)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>General expansion (Not applicable)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Elaboration (Not applicable)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: In the column of response time, “m” refers to minutes.
3.2.3. Scoring

The VET is thus far the only test that has applied automatic speech recognition technology to the rating of the speech sample. It provides information on four diagnostic subscores (sentence mastery, vocabulary, fluency and pronunciation) as well as an overall score. In assessing fluency and pronunciation, such elements as rate of speech, length and position of pauses, stress, and segmentation are measured. Scores are reported both numerically (in a range of 20 to 80) and in criterion-referenced descriptors. The first item in each section is considered a practice item and is not scored. The responses to the story-retelling section and the open-ended questions are not automatically scored, and in total, 64 responses are automatically scored.

3.2.4. Evaluation

In the 2008 technical report of the VET, published online by Pearson, various types of validity evidence were introduced in detail, which can be summarized from the perspectives of content, structural, generalizability, and external validity.

Content

Content validity of the VET was based on expert judgment on two aspects. First, the audio prompt in the test is not biased toward certain dialects. Second, the content of the questions is context-free, and examinees only need general rather than specific knowledge to answer the questions.

Structural

Correlations among test subscores on each scoring category—sentence mastery, vocabulary, fluency and pronunciation—range from 0.61 to 0.90. This indicates that the different scores are measuring distinct aspects of the test construct, and therefore offer useful diagnostics.

Generalizability

Since VET scores do not involve human raters, besides test-retest reliability, the generalizability aspect of the VET was supported by agreements of machine-produced scores, correlations with human scores, and a comparison of machine-produced scores with human scores. First, the correlation between the scores from the first and second administrations was found to be $r = 0.97$, indicating high test-retest reliability for the VET. Second, the agreements between ratings of the VET machine-generated scores, ranging from 0.88 to 0.97, were similar to those of human raters, suggesting that they are as reliable as human raters. Third, correlations between machine-generated scores and human scores, based on data from 50 examinees were high, ranging from 0.89 to 0.97. In particular, at the overall score level, the VET machine-generated scores are virtually indistinguishable from scoring based on careful human transcriptions and repeated independent human judgments. Finally, overall scores show an effective separation between native and non-native examinees. Test scores of a norming group of
775 native speakers of English were compared with those of a non-native norming group of 603 speakers. The score distributions for the different groups of examinees show that native speakers all perform well, while non-natives show a range of ability levels.

**External**

To establish the external validity of the VET, its overall scores were compared with other well-established large-scale tests. It was found that the correlations between the VET overall scores and those on other speaking tests, including the Test of Spoken English (TSE)\(^1\), the speaking section of the TOEFL iBT, and the International English Language Testing System (IELTS) Speaking Test\(^2\) were quite high, ranging from 0.75 to 0.94. On the other hand, the correlations were moderate between the VET scores and the TOEFL iBT overall scores. Based on these findings, the developers claimed that the VET scores are more closely related to those tests measuring speaking ability rather than general language ability; this seems to provide convergent and discriminant validity for the test.

**3.3. BEST Plus**

**3.3.1. General description**

The original Basic English Skills Test (BEST) was developed in the early 1980s as an easily administered assessment of the speaking abilities of non-English-speaking adult refugees and immigrants to the United States. In order to shorten administration time but maintain greater accuracy of measurement for accountability assessment, researchers at the Center for Applied Linguistics (CAL) developed BEST Plus, a computer-adaptive version of the BEST oral interview. The purpose of the test is to assess the ability to understand and use unprepared, conversational, everyday language within topic areas generally covered in adult education courses. The test is used to measure language gains in individuals as well as the overall effectiveness of a language program through a pretest-posttest process.

The BEST Plus is available in two face-to-face formats: computer-adaptive and print-based. Here, only the computer-adaptive format is introduced. In the computer-adaptive format, test items are delivered through a computer with a Best Plus CD, USB, or Network. The test is conducted before a test administrator who reads the question from the computer screen, listens to the candidate’s response, and enters the scores into the computer. The computer-adaptive system determines the difficulty level of the next question based on current estimates of the examinees’ ability. At times, the

---

\(^1\) The TSE is a tape-based test that measures the ability of nonnative speakers of English to communicate effectively. The test is used for employment, graduate assistantships, licensure, and certification purposes and has been discontinued as of March 31, 2010.

\(^2\) The IELTS speaking test, delivered by a certified examiner, consists of three parts: (1) answering everyday questions, (2) talking about a topic, and (3) having a discussion on the topic with the examiner.
examinee is invited to look at the computer screen (for example, when shown a picture to describe), but the examinee does not operate the computer.

### 3.3.2. Test task characteristics

During the test, the examinee is presented with a relatively easy item on a selected topic, followed by increasingly challenging questions if the examinee is able. There are 13 general topic domains, such as personal identification, health, housing, and transportation. The test comprises seven item types that vary in their cognitive and linguistic demands: photo descriptions, entry items, yes/no questions, choice questions, personal expansion, general expansion, and elaboration. Each time a new topic is begun, the first item presented is an entry question or photo description. Higher-ability examinees quickly move to open-ended expansion questions while lower-ability examinees move to questions that provide more support, such as choice or yes/no questions. An examinee may encounter up to seven different topically organized groups of questions in a test administration. The difficulty of test tasks was determined by analyzing the data collected from a field test that involved a large number of examinees. The test lasts between 5 and 20 minutes, depending on the proficiency level of the examinee.

### 3.3.3. Scoring

The BEST Plus evaluates examinees’ performance on a rating scale consisting of three criteria: listening comprehension, language complexity, and communication on a zero–ten scale. Scores are reported on a scale of 88–999.

### 3.3.4. Evaluation

Validity evidence for the BEST Plus is available on two aspects, generalizability and external validity, as reported in the BEST Plus technical report (2005).

**Generalizability**

Evidence for the generalizability aspect of test validity was provided with regard to occasions, ratings, and test forms. First, to establish test-retest reliability, 32 examinees took the test on two successive occasions; the correlation between the two sets of scores was 0.89. Second, during the test, both an expert scorer and a novice scorer rated all the responses, and the correlations for raw test scores between the rater pairings were consistently above 0.90. Third, parallel-form reliability was established through investigating the equivalence of the three print-based test forms in a study of 48 adult ESL students who each took two test forms. The correlation of scores between test forms ranged from 0.85 (Forms B and C) to 0.96 (Forms A and C). The rationale for using print-based test forms was that it would be difficult to predict the test items that are used for each form of the BEST Plus since it is a computer-adaptive test. Instead, test items on the print-based tests were drawn from the computer-adaptive item pool.
External

Evidence for this aspect was gathered from the relationship between test scores and Program Placement Levels and the relationship between test scores and scores from other measures of English proficiency. First, the field test was conducted with 24 programs that provided information about which classes 1866 of their students were placed into on the basis of existing measures and sorting procedures. The average correlation coefficient showing the relationship between placement levels and scores on the BEST Plus across all programs was 0.72. Developers claimed that this demonstrated the usefulness of the test as a placement instrument in ESL programs. Second, another study that established the relationship between educational level gain on the BEST Plus and the number of instructional hours received. It was found that 53% of examinees made a level gain if they received less than 60 instructional hours, as compared to 70% of examinees who made a level gain if they received more than 140 instructional hours.

3.4. Summary

In this section, we introduced the three computer-delivered speaking tests, each of which features a distinct application of computer technology: the TOEFL iBT uses both independent and integrated tasks and makes full use of multimedia for task prompts; the VET is the only one scored by machine; and the BEST Plus delivers test tasks in response to the examinee’s ability.

Table 3 summarizes validity evidence for the three tests and the followings are clear from the table. First, validity evidence is more abundant for the TOEFL iBT than for other two tests. Second, given that the tests have distinct features and accordingly different concerns about each test’s validity, test developers have conducted different types of studies that fall into the same categories. Finally, the studies reviewed have been mainly concerned with aspects of content, generalizability across occasions and across raters, and relationships with external scores. The other aspects of substantive, structural, and consequential have seldom been dealt with.

4. Challenges and future directions

4.1. Issues to consider for test development

As reflected in the three tests reviewed, new developments in the application of computer technology to testing speaking have made the delivery of speaking tests more efficient in many ways. However, when using computers to develop tasks, it is essential to consider the balance between efficiency and practicality, particularly for high-stake or small-scale tests. In the following, we consider the issues in test development on methods of test delivery, task delivery, rating, and, most importantly, test construct.
Table 3: Summary of validity evidence on the selected computer-delivered speaking tests

<table>
<thead>
<tr>
<th>Aspects of validity</th>
<th>Definition</th>
<th>Computer-delivered speaking test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>• Consistency and representativeness of test content with construct</td>
<td>- Development stage:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Expert judgment:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rosenfeld et al (2001)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Expert judgment:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>context-free;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>not biased on certain dialect</td>
</tr>
<tr>
<td>Substantive</td>
<td>• Consistency of response process with construct</td>
<td>- Speech sample analysis:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Think aloud:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Swain et al (2009)</td>
</tr>
<tr>
<td></td>
<td>• Existence of construct-irrelevant variance</td>
<td>- Questionnaire:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stricker &amp; Attali (2010)</td>
</tr>
<tr>
<td>Structural</td>
<td>• Consistency of scoring method with construct domain</td>
<td>- Think-aloud:</td>
</tr>
<tr>
<td></td>
<td>• Relationships among rating categories</td>
<td>- ( r = 0.61 - 0.90 )</td>
</tr>
<tr>
<td>Aspects of validity</td>
<td>Definition</td>
<td>Computer-delivered speaking test</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TOEFL iBT</td>
</tr>
<tr>
<td>Generalizability</td>
<td>- Replicability of results across tasks</td>
<td>- Generalizability theory analysis: Lee (2005), 0.88</td>
</tr>
<tr>
<td></td>
<td>- Replicability of results across raters</td>
<td>- Rater agreement monitored but not published</td>
</tr>
<tr>
<td></td>
<td>- Replicability of results across occasions</td>
<td>- Test-retest reliability: ( r = 0.88 )</td>
</tr>
<tr>
<td>External</td>
<td>- Convergent</td>
<td>- With self-assessment: Wang et al (2008) ( r = 0.55 )</td>
</tr>
<tr>
<td></td>
<td>- With other local ITA tests: Xi (2008)</td>
<td>- With other local ITA tests: Xi (2008)</td>
</tr>
<tr>
<td>Consequential</td>
<td>- Existence of intended positive or negative effects</td>
<td>- Wall &amp; Horak (2008)</td>
</tr>
</tbody>
</table>
4.1.1. Computer-based and Internet-based testing

Internet-based tests such as the VET are highly efficient as they can be taken anywhere and anytime as long as examinees have access to the Internet or a telephone. However, they have at least two potential problems: the risk of leaking of test items and difficulty of identifying the examinee. These issues have made it difficult for high-stake tests to be administered via the Internet, and tests such as the TOEFL iBT compromise by delivering the test through the Internet to designated sites.

4.1.2. Computer-adaptive, semi-adaptive, and non-adaptive testing

In computer-adaptive tests, the computer gives test tasks that are appropriate to the ability of the examinee. This can save administration time compared with non-adaptive tests that give the same tasks to all the examinees. Although widely used for testing receptive skills, computer-adaptive tests are not a very popular choice for testing speaking skills because of the high cost of developing speaking test tasks. The ideal item pool for a computer adaptive test would have a large number of highly discriminating items at each ability level. A prerequisite of such an item pool is field testing, which is labor-intensive and at a high risk for leakage of test items.

In addition, to make a speaking test completely computer-adaptive, the computer must be able to rate an examinee's speech accurately. So far, it is still difficult for computers to achieve high precision in recognizing the examinee's speech on open-ended tasks. Tests such as the BEST Plus, which require test administrators to input a score for a task in order for the computer to determine the next task, are a way around this problem. In this case, test developers must still weigh the costs of developing an item pool against the merits of saving administration time.

4.1.3. Automatic rating and human rating

Besides delivering test tasks, a computer can increase the efficiency of testing speaking by replacing human raters. However, test developers can face two difficulties in applying automatic rating technology to testing speaking. First, automatic rating, as in the VET, can currently only be used for rating tasks that have highly predictable responses, such as read-aloud or repetition tasks. Tasks that require extended responses from examinees are still difficult to rate satisfactorily by machine. A second impediment to the widespread use of automatic rating is the high cost of the development and application of speech recognition technology to testing.

4.1.4. The test construct of speaking: to include listening ability or not

In all three tests reviewed, listening comprehension ability plays an important role in the completion of test tasks. It is true that in order to interact successfully with others in English, we must first understand what others are saying. This trend may also be attributable to the increasing use of multimedia in task prompts. However, we must be careful with score interpretation. Also, test developers need to decide on the difficulty of the listening prompt, which may depend on to what extent a successful understanding
of listening prompts is intended to affect the completion of tasks.

4.2. Areas for future research

As reviewed in the previous section, a considerable amount of care and study has been devoted to specifying the content and scoring procedures of the tests and determining their generalizability across raters and their relationships with external scores. However, more research is needed on the substantive, generalizability, and consequential aspects of the validity of computer-delivered speaking tests in order to improve the validation process and provide more useful information to stakeholders.

4.2.1. Substantive aspect

Using computers in testing speaking is likely to cause serious construct-underrepresentativeness and construct-irrelevant variance from the perspective of examinees’ use of strategy as well as potential effects of computer-delivery method and examinee characteristics on test performance.

Examinees’ use of test-taking strategy

Although theoretically, the use of test-taking strategy is part of language ability, we lack information regarding the context of computer-delivered speaking tests. Studies on this topic are likely to be illuminating with respect to the construct validity of computer-delivered speaking test scores. Empirical evidence can help test developers decide whether to include the use of test-taking strategy as part of the test construct, and thus avoid the problem of construct-underrepresentativeness. Analysis of self-reporting by think-aloud protocol or questionnaire can reveal the cognitive and metacognitive processes involved when examinees respond to various speaking tasks.

Effect of computer-delivery method

Different from a traditional speaking test where an interlocutor is present to administer the test or act as the conversation partner, the computer-delivered one asks the examinee to speak to the computer. Exploration of the effects of delivery method on examinees’ test performance is thus important in order to eliminate construct-irrelevant variance. Also, how examinees actually use the planning time in a computer-delivered speaking test compared with a face-to-face test is another issue for further investigation.

Effect of examinee characteristics on test performance

Research on the effect of examinee characteristics on test performance could further our understanding of the constructs that tests assess. Future studies should explore the possible interaction of affective factors, such as attitudes toward the test, computer anxiety, and communication anxiety, with performance on a computer-delivered speaking test. Also, using computers for assessment may cause construct-irrelevant variance due to a lack of computer familiarity on the part of examinees. Considering the low exposure to computers for high school and university
students, computer familiarity is a factor that needs investigation.

4.2.2. Generalizability across tasks

So far, research conducted on performance assessment has shown that a large examinee-by-task variance exists that would limit the generalizability of scores from these tests (e.g., Brennan, Gao, & Colton, 1995). Since administration time for computer-delivered speaking tests usually lasts 15 to 20 minutes, increasing the number of tasks is often impractical. More studies are needed to determine the boundaries of interpretations and ways of reducing the interaction of examinees with tasks.

4.2.3. The consequential aspect of validity

Studies on the impact of administering computer-delivered speaking tests on teaching and learning English are still few in number. Empirical evidence about various learning contexts is needed. Further research should explore how learners prepare for speaking tests delivered by computer and how this differs from preparation for face-to-face tests in case there are any negative unintended consequences.

5. Conclusion

New developments in the application of computer technology have made the delivery of speaking tests more efficient in many ways. From summarizing the validity evidence available for the three tests, we hope that test developers and test users, who are ultimately responsible for fair and valid test use, will heed our calls for further validity work. Continuing efforts in this area are essential given the expected increasing use of computers in testing speaking.

Acknowledgement

I am deeply grateful to Prof. Negishi Masashi for encouraging me to write this article. I would also like to thank two anonymous reviewers for their helpful comments, and Dr. Koizumi Rie for her constructive advice on choosing the validity framework.

Reference


