JRC Scientific and Technical Reports

Computer-based Assessment (CBA) of Foreign Language Speaking Skills

Luísa Araújo (Ed.)

EUR 24558 EN - 2010

JRC EUROPEAN COMMISSION

ipsc Institute for the Protection and Security of the Citizen

Computer-based Assessment of Foreign Language Speaking Skills
CBA 2010

Luísa Araújo

The mission of the JRC-IPSC is to provide research results and to support EU policy-makers in their effort towards global security and towards protection of European citizens from accidents, deliberate attacks, fraud and illegal actions against EU policies.

European Commission
Joint Research Centre
Institute for the Protection and Security of the Citizen

Contact information
Address: Econometrics and Applied Statistics Unit, Via E. Fermi 2749, Ispra (VA), Italy
E-mail: luisa.borges@jrc.ec.europa.eu
Tel.: +39 0332 78 5268
Fax: +39 0332 78 5733

http://irmm.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

JRC 60919

EUR 24558 EN
ISBN 978-92-79-17173-4
ISSN 1018-5593
doi:10.2788/30519

Luxembourg: Publications Office of the European Union

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In the context of current efforts to establish an indicator of foreign language competence in Europe, based on the results of the first European Survey on Language Competences expected in 2012, a conference on computer-based assessment (CBA) of speaking skills was held in Brussels in June 2010. The sponsors of this event, the Centre for Research on Lifelong Learning of the European Commission in Ispra and the Directorate General of Education and Culture, are pleased to present this report with the Conference proceedings. It is our hope that it can serve not only to inform the implementation of the language survey but also to assist Member States and their representatives in the task of gathering current evidence on the benefits and challenges linked to the implementation of CBA.

During the meeting, various experts in the field of foreign language assessment presented state of the art research on the computer based assessment of speaking skills. The different scientific and technical issues discussed provide a good overview of current developments in the assessment of speaking proficiency in a foreign language in the context of large scale assessments. General practical benefits of CBA include the possibility of testing many examinees within a short time span and the assurance that standardized testing procedures are followed. Technical advances have made it possible to develop fully automated assessment instruments and we are now able to deliver computerized tests of oral proficiency that begin to emulate authentic interactions between human beings. The first oral language proficiency assessments relied solely on the interaction between humans established during the Oral Proficiency Interview (OPI). The SOPI, (Simulated Oral Proficiency Interview), introduced a simulated format whereby speech is elicited by a tape recorder, but rated by experts. Computer-assisted foreign language assessments, on the other hand, are of two sorts. One version requires no human intervention for test delivery, but the candidate's performance is assessed by expert raters. On the other hand, fully automated versions are both delivered and scored by machine and require no human intervention. The Computerized Oral Proficiency Instrument (COPI), like the SOPI also produced by the Centre for Applied Linguistics (CAL), represents the former development by reliably capturing oral language

ability, to be assessed by humans, whereas the latter, fully automated version, is represented by Versant (from Pearson Knowledge Technologies), which is scored entirely by computer.

The testing of foreign language speaking ability has evolved over time according to specific guidelines that define levels of language proficiency. In Europe, the *Common European Framework* (CEFR) sets the guidelines against which proficiency levels are rated and reported, while in the US the American Council on the Teaching of Foreign Languages (ACTFL) provides proficiency level definitions. Importantly, despite differences in the two speaking proficiency frameworks, the work developed by CAL according to the ACTFL guidelines has been influential in the development of current assessment instruments in Europe. Thus, the work carried out at CAL on the development of OPIs and SOPIs is frequently referred to by other authors of this report. From the paper by Dorry Kenyon and Meg Malone the reader can follow the developments in test design from the OPI to the COPI and get a sense of the challenges involved in capturing oral language ability using computer-based approaches.

Next, the paper by Evelina Galaczi argues that a "fitness for purpose" criterion should guide the adoption of test formats. Whereas computerized formats may not render the full range of speaking abilities in interactional contexts, they may offer a valid snapshot of linguistic ability. The paper by Ildikó Csépes describes how a paired exam mode following a SOPI format can be implemented without compromising test validity. Similarly, Thomas Eckes describes the design of a test of German as a Foreign Language as a particular implementation of the SOPI and discusses how scoring validity can be accomplished. Alistair Van Moere gives an account of how the fully automated Versant tests provide a set of performance-based measures linked to the CEFR. Finally, Carsten Roever challenges our thinking about how to capture pragmatic language ability using automated computer-based assessment and offers some examples for the contextualization of test items that elicit pragmatic responses.

As the papers included in this report clearly show, the field of foreign language assessment, and in particular the assessment of speaking ability, has come a long way and it is now possible to implement different assessment modes using computer-based technologies. Furthermore, regardless of whether a semi-automated or a fully automated system is implemented, we can be

confident that reliable tests are now available that can offer a valid measure of a learner's oral proficiency level. However, it is also clear that capturing the full range of the features of human spoken interaction is no easy task and future research efforts will likely focus on capturing complex linguistic interactions played out in a virtual world. Fortunately, it is already possible to create assessments that are fully delivered and scored by computers. In the context of large scale surveys, CBA clearly has the potential to facilitate implementation by reducing testing time and costs without compromising validity.

Luísa Araújo

Ispra, October 2010

INVESTIGATING EXAMINEE AUTONOMY IN A COMPUTERIZED TEST OF ORAL PROFICIENCY

Dorry Kenyon and Meg Malone

Center for Applied Linguistics

Abstract

This paper presents results from investigations into the effect of allowing examinee autonomy in a computerized test of oral proficiency. It begins with a description of the computerized assessment, the Computerized Oral Proficiency Instrument (COPI), as operationalized in Spanish and Arabic. It then presents results from research on allowing examinee autonomy in the assessment. The talk will present the results to these investigations. Overall, examinee attitudes were very positive towards the choices that the COPI provided them and that, though preparation time and response time varied by proficiency level and complexity of the task on the COPI, students felt that the assessment format allowed them the opportunity to provide a demonstration of their current proficiency levels. Overall proficiency ratings on both test formats (COPI and SOPI) were comparable across languages.

1. Introduction

When assessing students' attainment of foreign language skills, the testing of speaking poses formidable challenges. Feasible approaches to large-scale testing of second language skills in reading, writing, and listening have been available for decades. However, large-scale testing of the multi-faceted domain of speaking can be prohibitively resource-intensive in terms of the time and personnel required both to administer speaking tests and to score student responses.

Direct approaches to testing speaking have typically involved a test administrator/rater in a one-on-one interview testing format, or one or more administrators/raters facilitating paired or grouped testing. While these approaches appear to validly assess speaking, researchers have been clear to point out that most fail to capture many aspects of speaking that occur in real life outside of the testing situation. For example, the one-on-one oral proficiency interview approach has been criticized for not allowing the examinee to present strategies important in true conversational interactions (e.g., Johnson, 2000; Johnson, 2001). Group and paired oral interviews, intended to allow more interpersonal communication between examinees to take place, have also received their share of critical evaluation (e.g., Iwashita, 1999; O'Sullivan, 2002; Taylor, 2000; Taylor, 2001). Research into these speaking tests help reveal what type of oral language can be elicited by them—and what not.

Experience with testing speaking has repeatedly shown that the construct of speaking as operationalized by the assessment must be clearly articulated. Either the approach to testing speaking must be chosen to validly operationalize that understanding of what speaking is, or users of any approach must be satisfied with the *de facto* operationalization of what speaking is that is intrinsic to the testing format chosen. The fundamental validity question, as Messick (1989) so clearly stated, will be in relation to the inferences test score users make about students' speaking ability, or actions they may take in regards to students, teachers, and educational programs, on the basis of students' performances on the speaking assessment. Any assessment is highly contextualized and needs to be understood within its own situation.

The purpose of this paper is to describe how, within a specific context of testing oral proficiency in the United States using the *Speaking Proficiency Guidelines* of the American Council on the Teaching of Foreign Languages (American Council on the Teaching of Foreign Languages, 1999), technology has been successfully used to provide an alternative to a direct, face-to-face oral proficiency interview assessment. Earlier technology (tape-recordings) led to the first generation of this assessment, known as the Simulated Oral Proficiency Interview (SOPI). We will focus, however, on the more recent use

of computer technology in the second generation of this speaking test format, known as the Computerized Oral Proficiency Instrument (COPI). We begin by explaining some background to and context to the COPI. We then describe the COPI and summarize the results of published studies on it. We highlight some of the benefits to examinees, raters, test developers and researchers to the computer –based approach in this specific context. It is our hope that understanding this long-term research and development initiative at the Center for Applied Linguistics may provide a model for using computers in the testing of speaking skills in other contexts.

2. Background

In the United States, the testing of speaking skills in government and education has been dominated by the Oral Proficiency Interview (OPI). First developed by the U.S. Foreign Service Institute in the 1950s (Liskin-Gasparro, 1987), the OPI developed together with the guidelines that described speaking proficiency, as understood by the government at that time. In this sense, both the definition of the construct and the assessment procedure developed symbiotically, and to this day some would claim that the two are inseparable (Bachman, 1988).

The Foreign Service Institute's OPI and the Skill Level Descriptors for Speaking that defined the construct for the test were further developed and adopted by an intergovernmental agency, the Interagency Language Roundtable (ILR) and continue to be used extensively by U.S. government agencies as a common assessment and common description of levels of speaking ability across languages. In the 1980s, the ILR descriptors were revised for use in the U.S. educational context by the Educational Testing Service (ETS) and the American Council on the Teaching of Foreign Languages (ACTFL) (Liskin-Gasparro, 1987). ACTFL developed the ACTFL *Speaking Proficiency Guidelines* and its accompanying ACTFL OPI, and promulgated both the proficiency guidelines and the use of the OPI throughout the country. The so-called proficiency movement, built on the proficiency level definitions provided by the guidelines and on the OPI, took hold in American foreign language education where proficiency came to be seen as a unifying principle (see for example Hadley, 1990; Uber-Grosse & Feyton, 1991).

Criticism of the guidelines and the OPI appeared from the first (e.g., Bachman & Savignon, 1986; Bachman, 1988; Lantolf and Frawley, 1985, 1988), which some researchers claim have yet to be answered (e.g., Norris, 2001). Nevertheless, because of its widespread currency and link to the government's procedures and scales, the ACTFL OPI is considered by many in education, particularly in the United States, to be the "gold-standard" against which other speaking assessments are evaluated.

The ACTFL OPI relies upon a highly trained interviewer who administers the assessment following a structured procedure designed to allow the interviewer to determine the interviewee's global speaking proficiency as defined by the ACTFL *Speaking Proficiency Guidelines.* Through an individually determined series of questions, the interview assesses the examinee's ability to perform speaking functions at the four major ACTFL levels (Novice, Intermediate, Advanced and Superior), eliciting evidence in a ratable speech sample that probes the examinee's highest level of functional speaking ability as defined by the guidelines.

However, as the ACTFL OPI and its accompanying guidelines began to be promulgated in the United States, it became clear to researchers at the Center for Applied Linguistics (CAL) that a dearth of interviewers, particularly in the less commonly taught languages, would limit the benefits that could be gained from having a common set of guidelines that could be used across languages. Thus, beginning in the mid 1980s, language testing specialists at CAL explored methods of assessing oral proficiency following the ACTFL guidelines that did not require a trained interviewer conducting the interview in live time. Because they did not involve an interviewer, these methods came to be known as semi-direct assessments of speaking skills. They could be administered simultaneously to a group of examinees and required no trained tester for administration (although training in the ACTFL guidelines would be required to rate them).

Early efforts toward semi-direct assessment relied upon a cassette tape recorder and test booklet to elicit speaking performances ratable using the ACTFL guidelines from examinees. The speaking performance was recorded on cassette tape and sent to trained raters to rate at a later time. The first such test was in Chinese (Clark, 1988), followed by Portuguese (Stansfield et al, 1990), Hausa (Stansfield & Kenyon, 1993), Hebrew (Shohamy et. al, 1989) and Indonesian (Stansfield & Kenyon, 1992a). During this time, the format was also used to develop an assessment in Spanish and French used to assess teachers' proficiency in those languages (Stansfield & Kenyon, 1991). In the early 1990s, the Chinese Speaking Test was updated and tests also appeared in Russian, Spanish, French, and German.

Because the major validity claim of the SOPI was that the semi-direct format elicited an examinee speech sample that could be rated according to the criteria of the ACTFL guidelines (which focus on function, discourse, and accuracy), the research on the test focused on studies involving the same set of students taking both the SOPI and OPI. Published research (cited above) showed high correlations between examinee performances as rated by the criteria of the ACTFL guidelines on the SOPI and the

OPI. Correlations between the OPI and SOPI ranged from .89 (Israeli Hebrew) to .93 (Mandarin Chinese and Portuguese) to .94 and .95 (in American Hebrew and Indonesian, respectively). Later research in German (Kenyon & Tschirner, 2000) focused on performances at the lowest proficiency levels and likewise found that examinees were rated similarly using the ACTFL criteria, whether the speech performance was elicited via the OPI or the SOPI. For only two subjects in the study did examinees have different outcomes, although both were at adjacent levels on the scale.

Outside of CAL, the SOPI format was adapted in different contexts. In the U.S., it informed the development of the speaking portion of the assessments of the Minnesota Articulation Project (Chalhoub-Deville, 1997), while in Germany it had influence on the TestDaF, used for admission of students for whom German is a foreign language to German universities (Kenyon, 2000b; TestDaF-Institut, 2002). It was also used at Sookmyung Women's University in Korea as a model for the Multimedia Assisted Test of English Speaking, designed to assess global speaking competence of Korean speakers of English (Lee, 2007).

As computer technology expanded in U.S. education towards the end of the 1990s, researchers at CAL explored the possibility of using a computer rather than cassette tape and test booklet to elicit speech samples ratable by the criteria of the ACTFL guidelines. Dorry Kenyon, Valerie Malabonga and others at CAL (Kenyon, 2000a; Malabonga and Kenyon, 1999; Malabonga and Kenyon, 2000) conducted a research program from 1997-2000 to determine the feasibility of eliciting and gathering language ratable according to the ACTFL guidelines via computer. In particular they investigated the extent to which the use of computer technology could overcome some of the limitations of the SOPI method. The assessment format was called the Computerized Oral Proficiency Instrument (COPI) and their studies, in Spanish, Arabic and Chinese, likewise indicated high correlations between the OPI and COPI and the SOPI and COPI. Following a description of the COPI, the results from other aspects of their research, focusing on examinee and rater behavior and attitudes, are summarized in this paper. It may be noted that the COPI has now become operational in Spanish and Arabic.

3. Description of the COPI

To understand the COPI, it is important to first have some background on CAL's SOPI. The SOPI is a performance-based, tape-mediated speaking test that relies on audiotaped instructions and a test booklet to elicit oral performances from the examinee. All tasks on CAL's SOPI are designed to maximize the likelihood that the examinees' speech performance may be rated using the criteria of the

ACTFL *Speaking Proficiency Guidelines.* They are also contextualized to ensure that they appear as authentic as possible within the constraints of the testing format.

The prototypical CAL SOPI follows the same four phases as the ACTFL OPI: warm-up, level checks, probes, and wind-down. The warmup phase is designed to ease examinees into the test format and contains simple personal background questions. The tasks in the level checks and probes allow examinees to demonstrate their ability to perform different speech functions defined in the ACTFL *Speaking Proficiency Guidelines* at the Intermediate, Advanced, and Superior levels. Such functions include asking questions, giving directions based on a simple map, describing a place, narrating a sequence of events based on the illustrations provided, speaking about selected topics or performing in simulated real-life situations to apologize, describe a process, support an opinion, or speak persuasively. Because these latter tasks may include functions too complex for lower-level examinees, there is the opportunity to administer only the first half of the test form to lower proficiency students. The wind-down allows the examinee to respond to a less challenging task at the end of the assessment.

Because the SOPI is administered via a recording and a test booklet, the test controls the amount of time examinees have to read the directions (following along with the directions being read aloud), plan their response, and deliver their response. For any SOPI form, all examinees are administered the exact same tasks. The specific task directions are provided in English, the examinees' native language, in order that specific nuances in the directions that ensure the elicitation of ratable speech samples will be understood. In cognitive labs with examinees, researchers have discovered that comprehension of these subtle nuances is critical to ensuring that examinees know exactly what is expected of them and in scaffolding their mental attitude towards accomplishing the simulated task in the expected manner. However the prompt, to which the examinees' reply is a logical rejoinder, is always in the target language of the test.

The feasibility study on the COPI explored how limitations on the SOPI might be overcome. Unlike the SOPI, which cannot be stopped or changed once it begins, the COPI was designed to be more adaptable and tailored to the individual examinee in the following ways.

First, through the use of a self-assessment, the examinee can choose the level of the first task on the assessment. In other words, the examinee can state whether he or she wants to respond to a task designed to elicit functions at the ACTFL Novice, Intermediate, Advanced, or Superior level first. Tasks are selected by the computer from a large underlying pool of tasks.

Second, through the use of a large underlying pool of tasks, examinees can choose among topics they'd like to talk about to provide evidence of fulfilling the necessary language functions. That is, the underlying pool contains similar tasks (e.g., comparing) that address different topics.

Third, examinees can provide input into the level of challenge provided by every other task presented. They can select at a task at the same level of challenge (i.e., Novice, Intermediate, Advanced or Superior) as the last one they answered, at a lower level of challenge, or at a higher level. The computer has an underlying algorithm to ensure that, given the examinee's starting level, the examinee receives enough tasks at each level to ensure ample evidence is provided to establish a baseline and a ceiling for the examinee.

Fourth, within a maximum upper limit, the examinee can control how much time he or she uses to think about and plan a response. This allows the examinee to determine when he or she feels "ready" to give the response. While in the SOPI cognitive labs were used to determine a time limit that would be appropriate for the majority of examinees, some examinees could become nervous if they had finished thinking through what they were going to say and had lots of time left over, while others at times felt "rushed by the beep" alerting them it was time to deliver their response.

Fifth, again within a maximum upper limit, the examinee can control how much time he or she spends giving the response. Again, while in the SOPI response times were standardized and studies were determined to provide an appropriate time for the majority of examinees, students responded in cognitive labs that sometimes they felt uncomfortable when they had completed their response and there was still much time left. They felt that this seemingly long period of time indicated to them they probably hadn't said enough. Likewise, although examinees hear a five-second warning before time runs out on the SOPI, they sometimes felt that they were being cut off before they finished saying what they wanted to say and were unnerved by the warning tone.

Sixth, examinees were given a choice whether to hear the instructions to the Advanced and Superior level tasks in English or in the target language. This decision was made again on the basis of findings from across the SOPI projects at CAL. Lower-level examinees recognized that they might not have understood the instructions had they been in the target language, while some upper-level examinees felt that they would have felt more comfortable if they didn't experience going back and forth between the prompt and their response in the target language and the task instructions in English.

We now describe the 6 phases of the COPI.

1. Welcome/Information on the Purpose and Structure of the COPI

The purpose of this phase is to introduce examinees to the COPI and help them feel more at ease in the new testing environment.

2. Personal Information

In this phase examinees enter their personal data and are given an opportunity to check and correct any wrong information. The information provided by the examinees is used to identify the examinees and to ensure that the tasks presented to the examinees are tailored to the examinees' profiles. For example, in Arabic the word "you" carries gender. The underlying task pool thus contains both "male" and "female" versions of each target language prompt and task as appropriate. An algorithm in the COPI ensures that the male or female version of these tasks is presented to the examinee, depending on whether the examinee has identified him-or herself as a male or female, respectively.

3. Self-Assessment

The next phase is the examinee self-assessment. In the feasibility study, the examinee answered 18 questions, one at a time. Reviewers suggested this could be simplified and now in the operational version the examinee views a list of four sets of 5-8 "can-do" statements about functional tasks in the language. Each set of statements corresponds to one of the major ACTFL proficiency levels (Novice, Intermediate, Advanced or Superior). Proficiency levels are color-coded throughout the COPI and are labeled A, B, C and D. Examinees are asked to select the set of can-do statements that best matches their speaking ability. Based on the results of the examinee's self-assessment, the COPI suggests a level at which the examinee should begin (Kenyon and Malabonga, 2001).

4. Sample Task

In the fourth phase, the examinee is able to see and listen to a sample task designed for the examinee's self-assessed level of proficiency. The examinee can also listen to a sample performance on this task that was rated at the level of the task. The examinee can also respond to this task, listen to his or her own performance and compare it to the sample response (Kenyon and Malabonga, 2001). The examinee can then decide whether the selected task level is appropriate for him or her. If not, he or she may select a higher or lower ACTFL level at which to begin the test. The examinee can again listen to

a sample task and performance at the new level, as well as respond to the sample task at this level. After this experience with one or sample tasks, the examinee confirms his or her starting level.

5. Responding to Performance Tasks (The Actual Test)

An algorithm controls the ACTFL level of tasks presented to the examinee. The level of the examinees' first performance task is determined by the level chosen by the examinee at the end of the fourth phase. In the feasibility study, the underlying algorithm presented examinees a choice of three different speaking functions, each with their own topic, and kept track of examinee choices, such that once a topic or speaking function had been chosen, it was never repeated. (Note that in the current operational COPI, all three tasks have the same speaking function but different topics or content areas.)

As mentioned earlier, in opposition to the SOPI, the COPI provides examinees with some choice of topic or content area to talk about as they display their current level of proficiency, as well as some selection in the difficulty level of every other task. That is, the underlying computer algorithm ensures that examinees are presented with a minimum of four tasks at their self-assessment level, which allows raters to confirm the floor of the examinees' current proficiency level, and three at the next higher level (or next lower level if their self-assessment level is at Superior), to allow raters to confirm a ceiling (i.e., where their current proficiency level is not high enough to complete the task requirements according to the criteria of the ACTFL guidelines). Thus, examinees are administered a minimum of seven tasks. However, depending on the choices they make, examinees may get a maximum of 11 tasks. These numbers compare to 15 tasks on a typical full-range CAL SOPI.

Also as mentioned earlier, in this phase examinees can control their planning and response times and, for tasks at the Advanced and Superior levels, whether to receive the task instructions in English or in the target language.

6. Feedback about the Levels of the Tasks that Examinees Took and Closing

In the last phase, examinees are congratulated on completing the test and get feedback about the number of tasks to which they have responded and to which of the four levels the tasks corresponded.

4. Examinee Considerations

Kenyon & Malabonga (2001) and Malabonga, Kenyon and Carpenter (2005) present the results of research during the feasibility studies on examinee performance and attitudes toward the COPI. Their results are summarized in this section of the paper.

Fifty-five undergraduate students studying Arabic (15), Chinese (16) or Spanish (24) participated in the research. All students completed both a SOPI and a COPI in their language. In addition, a subset of the Spanish students were also administered a face-to-face ACTFL OPI. Due to technological difficulties, not all tests could be rated. The correlation between the SOPI and the COPI for the 46 students with both scores was .95. For the 16 Spanish students who had both a COPI and an OPI, the correlation was .92. Examinees scored very similarly across the tests. However, in the few cases of disagreement for the Spanish students' ratings, there was a tendency for the SOPI or COPI rating to be higher than the OPI.

In conducting research on examinee attitudes toward the COPI, CAL researchers focused on the comparison of examinee attitudes toward the COPI and SOPI for the 55 examinees across the three language who took both tests. The six categories of comparison were (1) opportunity to demonstrate strengths and weaknesses in speaking, (2) test difficulty, (3) test fairness, (4) nervousness, (5) clarity of instructions, and (6) representativeness of the performance.

Table 1 (originally Table 2 from Kenyon & Malabonga, 2001), presents the descriptive statistics on six four-point Likert scale items (4 = Strongly Agree, 3= Agree, 2= Disagree, 1 = Strongly Disagree) contained on a questionnaire administered separately on the COPI and SOPI. Each was administered to the subject following the administration of the respective test. Ratings were always more in the desired direction on the COPI than the SOPI (e.g., lower on number 4 for nervousness on the COPI and higher on number 5 for clarity on the COPI). However, paired t-tests showed that the only statistically significant difference was on question 2 about perceived test difficulty. Examinees felt that the SOPI was more difficult than the COPI.

Table 1 - Comparative Results on Questionnaire 1 (COPI and SOPI)

Question	**N**		**COPI**	**SOPI**
1. I feel I had the opportunity to adequately demonstrate both my strengths and my weaknesses in speaking (*language*) on the (*test*).	55	Mean	2.98	2.84
		Std. Dev.	.62	.66
2. I feel the (*test*) was difficult.	55	Mean	2.30	2.71
		Std. Dev.	.60	.74
3. I feel there were questions asked or speaking situations required in the (*test*) that were unfair.	55	Mean	1.80	1.89
		Std. Dev.	.56	.53
4. I felt nervous taking the (*test*).	55	Mean	2.24	2.49
		Std. Dev.	.84	.84
5. I feel the directions for the (*test*) were clear (i.e., I felt I always knew what I needed to do).	55	Mean	3.58	3.44
		Std. Dev.	.69	.67
6. I feel someone listening to my (*test*) responses would get an accurate picture of my current ability to speak (*language*) in real life situations outside the classroom.	52	Mean	2.77	2.73
		Std. Dev.	.67	.77

Table 2 (originally Table 3 from Kenyon & Malabonga, 2001) shows results on a questionnaire that asked the 55 subjects to directly compare the two test formats. Subjects completed this questionnaire after completing all tests. In this table, we see that examinees felt strongly that the COPI better allowed them to demonstrate their strengths and weaknesses in speaking (question 1) and would give someone listening to their responses a more accurate picture of their current ability (question 6), while they felt the SOPI was more difficult (question 2) and made them feel more nervous (question 4). While many subjects felt both the COPI and SOPI were about the same in terms of clarity (question 5) and fairness (question 3), there was an edge for the COPI in these two areas among the remaining students who indicated one over the other.

Table 2. Comparative Results on Questionnaire 2 (COPI and SOPI)

Question	**COPI**	**SOPI**	**BOTH**	**Missing**
1. Which technology-based test did you feel better allowed you to demonstrate both your current strengths and weaknesses in speaking (*language*)?	60.0%	25.5%	14.5%	0.0%
2. Which technology-based test did you feel was more difficult?	29.1%	56.4%	14.5%	0.0%
3. Which technology-based test did you feel was fairer?	47.3%	20.0%	32.7%	0.0%
4. Which technology-based test did you feel more nervous taking?	32.7%	47.3%	18.2%	1.8%
5. Which technology-based test did you feel had clearer directions?	34.5%	10.9%	54.5%	0.0%
6. Which technology-based test do you feel would give someone listening to your performance a more accurate picture of your current ability to speak *(language)* in real life situations outside the classroom?	45.5%	29.1%	21.8%	3.6%

Comments made by examinees on their questionnaires, as reported in Kenyon and Malabonga, 2001, indicated that the enhancements made on the COPI were having the desired effects. The following comments are illustrative:

- I choose the COPI because, as I wrote on the feedback sheets to the test, the SOPI left me feeling greater pressure in having to respond in a restricted time.
- I was more comfortable with the COPI because it allowed me to move at my own pace and took proficiency –level [according to the student] into account.
- The SOPI was more difficult because you were unable to choose your situation and the times in which to start/finish speaking.
- SOPI [was more difficult]. This is because the SOPI tested all levels, whereas the COPI was more realistic and enabled me to show what I knew.
- COPI allowed me to choose my level. I never felt inferior.
- COPI allowed choice allowing speaker to demonstrate in areas they feel comfortable with.
- COPI—I was able to apply more of the vocabulary I know by choosing questions rather than being forced to answer certain ones.

On the other hand, not all examinees felt comfortable with choices. Kenyon and Malabonga, 2001, report on two interesting comments. One student complained that that the COPI offered too many choices. Another felt that because the SOPI didn't demand interaction (i.e., clicking on the screen to "continue"), they could "just relax and concentrate on Spanish."

It should also be noted here that for the Spanish students who could compare the three tests (COPI, SOPI, and face-to-face OPI), there was still a definite preference for the OPI over either technologically-mediated assessment in terms of feeling that the OPI provided them a better opportunity to demonstrate their strengths and weaknesses in speaking Spanish (question 1), and that their performances on the test would provide a more accurate picture of their current ability to speak Spanish outside of the classroom (question 6). From these results and examinee comments, Kenyon and Malabonga (2001) remark that "it appeared that both technologically-mediated tests were unable to replicate the interactive, conversational and personal nature of the face-to-face interview for the Spanish examinees who had experienced all three tests."

Malabonga, Kenyon, and Carpenter (2005) present further results from the feasibility study that sheds light on the functioning of the COPI. Regarding the self-assessment, they report that the correlation between the self-assessment administered as the third phase of the COPI and the results on the COPI

was .88 for the 50 subjects for whom complete data was available, with 54% agreement on the exact or adjacent sublevel rating. The vast majority of subjects (89%) in the study started the COPI with a task at the level suggested by their self-assessment. Of the six subjects who didn't, three started at one level below and three started at one level above. Two of the three who started lower were recommended to start with a task at the highest level, Superior.

Analyses of the tasks at which students started the COPI and their final ratings revealed that none of the examinees produced a set of responses to tasks which did not clearly show their floor and ceiling. Indeed, the analysis indicated that the starting level chosen by examinees tended to be a conservative lower boundary on actual examinee proficiency, which was the intent of the design.

In terms of the use of the planning time, Malabonga, Kenyon and Carpenter (2005) show that its use varied by global proficiency level of the examinees, as well as by level of the COPI task. As expected, students on average used the least amount of planning time for Novice-level tasks (18.81 seconds), then Intermediate-level tasks (26.37 seconds), then Advanced-level tasks (27.73 seconds) and the most time for Superior-level tasks (32.94 seconds). However, in terms of the global proficiency level, examinees at the highest two sublevels (Superior and Advanced-High) used ***less*** time to plan on average (17.92) than students in the Advanced sublevels (Advanced-Mid, Advanced-Low) who used 31.08 seconds to plan for their tasks on average, and students in the Intermediate sublevels (Intermediate-High, Intermediate-Mid, and Intermediate-Low), who used on average 27.49 seconds to plan their responses. Only students in the Novice sublevels (Novice-High, Novice-Mid and Novice-Low) used less time on average (11.41 seconds). These results indicate the wide variety of differences in the use of planning time. The short amount of time used, on average, by high proficiency students may be indicative of the automaticity they have acquired in the language.

Similarly, response times varied widely. Tasks at the Novice level had on average the shortest response times (31.73 seconds), followed by tasks at the Intermediate-level (57.86 seconds), then tasks at the Advanced level (100.27 seconds), with tasks at the Superior level having on average the longest response times (141.11 seconds). This time, however, the subjects with the highest global proficiency levels (Superior and Advanced-High) used the greatest amount of response time on average (140.37), with each subsequent group of lower proficiency using less response time. Students with performances at the Advanced sublevels (Advanced-Mid, Advanced-Low) used 96.17 seconds on average to respond, while students in the Intermediate sublevels (Intermediate-High, Intermediate-Mid, and Intermediate-Low) used 52.22 and students in the Novice sublevels (Novice-High, Novice-Mid and Novice-Low) used 33.34. In further analyses on the interaction between global proficiency

level of the examinee and response time, it was clear that on average, at tasks at the same level, examinees with higher proficiency ratings provided longer responses than examinees with lower proficiency ratings, and for examinees at the same global proficiency level, longer responses were given to tasks targeted at higher proficiency levels than those targeted at lower proficiency levels.

Interestingly, when examinee use of planning and response time on the COPI was compared to the planning and response time provided on the SOPI, planning time on the COPI was generally about 10 seconds more than that provided on the SOPI and response time used on the COPI for Superior and Advanced tasks was much longer than the time provided on the SOPI. Average COPI response time for Superior-level tasks was 142.12 seconds, while the SOPI provided only 88.67 seconds for Superior-level tasks. Likewise, average COPI response time for Advanced-level tasks was 100.27, compared to 71.75 seconds provided on the SOPI. Only for Intermediate level tasks did the SOPI provide more time (64.57 seconds) than the average response time for Intermediate level tasks on the COPI (57.86 seconds). While great care went into determining the allotted times on CAL's SOPIs, including collecting feedback from hundreds of examinees on their feeling of the adequacy of the planning and response times provided, it appears that when allowed to choose their planning and response times, examinees generally take more time than the levels provided for by the SOPI. This may be one reason why subjects in the feasibility study generally felt less nervous on the COPI and generally felt the SOPI was more difficult.

On the whole, while not affecting overall student ratings, we were pleased that the proposed improvements to technologically-mediated speaking tests provided by the COPI model in CAL's studies, when compared to CAL's SOPI, appeared to be noted and taken advantage of by examinees. We believe that by providing some measures of examinee choice into the test administration, we have taken some steps to "humanize" the test-taking experience.

5. Rater Considerations

In this section, we briefly want to comment on what we feel are improvements in the rating experience provided by CAL's COPI model versus both CAL's SOPI model and the OPI. For face-to-face or telephonic ACTFL OPIs conducted in live time, the first rating is conducted as the interview takes place; all ACTFL OPIs are also audio-taped for later confirmation by a second rater. The second rater must listen in real time to the entire length of the interview, which can be a time consuming process as interviews, depending on the proficiency level of the examinee, can range from 15 to 45 minutes.

CAL's cassette-based SOPIs, likewise, require the rater to listen to up to 45 minutes of tape. Although SOPI many raters fast-forward through the directions and non-examinee response parts of the examinee response tape, many reported that the time required to listen to and fast-forward through the examinee response tape was still nearly 45 minutes. In addition, time was lost if the rater wished to re-listen to a particular task response and under- or over-rewound the cassette tape and spent additional time hunting for the examinee response.

In developing the COPI, CAL designed a rating interface that addressed challenges reported by raters of the cassette-based SOPI. The COPI rating interface thus presents a new approach to rating compared with traditional OPI and cassette approaches. Responses are also digitally stored for immediate or later retrieval, so rating can be accomplished at any time and at an individually determined pace. The computer interface displays examinee responses as separate files, named by examinee and by task. This means that, for each examinee, raters can listen to task responses in any order and not necessarily in the order the examinee took the tasks. Also, with a click on a file name raters can replay examinee responses for a particular task. They can also easily return to any previously rated tasks for review.

In addition, the computer can easily provide raters with quick access to materials that will facilitiate the rating process. For example, for the task currently being rated, raters can read the task directions on the computer screen and listen to the instructions and the target language prompt were presented to the examinee. They also see any accompanying graphics that the examinee would have seen. In addition, the rater can easily call up the ACTFL criteria (i.e., the task-level expectations) for any task, and, as provided, audio benchmark scoring samples against which to compare the examinee's response. To provide more informative feedback, raters can also enter notes to examinees on this screen in order to provide general comments and/or task-specific feedback to the student.

In scoring the SOPI and COPI, the performance on each task is separately rated. Then, a rather complex algorithm is used to compute the examinee's overall global proficiency level from the individual task ratings. While this is somewhat tedious to do by hand, the COPI rating program automatically calculates a global rating, error-free, based on the set of individual task-level ratings. Because this global rating is based on consistency of performance across tasks at a given ACTFL level, the technology of the COPI rating program can increase rating efficiency by allowing raters to listen only to those tasks that are necessary to give an accurate assessment of the examinees' current proficiency level. Because the tasks do not need to be scored in order of administration, raters can facilitate their rating task by, for example, beginning with the most challenging tasks to determine the

examinee's performance ceiling (Malabonga & Kenyon, 1999). The efficiency produced by these innovations in the COPI rating program (versus the traditional approach to scoring the SOPI) was demonstrated during the feasibility study discussed above. According to Malabonga, Carpenter, and Kenyon (2002, December), the time to score the COPI tests was on average *three times shorter* than the time required to score the SOPI tests in that study.

The rating of CAL's SOPIs in educational settings for institutional use has been facilitated by cassette and manual materials known as Self-Instructional Rater Training Kits (Kenyon, 1997). These are available in Spanish, French, German, Russian, Chinese, and Arabic. CAL researchers have taken advantage of the computer in increasing the efficiency and effectiveness of these programs through CAL's Multimedia Rater Training Programs (MRTP), available in Spanish, French and German. This software provides background information on the structure of the COPI and the ACTFL rating scale, samples of pre-rated student responses to COPI tasks, tips for rating by scoring level and by task, calibration sets at all levels, and an audio reference library of benchmark performances. The rater training is self-contained and self-paced with adaptive exercises that provide plenty of scoring practice. Program users are provided with activities to become familiar with the rating scale and are able to listen to and rate more than 200 authentic examinee responses that have been pre-rated by certified SOPI and ACTFL OPI raters. Accessed via CD-ROM or downloadable file hosted online, these programs provide a convenient and effective means for interactive rater training without the time and travel costs inherent to live rater training workshops.

The use of the COPI model, with the automatic digitization of examinee responses and the computerized collection of rating data, can facilitate the rapid development of scoring programs similar to the MRTP for other COPI languages and in other contexts.

6. Test developer considerations

The use of a technologically-mediated speaking assessment with clearly delineated tasks, such as the SOPI or COPI, provides the opportunity for tailored research and development into optimizing the characteristics of the tasks to enable examinees to give their optimal performance. Unlike the OPI, in which examinees can ask for clarification if they do not understand what to do, or interviewers can modify their questions if an initial one is unsuccessful, in the SOPI (and COPI) the examinee only gets one chance to understand what he or she must do. Thus, in the early days of SOPI development, there were many cognitive labs held with students in each new language in which the SOPI appeared, as well as larger-scale questionnaires completed by test-takers (e.g., Kenyon and Stansfield, 1993), in

addition to evaluations of raters. Early descriptions of the characteristics of SOPI tasks, drawn from this research, can be found in Pavlou and Rasi (1994) and Stansfield (1996). In this section, we illustrate further research that facilitates test development with an example from the *CAST Project*, which followed the COPI feasibility study and is described below. Again, the context for this project is the assessment of speaking within the framework of the ACTFL *Speaking Proficiency Guidelines* in which the ACTFL OPI is regarded as a "gold standard."

As discussed earlier, beginning with its first SOPIs in the 1980s, CAL has amassed a great deal of data, knowledge and experience about test takers' experience with both the technologically-mediated semi-direct assessment and the face-to-face direct oral proficiency interview. The *CAST Project* (2002-2005), under the direction of Margaret Malone, allowed CAL researchers, working in cooperation with San Diego State University, ACTFL, Brigham Young University and the Defense Language Institute, to integrate findings from over 20 years of experience in developing semi-direct tests of speaking skills. The decades of experience included both successes of the semi-direct approach as well as challenges identified by stakeholders.

The main goal of the project was to address how to develop the most effective technologically-mediated oral proficiency test possible within the ACTFL framework. The study examined the characteristics of successful oral proficiency tasks on both direct (OPI) and semi-direct (SOPI and COPI) tasks, as well as the functions necessary to produce a ratable sample at each ACTFL proficiency level. This research is described in the next section.

The CAST Project

The project team endeavored to understand the components of a successful oral proficiency test task within the ACTFL speaking proficiency framework and what speaking functions are necessary to elicit a ratable sample at each major ACTFL level (Malone and Rasmussen, 2006). The project thus examined tasks both independently and as a part of a test framework that would produce sufficient evidence of examinee performance at each proficiency level. The team defined successful tasks as those that allow a speaker at a specific ACTFL level of proficiency (e.g. an Advanced-level speaker) to give a performance demonstrating the most salient features that raters need to see to confirm a rating at that proficiency level (e.g. a rating at one of the three Advanced sublevels: Advanced-Mid, Advanced-Low, and Intermediate-High) on a task designed to target that proficiency level (e.g. an Advanced-level task). Conversely, using our example, unsuccessful Advanced-level tasks were those whose characteristics limit the ability of an Advanced-level speaker to demonstrate the most salient rating

features of Advanced-level proficiency level, although the task was intended to be at the Advanced level. These definitions gave rise to three fundamental research questions:

- Within the ACTFL context, what are the overall characteristics of successful assessment tasks; that is, tasks that elicit optimal performances from speakers whose proficiency is on a par with the level of the task? What are the specific characteristics of successful tasks at each proficiency level?
- What are the overall characteristics of unsuccessful assessment tasks within the ACTFL context; that is, tasks that fail to elicit optimal performances from speakers whose proficiency is on a par with the intended level of the task? What are the specific characteristics of unsuccessful tasks at each proficiency level?
- Which speaking functions are necessary (mandatory) to produce a ratable performance sample (i.e., one with enough clear evidence of the speaker's proficiency level) at each major ACTFL proficiency level?

To answer these questions, the research team collected and examined tasks and accompanying performances at the four major proficiency levels from completed OPI, SOPI, and COPI assessments in Arabic and Spanish. For this study, 27 Spanish OPIs, 20 Arabic OPIs, 17 Arabic COPI and SOPI tasks, and 18 Spanish COPI and SOPI tasks, along with examinee responses, were analyzed. In a SOPI or COPI, the term *prompt* refers to the question or statement to which the examinee responds (Malone and Rasmussen, 2006). The question or statement provides the content and context for the task, and frames the response in terms of the function of the task to which the examinee is responding as well as the expected discourse type. On the SOPI/COPI, the prompt is one-way and non-negotiable; the examinee cannot ask for clarification, repetition or restatement. A task is thus identifiable as the (single) prompt and the examinee's response to this prompt.

An OPI prompt is different from a SOPI/COPI prompt because the OPI is a two-way conversation between the examinee and the interviewer. As in the SOPI/COPI, the OPI prompt must include content and context, as well as an explanation of the function and discourse type expected in the response, but this information may be novel to the prompt at hand or a continuation of the previous part of the interview. More significantly, in an OPI the examinee can ask for clarification, repetition, or re-statement, and the interviewer can clarify, repeat, or restate the prompt if the examinee is not responding as the interviewer intended. Because an OPI prompt may be negotiated between the interviewer and examinee, it may consist of one statement or question or a series of statements, questions, clarifications, repetitions, and restatements that ultimately yields a ratable response. The 47

OPIs were divided into separate prompts and responses, with guidance from the entire research team. Each prompt and response comprised a task that could be compared the SOPI and COPI tasks for the analysis.

To select SOPI and COPI tasks for analyses, the researchers examined 10 years' worth of feedback from examinees and raters on Arabic and Spanish SOPI tasks, as well as feedback from raters, examinees, and students who participated in the feasibility study of the Arabic and Spanish COPI. Based on this feedback, researchers selected those tasks that were deemed either highly successful or highly unsuccessful at eliciting speech at the targeted proficiency level. The nature of SOPI and COPI tasks meant that these could be subdivided by task function at each proficiency level. All SOPI and COPI tasks selected for analysis had at least two to three examinee responses associated with it.

When all of the prompts and responses had been identified, they were uploaded into a Web-based database for the collection of evaluative data. Raters certified to conduct ACTFL OPIs interacted with the database through a specially designed interface which presented one task at a time. It showed the text of the prompt in the target language and in English and asked the rater to identify prompt level; target function; content area/topic; context (formal or informal); and discourse type from a drop-down list. The interface then asked the rater to evaluate the prompt as to whether sufficient background information was given and whether the prompt was appropriate for the function the rater had identified.

Next, each rater was asked to evaluate each examinee response, stating whether it was below the intended ACTFL level of the prompt, approaching the level, on level, or above the level, and whether the determining factor in that rating was function, discourse, or accuracy. Finally, if the response was not at the appropriate proficiency level, the rater was asked to judge whether that was due to problems with the prompt that hindered the examinee's performance, or whether the response clearly showed the examinee was not yet at that level of proficiency.

Using the interface, raters entered their evaluations into a prompt evaluation database. The CAL analysis team then used the following process to analyze raters' evaluations:

1. Identified prompts on which at least two of the three raters agreed on the prompt level (Novice, Intermediate, Advanced, Superior). If the prompt level was identified differently by all three raters, the prompt was discarded as an unsuccessful prompt.

2. Identified the function(s) that raters had assigned to the prompts identified in step 1. If raters agreed on the function, that function was identified as consistent with the proficiency level. If raters chose different functions, staff referred to the transcript. If the functions were similar, the staff proceeded to step 3. If the two functions appeared very different, that function was not included at that proficiency level.
3. Identified the content area(s) assigned by the raters to the functions identified in step 2. If raters agreed on the content, that content area was identified as consistent with the proficiency level. If raters chose different content areas, staff referred to the transcript. As with the functions, if the content areas were similar, staff identified those areas as consistent with the proficiency level, and if the content areas were markedly different, they were not included.

This analytical process allowed the research team to identify criteria for successful and unsuccessful SOPI and COPI tasks across proficiency levels and to identify the features of successful and unsuccessful tasks at each level. As a result, CAL researchers were able to eliminate from the COPI task pool which had been developed for the feasibility study tasks that were deemed unsuccessful. In addition, CAL researchers were able to revise COPI tasks, based on the analysis of the evaluative data collected by this study, to improve COPI tasks and optimize elicitation of ratable speech samples. In addition, CAL was able to examine the functions common to ratable oral proficiency performances across the three types of tests (OPI, SOPI and COPI). From these data, the researchers were able to develop a framework for developing technologically-mediated speaking tasks that provide examinees the appropriate opportunities to demonstrate the functions, discourse, and accuracy necessary to produce a ratable sample at each ACTFL proficiency level.

The development of the first operational COPIs in Spanish and Arabic proceeded from both the findings of the COPI feasibility study, which itself had been built based on 15 years experience with the SOPI, and the *CAST Project*. From the feasibility study, the maximum amount of planning and response time in the operational COPI was determined based on examinee and rater feedback, as well as the results of Kenyon, Malabonga and Carpenter (2005). Similarly, examinee choice with regard to task topic and difficulty were maintained in the operational version. Additional features explored in the feasibility study, such as the option to listen to and read test directions in the target language for test takers with high levels of proficiency, were also incorporated into the final versions. Of equal importance, the results of the *CAST Project* allowed CAL researchers to incorporate both feedback from examinees and raters into the task and test design to optimize the opportunity for examinees to demonstrate their current level of proficiency. Finally, the research conducted on the *CAST Project* allowed CAL to determine which functions were necessary to comprise a set of tasks at each major ACTFL proficiency level.

Although for the COPI the definition of speaking proficiency is provided by the ACTFL *Speaking Proficiency Guidelines*, any assessment of speaking skills needs to start with a clear definition of what is understood by "speaking" and how it will be operationalized in the assessment. The *CAST Project* provides one example of an approach to refining the development of technologically-mediated speaking tasks. This approach incorporated an analysis of tasks, an evaluation of examinee responses, and the elicitation of evaluative data by experts to define and identify task characteristics that allow examinees to provide raters with the evidence required for raters to evaluate their speaking performances on the basis of the assessment's operationalization of what speaking is. Using such analytical approaches help test developer optimize the elicitation of speech performances given constraints due to the use of technological media.

7. Researcher Considerations

In addition to facilitating the large-scale administration of a speaking assessment, the use of a standardized set of tasks as appears on the SOPI and COPI has the potential to facilitate research. For example, the SOPI has been used to as a pre- and post-test speaking proficiency assessment in a project to study gains in oral proficiency in a study abroad program (Vande Berg, Connor-Linton and Paige, 2009). The use of standardized tasks facilitated the rating of student performances in the research study; in addition, it allowed for qualitative analysis of the comparable linguistic performances across tasks and students.

While researchers at CAL have not yet been able to mine the full possibilities of the COPI approach to facilitate research into the development of speaking skills of students, we believe that the use of the COPI, with its automatic digitization of student responses and easy storage and retrieval of sound files, together with databasing ratings and rater comments on each task, has the potential to facilitate research on student oral proficiency. The standardization of the COPI tasks sets it apart from the OPI, in which many factors play a role in variations among performances (see, for example Brown, 2003). With the COPI methodology, student performances can be qualitatively and linguistically analyzed across tasks, across contexts of student learning and student language background, and across raters (e.g., how raters apply the scale). Insight into examinee proficiency may also be gained by analyzing examinee behavior on tasks (e.g., amount of planning time and response time used, language selection of the task instructions). With the computerized COPI rating, rater behavior, such as how often raters listen to benchmark samples, how much time raters spend on the rating task, and so on, can be easily tracked for analysis that may give insight into improving rater training, differences in the behavior

between new versus experienced raters, and other similar issues. Finally, as seen in the *CAST Project*, the ease with which data is captured on the COPI can facilitate research on the tasks themselves.

Perhaps of more importance, however, is that fact that the use of standardized tasks, as presented in the COPI model, may make it easier to develop a new generation of automatic oral proficiency scoring on constrained tasks. This potential is suggested, for example, by Levow and Olsen (1999). Many programs currently exist to electronically score written response to essay tasks, examining a variety of textual and linguistic features using natural language processing. When the time comes that digitized examinee speech can be reliably, accurately, and cost-effectively converted into text, then linguistic features beyond the physics of the speech sounds that are important in communication may be able to be analyzed by computer. Databases of large numbers of examinees replying to the same tasks, as can be collected by the use of the COPI model, may help facilitate research into this development.

8. Conclusion and discussion

In this paper we have presented research and development work at CAL on a computerized oral proficiency assessment that is based in a very specific historical context of assessing speaking skills. Since the early 1980s, researchers at CAL have explored using technology to provide another means to assess the speaking proficiency of foreign language learners according to the criteria of the ACTFL *Speaking Proficiency Guidelines*, a means which addresses some of the daunting logistics of large-scale face-to-face oral testing. While the use of the ACTFL guidelines to assess speaking in the United States is ubiquitous, we recognize that there are also criticism as to their validity, as Norris (2001) writes, "for informing interpretations about learners' language abilities or for making decisions and taking actions within language classrooms and programs." With Norris, we support further investigation in the area of the use of such scores.

We also agree with Norris (2001) in his review of the COPI, that "language test developers need to begin their deliberations about speaking assessment not by asking what computers are capable of doing, but rather by asking (a) what kinds of interpretations actually need to be made about L2 speaking abilities; (b) what kinds of evidence a test will need to provide in order to adequately inform those interpretations; and (c) what kinds of simulation tasks will provide the required evidence." This is the crux of evidenced-centered assessment design (Mislevy, Steinberg, & Almond, 1999; 2002), which forms the foundation of test-development activity at CAL. Whether speaking is assessed face-to-face or through a technologically-mediated modality, test developers must ensure that the speaking tasks and rating protocol allow raters to evaluate what is truly valued and of import for the interpretations and actions that may be made on the bases of the test scores. As we stated at the

beginning of the article, speaking is multi-faceted and at present no assessment can capture all its richness.

Nevertheless, we hope that the COPI can stimulate ideas about what can be done using computer-technology to assess speaking skills. Indeed, it may be beneficial to consider which aspects of second language speaking abilities can best be captured through currently available automated scoring technologies and then design COPI-like tasks, with appropriate scoring rubrics, to capture important aspects of speaking skills over and above those that can be elicited with tasks that can use automated scoring technology. A combination of the use of automated scoring technology with a few choice rater-scored COPI-like tasks targeted to provide additional information about students' speaking skills may make for a much richer speaking assessment given the current state of knowledge and provide richer data for research and analysis.

References

American Council on the Teaching of Foreign Languages. (1999). *ACTFL Proficiency guidelines – Speaking (Revised, 1999).* Yonkers, NY: American Council on the Teaching of Foreign Languages.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition, 10*(2), 149-164.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal, 70*(4), 380-390.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-25.

Chalhoub-Deville, M. (1997). The Minnesota Articulation Project and its proficiency-based assessments. *Foreign Language Annals, 30*(4), 492-502.

Clark, J. L. D. (1988). Validation of a tape-mediated ACTFL/ILR-scale based test of Chinese speaking proficiency. *Language Testing, 5*(2), 197-205.

Hadley, A. O. (1990). The concept of proficiency and its impact on foreign language teaching programs; Le Concept de competence fonctionnelle et son impact sur les programmes et l'enseignement des langues etrangeres. *Etudes de Linguistique Appliquee, 77*, 85-96.

Iwashita, N. (1999.) The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing* 8(1), 51-66.

Johnson, M. (2000). Interaction in the Oral Proficiency Interview: Problems of validity. *Pragmatics, 10*(2), 215-231.

Johnson, M. (2001). *The art of non-conversation: A reexamination of the validity of the Oral Proficiency Interview*. New Haven: Yale University Press.

Kenyon, D. M. (1997). Further research on the efficacy of rater self-training. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 257-273). Jyväskylä, Finland: University of Jyväskylä.

Kenyon, D. M. (2000a). Enhancing oral proficiency assessment through multimedia: A model, applications, and research needs. In E. Tschirner, H. Funk, and M. Koenig (Eds.), *Schnittstellen: Lehrwerke zwischen alten und neuen Medien* (pp. 171-201). Berlin, Germany: Cornelsen Verlag.

Kenyon, D. M. (2000b). Tape-mediated oral proficiency testing: Considerations in developing Simulated Oral Proficiency Interviews (SOPIs). In S. Bolton (Ed.), *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar* (pp. 87-106). Köln (Cologne), Germany: Gilde Verlag.

Kenyon, D. M. & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology 5*(2), 60-83.

Kenyon, D. M. & Stansfield, C.W. (1993). A method for improving tasks on performance-based assessments through field testing. In Huhta, A., Sajavaara, K. and S. Takala (Eds.), *Language Testing: New Openings* (pp. 90-102). .Jyväskylä, Finland: Institute for Educational Research.

Kenyon, D. M. & Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *The Modern Language Journal 84* (85-101).

Lantolf, J. P. & Frawley, W. (1985). Oral proficiency testing: a critical analysis. *Modern Language Journal 69 (4), 337-349.*

Lantolf, J. P. & Frawley, W. (1988). Proficiency: Understanding the construct. *Studies in Second Language Acquisition, 10*(2), 181-195.

Lee, Y. (2007). The multimedia assisted test of English speaking: The SOPI approach. *Language Assessment Quarterly, 4*(4), 352-366.

Levow, G. & Olsen, M. B. (1999). Modeling the language assessment process and result: Proposed architecture for an automatic oral proficiency assessment. In M. B. Olsen (Ed.), *Computer mediated language assessment and evaluation in natural language processing: Proceedings of a symposium sponsored by the Association for Computational Linguistics and International Association of Language Learning Technology* (pp. 24-31). New Brunswick, NJ: Association for Computational Linguistics.

Liskin-Gasparro, J. (1987). Testing and teaching for oral proficiency. Boston: Heinle & Heinle Publishers, Inc.

Malabonga, V. Carpenter, H., & Kenyon, D. (2002, December). *Computer assisted rating: Reliability, efficiency and perceptions on the COPI.* Paper presented at the 24th Annual Language Testing Research Colloquium, Hong Kong, PRC.

Malabonga, V., & Kenyon, D. (1999). Multimedia computer technology and performance-based language testing: A demonstration of the Computerized Oral Proficiency Instrument (COPI). In M. B. Olsen (Ed.), *Computer mediated language assessment and evaluation in natural language processing: Proceedings of a symposium sponsored by the Association for Computational Linguistics and International Association of Language Learning Technology* (pp. 16-23). New Brunswick, NJ: Association for Computational Linguistics.

Malabonga, V. & Kenyon, D. (2000). Multimedia Performance Based Language Assessment: The Computerized Oral Proficiency Instrument (COPI). In J. Bourdeau & R. Heller (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2000* (pp. 1442-1443). Chesapeake, VA: AACE.

Malabonga, V., Kenyon, D. M, & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing, 22*(1), 59-92.

Malone, M. & Rasmussen, A. (2006) Computer Assisted Screening Tool (CAST) Framework: *Guidelines for the development of level-specific oral proficiency assessment tasks* Unpublished manuscript. Center for Applied Linguistics: Washington, DC.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd. ed., pp. 13-103). New York: Macmillan.

Milleret, M., Stansfield, C. W., & Kenyon, D. M. (1991). The validity of the Portuguese speaking test for use in a summer study abroad program. *Hispania, 74*(3), 778-787.

Mislevy, R.J., Steinberg, L.S. and Almond, R.G. (1999). Evidence-centered assessment design. Princeton, NJ: Educational Testing Service.

Mislevy, R.J., Steinberg, L.S. and Almond, R.G. (2002). Design and analysis in task-based language assessment. Language Testing, 19 (4), 477-496.

Norris, J. M. (2001). Concerns with computerized adaptive oral proficiency assessment. *Language Learning & Technology, 5*(2), 99-105.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing* 19(3), 277-295.

Pavlou, P. Y., & Rasi, S. B. (1994). *Considerations in writing item prompts for Simulated Oral Proficiency Interview (SOPI)*. Paper presented at the Annual language Testing Research Colloquium, Washington, DC. (ERIC Document Reproduction Service No. ED 371 622).

Shohamy, C. W., Gordon, C., Kenyon, D. M., & Stansfield, C. W. (1989). The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Higher Hebrew Education, 4*, 4-9.

Stansfield, C. W. (1996). *Test development handbook: Simulated Oral Proficiency Interview.* Washington, DC: Center for Applied Linguistics.

Stansfield, C. W., & Kenyon, D. M. (1991). *Development of the Texas Oral Proficiency Test (TOPT).* Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 332 522).

Stansfield, C. W. & Kenyon, D. M. (1992a). The development and validation of a simulated oral proficiency interview. *The Modern Language Journal, 76*, 129-141.

Stansfield, C. W., & Kenyon, D. (1992b). Research on the comparability of the Oral Proficiency Interview and the Simulated Oral Proficiency Interview." *System, 20*(3), 347-364.

Stansfield, C. W., & Kenyon, D. (1993). Development and validation of the Hausa speaking test with the ACTFL proficiency guidelines. *Issues in Applied Linguistics, 4*(1), 5-31.

Stansfield, C. W., Kenyon, D. M., Paiva, R., Doyle, F., Ulsh, I., & Cowles, M. A. (1990). The development and validation of the Portuguese speaking test. *Hispania, 73*(3), 641-651.

Taylor, L. (2000). Investigating the paired speaking test format. *Cambridge ESOL Research Notes, 2*, 14–15.

Taylor, L. (2001). The paired speaking test format: Recent studies. *CambridgeESOL Research Notes, 6*, 15–17.

TestDaF-Institut 2002: *The TestDaF: a new measure of German language proficiency*. Brochure. Hagen, Germany: Test-DaF-Institut.

Uber-Grosse, C., & Feyten, C. M. (1991). Impact of the proficiency movement on Florida. *Hispania, 74*(1), 205-209.

The authors:
Dorry M. Kenyon
Center for Applied Linguistics
4646 40th Street, NW
Washington, DC 20016
E-mail: dkenyon@cal.org

Dorry M. Kenyon holds a Ph.D. in Educational Measurement and Applied Statistics from the University of Maryland. He is the Director of the Language Testing Division at the Center for Applied Linguistics (CAL) in Washington, DC, USA. At CAL he advises or directs a variety of projects related to the assessment of the English language and foreign language skills of language learners spanning the ages of pre-school to adult. He also serves as CAL's chief psychometrician. Since joining CAL in 1987, he has worked in all aspects of designing, developing, validating, and operationalizing language tests through many large state and national projects. With a particular interest in using technology in the assessment of speaking skills, he worked on CAL's *Simulated Oral Proficiency Interviews* (SOPIs) and led the initial development of CAL's *Computerized Oral Proficiency Instruments* (COPIs). He also served as the lead developer of CAL's *BEST Plus,* a computer-assisted oral assessment for adult English language learners.

Margaret E. Malone
Center for Applied Linguistics
4646 40th Street, NW
Washington, DC 20016
E-mail: mmalone@cal.org

Margaret E. Malone (Ph.D., Linguistics, Georgetown University) is Senior Testing Associate at the Center for Applied Linguistics and Co-Director of the National Capital Language Resource Center. She directs projects on assessment research, technology-mediated tests of oral proficiency and provides professional development on language assessment to teachers. Her projects include survey research on user beliefs about the TOEFL, developing assessment literacy for language instructors, evaluating the effectiveness of a national effort to teach critical languages during the summer and developing assessments for US short-term language program. Before re-joining CAL in 2000, she directed language testing for Peace Corps-Worldwide and designed and delivered technical assistance to language programs in six U.S. states. She has taught graduate level courses in language testing and teaching methods at Georgetown and American Universities. A member of the Editorial Review Board of the Language Assessment Quarterly, she is the co-founder of the East Coast Organization of Language Testers.

FACE-TO-FACE AND COMPUTER-BASED ASSESSMENT OF SPEAKING: CHALLENGES AND OPPORTUNITIES

Evelina D. Galaczi

University of Cambridge ESOL Examinations

Abstract

Computer-based assessment of speaking has become more widespread in the last five years and has presented an alternative and a complement to the more traditional 'direct' face-to-face approach to speaking assessment. This paper briefly overviews the current academic literature on face-to-face and computer-based assessment of speaking, and explores the test features of these two different test modes from the perspective of two test-quality frameworks, namely Bachman & Palmer's (1996) 'test usefulness' and Weir's (2005) socio-cognitive framework for test validation. The paper discusses the advantages and challenges offered by computer-based speaking assessment and provides an illustration of a CB-delivered oral test (Cambridge ESOL's BULATS online speaking test). An argument is made for the fundamental importance of contextualizing the debate within the notion of 'fitness for purpose' of a given test.

Introduction

Computers and related technology have acquired considerable importance in language assessment in the last few decades, and there is no doubt that the use of computer-based (CB) tests will become even more predominant in the future. The newest addition to the CB assessment field has been the assessment of speaking, largely influenced by the increased need for oral proficiency testing and the necessity to provide speaking tests which can be delivered quickly and efficiently whilst maintaining high-quality. Computer-based assessment of speaking presents a viable alternative and a complement to the more traditional face-to-face approach to speaking assessment and is gaining in importance in the last five years, as seen in the introduction of CB speaking tests by several large examinations boards (e.g., ETS's TOEFL iBT speaking test, Pearson's PTE Academic test, and Cambridge ESOL's forthcoming BULATS online speaking test). This paper will address the issues surrounding the use of technology in the assessment of speaking (both in terms of delivery and scoring) and its impact on test qualities. These issues will be explored with reference to theoretical conceptualisations of speaking ability (Bachman & Palmer 1996; Canale & Swain 1980; Canale 1983) and to frameworks for evaluating test qualities and test usefulness (Bachman and Palmer 1996; Weir 2005). The paper is

organised around guiding principles of test usefulness, which will be used to structure the discussion and explore the promises and limitations of CB speaking tests in comparison with face-to-face speaking tests. Finally, the crucial question of whether the advantages of CB oral assessment outweigh any disadvantages will be addressed. Throughout the paper, an argument will be made about the importance of contextualising the debate on CB and face-to-face speaking assessment within the concept of 'fitness for purpose'. We will argue that CB assessment of speaking, despite its limitations, can be a viable alternative and complement to the face-to-face test mode, provided test developers can support a 'fitness for purpose' validity argument.

Let's start with some key terminological distinctions. The two most widespread test modes (also referred to as 'formats' or 'channels') in the assessment of speaking are '*direct*' speaking tests, which involve interaction with a human examiner, and '*semi-direct'* tests, in which test-taker speech is elicited with pre-recorded examiner questions. In direct tests of speaking the test taker is required to interact with another person, who is either an examiner or another test taker, or both, typically in a face-to-face setting (in the interest of simplicity, 'face-to-face' tests will be used interchangeably with 'direct' tests, even though technically direct tests could also be conducted over the phone or using VOIP). In semi-direct tests the test taker is required to respond to a series of prompts delivered by either audio/video tape or, more commonly, through a computer either online or CD-ROM- based. For the purposes of this paper, Web-based and CD-ROM speaking tests are conflated under one category, since the aim of the paper is not to discuss the differences between them (see Roever 2001 for a useful comparison). In addition to test delivery, a further distinction needs to be made between the use of human raters in the scoring process and computer-scored speaking assessment, also known as automated scoring assessment. Figure 1 presents a visual picture of the different possibilities available.

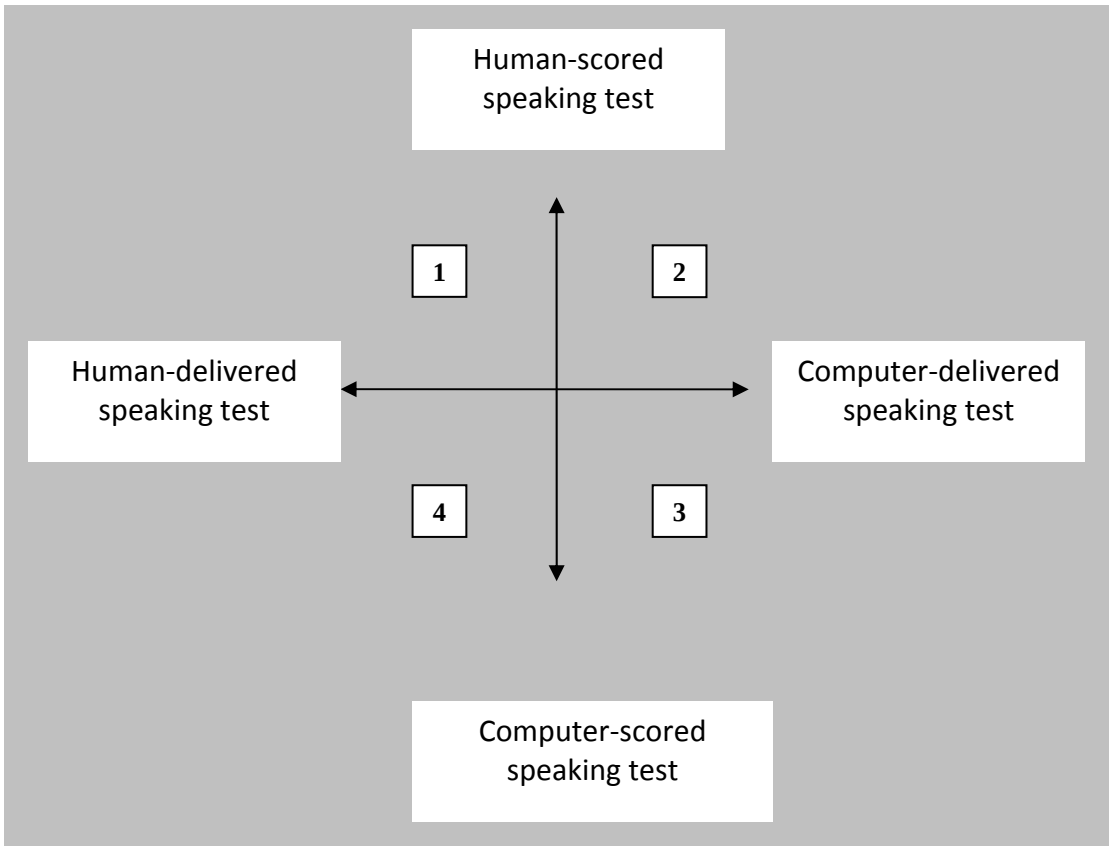


Figure 1 - Delivery and scoring possibilities in speaking assessment

As can be seen, different configurations are available in speaking assessment regarding both delivery and scoring. Quadrant 1 represents the traditional category of speaking tests which involve trained and certified human examiners acting as interviewers and raters. Some examples are Cambridge ESOL's suite of General speaking tests or the ACTFL OPI tests. The second quadrant represents speaking tests which are delivered by computer, but scored by a human rater. Examples include the forthcoming BULATS online speaking test, the TOEFL iBT speaking test and the SOPI and COPI produced by the Centre for Applied Linguistics. In quadrant 3 we find speaking tests which are fully automated and both delivered and scored by computer, as exemplified by Pearson's PTE Academic speaking test, the Versant tests, and the TOEFL iBT speaking practice tests. Finally, quadrant 4 represents tests which are delivered by a human interlocutor, but scored by machine. No such tests currently exist, simply due to the limitations of automated scoring systems to deal with the complexities of human-human interaction.

Test qualities

The choice of test mode for test delivery or scoring – CB or face-to-face – has fundamental implications for all aspects of a test's validity, including the construct, the task characteristics and cognitive demands triggered by the tasks, as well as the social consequences of the test. Several useful

frameworks for evaluating test qualities have been proposed in the literature, notably Bachman and Palmer's (1996) framework for test usefulness and Weir's (2005) socio-cognitive framework for validating language tests. While these two frameworks differ in their approach and presentation of the issues which need consideration, fundamentally they draw attention to similar issues which are at play when evaluating the usefulness of a test. In both frameworks, the main overarching issue which needs consideration is the **construct validity** of the test and the relevant **construct definition**. This overarching and unifying concept involves a definition of the traits of language ability to be assessed. The definition of the test construct determines all other decisions about the qualities of the test, which include:

- the **cognitive processes** required to complete the tasks in the test ('cognitive validity' in Weir 2005; 'interactiveness' in Bachman & Palmer 1996)
- the characteristics of the **test tasks** ('context validity' in Weir 2005; 'authenticity' in Bachman & Palmer 1996)
- the **reliability** of the test ('scoring validity' in Weir 2005; 'reliability' in Bachman & Palmer 1996)
- the test **impact** ('consequential validity' in Weir 2005; 'impact' in Bachman & Palmer 1996)
- the **practicality** dimensions of the test ('practicality' in Bachman & Palmer, 1996)

A discussion of the differences between these two frameworks is outside the scope of this paper (for a relevant discussion of their features, see Weir and O'Sullivan, forthcoming). Instead, the conceptual overlap in the two frameworks, as seen in the listed issues above, will be used as signposts to structure the paper.

Let us now review these test qualities in turn in the context of CB and face-to-face speaking tests, and focus on the impact of test mode on each one.

Construct validity

Construct validity is an overriding concern in testing and refers to the underlying trait which a test claims to assess – in this case, Communicative Language Ability. Construct definition is the specific definition of the trait which is (claimed to be) measured by the test. There isn't a single definition of a construct; instead, each test developer defines and delimits the specific trait, i.e. construct, which will be measured in an exam. A fundamental issue since the advent of computer-based language testing has been whether and how the delivery medium changes the nature of the construct being measured (Chapelle & Douglas 2006; Xi 2010).

The main characteristic of the direct face-to-face test mode is its interactional, multi-directional and co-constructed nature. The vast majority of speaking is reciprocal and jointly constructed by interlocutors who share a responsibility for making communication work and who accommodate their contributions to the evolving interaction (Lazaraton 2002; Sachs, Schegloff and Jefferson 1978). The underlying construct here is related to spoken *interaction*, which is integral to most frameworks of communicative language ability (Bachman & Palmer 1996; Canale & Swain 1980; CEFR Council of Europe 2001) and direct tests of speaking. Hymes (1972: 283), the father of the social view of language, has argued about the importance of considering the social aspect of speaking ability, noting that "the performance of a person is not identical with a behavioural record. It takes into account the interaction between competence (knowledge, ability to use), competence of others, and the … emergent properties of events themselves". In a similar vein, Savignon (1983: 8) has noted that communication is "dynamic, rather than … static" and involves negotiation of meaning between two or more persons within a specific context. Such a socio-cognitive approach to language use which views speaking both as a cognitive trait and a social one is consistent with the principles of the Common European Framework of Reference (2001: 26), which views speaking as comprising two skills, namely spoken production and spoken interaction.

Direct tests of speaking, such as the majority of Cambridge ESOL's speaking tests, are based on a socio- cognitive theoretical model with an emphasis not just on the knowledge and processing dimension, but also on the social, interactional nature of speaking (Taylor 2003, forthcoming). CB oral assessment is, in contrast, uni-directional and lacks the element of co-construction, since the test takers are responding to a machine. (It is worth noting that this is currently the case, but mobile and VOIP technology will allow future CB tests to have an interactional element.) In a semi-direct speaking test the construct is defined in psycho-linguistic terms with an emphasis on its cognitive dimension and on production. The construct underlying a computer-based test of speaking is, as such, not as broad as the construct underlying a face-to-face test, since by necessity it lacks an interactionist nature, which results in narrower construct definitions of this assessment mode, and has, indeed, been a point of critique (see, for example, Chun 2006). A key question, therefore, is whether the use of CB oral tests leads to construct under-representation, and what the consequences are.

Direct tests of speaking provide opportunities for all aspects of language to be evaluated which can be evaluated in a computer-based test, namely lexico-grammatical control, range and complexity, as well as fluency, coherence, cohesion and pronunciation. This core linguistic knowledge – also referred to as "facility in L2" (Bernstein, van Moere and Cheng, 2010: 356) comprises essential units of knowledge which every speaker of a language needs to have (some) mastery of. Additionally, and

crucially, face-to-face tests can do so both in monologic and dialogic (singleton, paired or group) contexts. Both test modes draw on essential qualities, but one – the computer-delivered scenario – is relatively narrow since it doesn't assess the ability to deal with the interactional and functional demands of speech. These involve the ability not just to construct one's own message, and use a range of informational functions (e.g., providing personal information, describing or elaborating), but also interactional functions (e.g., persuading, agreeing/ disagreeing, hypothesizing) and interaction management functions (e.g., initiating an interaction, changing the topic or terminating the interaction), as discussed in O'Sullivan, Weir and Saville (2002). In terms of interaction management functions, our understanding has expanded and we now know that they comprise a wide range of skills, such as the ability to develop topics through initiating, maintaining and closing turns; show comprehension through backchannels and through substantive comments and responses; produce turns which are relevant to previous turns; show listener support through backchannelling and appropriate gaze and body language; show anticipation of end-of-turn as evidenced by appropriate latching and overlapping; (Ducasse & Brown 2009; Galaczi 2008; Lazaraton 2002; May 2009; Riggenbach 1998; Storch 2002). It is all of these interactional and functional features which enable face-to-face speaking tests to represent a broader construct. In addition, the face-to-face test mode allows the examiner to probe the ceiling of a test taker's ability, thus allowing for a broader performance range to be elicited, which in turn samples from a broader construct.

In terms of construct definition, therefore, the face-to-face mode offers the advantages of a broader construct. CB tests narrow down the construct since currently they lack an interactional component (but could, as noted earlier, offer great advantages in the future by enabling CB speaking tests to have an interactional component through remotely connecting test takers and examiners). A further narrowing down of the construct is seen when tests are not just computer-delivered, but also computer-scored. Natural language processing has provided valuable tools for automated assessment of learner performances, but has also narrowed down the oral test construct, simply because it cannot at present capture the complexities of natural language use (Douglas & Hegelheimer 2007). The construct definition of an automated scoring speaking test, therefore, would be driven by the features that can be measured by a machine, and not the breadth of linguistic features underlying communicative language ability. As Xi (2010) notes, automated scoring systems need to address the issue of domain representation and the threat of under- or misrepresentation of the construct. Research has alerted us, for example, about the limitations of automated analysis of lexical features. Schmitt (2009, cited in Galaczi & ffrench, forthcoming) and Iwashita, Brown, McNamara and O'Hagan (2008), for example, have found that the only reliable predictor of lexical progression across proficiency levels is the number of tokens (i.e. words) and types (i.e. different words) used. No other lexical measures have so

far showed significant differences across levels, leading Schmitt to caution us about the limitations of automated assessment of lexical resources until more work is carried out on the use of multi-word lists (e.g., Martinez 2009).

Comparisons between the face-to-face and CB mode have received some attention in the academic literature, with various studies probing the differences and similarities between these two channels of communication and indirectly addressing the constructs underlying these test modes. Some studies have indicated considerable overlap between direct and semi-direct tests, at least in the statistical correlational sense that people who score high in one mode also score high in the other. For example, Stansfield and Kenyon (1992) compared the OPI (a direct test of speaking) and the SOPI (a tape-mediated test of speaking) and concluded that "both tests are highly comparable as measures of the same construct – oral language proficiency" (1992:363). Wigglesworth and O'Loughlin (1993) also conducted a direct/semi-direct test comparability study and found that the candidate ability measure strongly correlated, although 12% of candidates received different overall classifications for the two tests, indicating some influence of test method. More recently, Bernstein et al (2010) have investigated the concurrent validity of automated scored speaking tests and have also reported high correlations between human administered/human scored tests and automated scoring tests. The argument in these concurrent validity studies is that semi-direct and, especially, automated scoring speaking tests 'intend to *predict* speaking proficiency, but do not *directly measure* communicative competence' (Xi 2010: 294). In contrast, others have argued that as discourse events and assessment experiences, the direct and semi-direct modes "are not interchangeable as tests of oral proficiency" (O'Loughlin 2001:169). Similarly, Shohamy (1994) observed discourse-level differences between the two channels, and found that when the examinees talked to a tape recorder, their language was a little more literate and less oral-like; many of them felt more anxious about the test because everything they said was recorded and they only had a one-way for communicating was speaking – no requests for clarification and repetition could be made.

To summarise, although great strides have been made with CB-delivered and computer-scored speaking tests, the caveat of construct under-representation still remains, since at present computer-delivered tests cannot elicit the interactional co-constructed aspect of communication, and automated scoring systems do not capture the complexities of natural language use. Computer-delivered tests, however, can very effectively tap into speech production.

Task types and cognitive processes

The choice of test mode has key implications for the task characteristics of a test. This in turn impacts on the cognitive processes which a test can activate and tap into and has implications for the interactional authenticity (Bachman & Palmer 1996) and cognitive validity (Weir 2005) of a test. Weir (2005) has convincingly argued for the symbiotic relationship between cognitive validity (i.e., cognitive processes activated during a test) and context validity (i.e. task characteristics), which are core considerations in test validation. Similarly, Bachman and Palmer (1996) discuss the interactiveness of a test, which refers to the interaction between cognitive processes and test content, as a crucial test quality which is determined by test/task characteristics.

The tasks in computer-based speaking tests are typically short, constrained, production tasks, where one speaker produces a turn as a response to a prompt. The turns typically vary in length from brief one-word responses to longer responses lasting approximately a minute (although technically there is no constraint on time: the turn can be as long as test developers or test takers want). Some examples of CB speaking response tasks include reading a passage, repetition of utterances, describing visuals, responding to questions. Clearly, these task types are limited by the available technology and the flexibility of spoken interaction is not captured. Various voices in the literature have cautioned us about some caveats associated with the task characteristics of CB speaking tests: Alderson and Bachman (2006:xi) noted that the use of technology in speaking assessment could introduce a "conservative element" to the assessment and tasks, since test items are limited to those that can be delivered and marked by a machine. Following along similar lines of thought, Luoma (2004: 50) indicated that computer-based speaking tasks run the risk of becoming "the speaking equivalent of multiple choice tasks". Care needs to be given, therefore, to the design of CB speaking tasks so they have some (pseudo) communicative features. One example is a task from the BULATS Online speaking test which involves responding to a set of questions. All the questions are thematically related, which gives the task some authenticity, unlike an equivalent task with thematically unrelated questions.

Despite its limitations, the use of technology in speaking assessment has supported the creation of innovative tasks which use multi-media in the assessment of integrated skills. As Jamieson (2005: 233) suggests, these are "tasks that better represent authentic language use," since they approximate the target language use domain more closely (Bachman & Palmer 1996) and enhance a task's authenticity, especially in the case of EAP tests. Integrated tasks, as seen for example in the TOEFL iBT speaking test, involve several language skills in the completion of the task. For example, a test taker may be asked to read a short passage, listen to a short lecture and then provide an oral summary

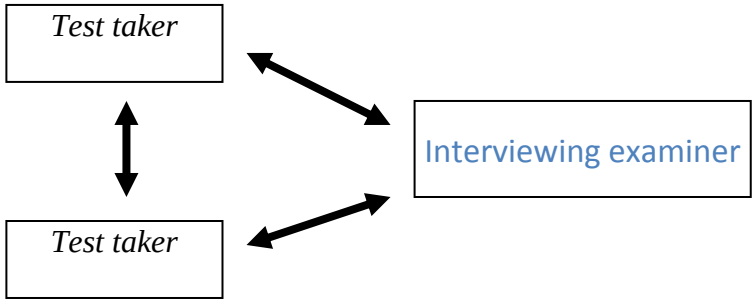
and compare and contrast the main points from the reading and listening. Integrated tasks raise the question of how the construct changes as a result of testing integrated skills. It's worth noting that there is still limited research on integrated tasks, and what research there is has revealed some interesting insights into the role of integrated tasks in language assessment. Research in a different context – reading and listening - has shown that even though integrated tasks required the test taker to adopt

non-test reading and listening strategies, the test takers approached them as test tasks and were interested in getting the right answer, instead of engaging in academic listening, for example (Cohen and Upton 2006; Douglas & Hegelheimer 2006).

CB speaking tasks could present a threat to a test's validity since a more limited range of task types is used; this is a result of the limitations the medium places on test developers to create tasks which can be delivered and, in some cases, scored by computer. Douglas and Hegelmeimer (2007: 125) note that computers are generally thought to "enhance the possibilities for presenting lifelike language tasks," and yet this contention has limited applicability to CB speaking assessment (with the exception of integrated tasks), since no existing CB programme can reflect truly interactive oral language. Interestingly, they note that chat bots (i.e. computer programmes designed to simulate an intelligent conversation with one or more human users via auditory or textual methods) are beginning to explore the potential. While obviously not examples of artificial intelligence, chat bots have an advantage in that most people prefer to engage with programs that are human-like, and this gives chatbot-style techniques a potentially useful role in interactive systems that need to elicit information from users, as long as that information is relatively straightforward and falls into predictable categories.

In contrast to CB speaking tasks, interactive speaking tests allow for a broader range of response formats where the interlocutors vary, and in addition to basic production tasks also include interactive tasks co-constructed by the examiner and test taker(s). For example, the interaction channels and, therefore, possible tasks types in Cambridge ESOL's Advanced speaking test include several interlocutors, as seen in Figure 1.

Figure 2 Interaction channels in Cambridge ESOL's Advanced speaking test



The tasks in this (paired) test format involve interaction between the examiner and each test taker, between the candidates themselves, between the candidates and examiner, and also include the chance for each candidate to carry out a monologic task on their own. Complex interaction is elicited in such a paired face-to-face format, as shown by Brooks (2009). The variety of tasks and response formats which a direct test contains allows for a wider range of language to be elicited and so provides broader evidence of the underlying abilities tested and consequently contributes to the exam's validity.

A caveat which needs to be considered in relation to face-to-face speaking tests, and which comes at the expense of eliciting a richer and more complex interaction, is the role of interlocutor variability, which could introduce construct-irrelevant variance and present a potential threat to the validity of the test. Such variability has been well-documented in the literature (e.g. Brown 2003; Galaczi 2008; Nakatsuhara 2009; O'Sullivan 2002), highlighting the need for test developers to address such inherent interlocutor variability and control for it in their tests. Swain (cited in Fox 2004: 240) wisely argues that variability related to different characteristics of conversational partners is "all that happens in the real world. And so they are things we should be interested in testing" and further contends that eliminating all variability in speaking assessment is "washing out … variability which is what human nature and language is all about". Coping successfully with such real-life interaction demands, therefore, becomes part of the construct of interactional competence and brings about an ethical responsibility on the part of test developers to construct tests which would be fair and would not provide (intentionally or unintentionally) differential and unequal treatment of candidates based on interlocutor variability. In the face-to-face speaking test context, the use of multi-part tests which include different types of talk (e.g., test-taker/examiner, test-taker/test-taker, and test-taker only), as is the practice at Cambridge ESOL, optimises the benefits of the paired format, while controlling for possible limitations. The use of a script for examiners also controls for some of the variability.

To summarise, CB speaking tasks are very good at eliciting production on constrained tasks and at the automated scoring of core lexico-grammatical linguistic features. Face-to-face tests, on the other hand, provide opportunities for interaction which could involve several interlocutors. Task characteristics have a strong, 'symbiotic' (Weir 2005) relationship with construct validity, in that they determine/delineate the cognitive processes activated by the tasks, and ultimately the construct which the test is measuring. As such, the differences between CB and face-to-face speaking tests entail not just different task types, but also different cognitive demands (Chun 2006; Field forthcoming). A fundamental question which arises is how representative these task features are of the real-life domain the test is trying to replicate. The key issue, therefore, extends beyond task characteristics to whether the test in question, be it CB or face-to-face, is appropriate for the inferences and decisions the test

users wish to make. As Weir (2005: 72) contends, "clearly, if we want to test spoken interaction, a valid test must include reciprocity conditions." In contrast, if we are mostly interested in the narrower constructs of speech production or "facility in L2" (Bernstein et al. 2010: 256), then a computer-based automated test would be appropriate. It is a consideration of the purpose of the test which is crucial, and we shall return to that later in the paper.

Reliability

Test reliability ('scoring validity' in Weir 2005) relates to the dependability of test scores, and is another test quality which needs to be considered in investigating the characteristics of CB and face-to-face speaking tests. The two main forms of variation, and therefore, threats to reliability in speaking assessment are (a) the role of interviewer variability in delivering the test, and (b) the influence of rater variability in scoring the test. Both of these types of variability and sources of measurement error are reduced in CB oral assessment, and eliminated in automated-scoring tests. One of the undoubted advantages of computer-delivered speaking tests is their high reliability due to the standardisation of test prompts and delivery, which naturally eliminates any interviewer variability. Each prompt is delivered in an identical way, regardless of where the candidate takes the test, which in turn eliminates measurement error due to interviewer variability. Computer-scored oral tests also enjoy very high reliability, since the same algorithms and standards are always applied by the computer (which, however, can only deal with a narrower sample of speech, as discussed earlier).

In addition, computer-delivered speaking tests (or performance tests in general) have logistical advantages which enhance reliability, since they allow the more efficient deployment of different marking models (e.g., full double marking, partial double marking), and more rater monitoring checks to be employed as part of the assessment process. For example, it would be relatively simple to give raters worldwide access to test-taker performances in the form of electronic files as part of a multiple-marking model, thus increasing the reliability (and also efficiency) of the test.

One of the challenges and frequent points of criticism of human-marked tests (face-to-face or computer-based) is the role of rater effects such as rater harshness/leniency, consistency, central tendency (Myford & Wolfe 2003, 2004), which present an inherent threat to reliability and could compromise the integrity of the test scores. Despite such caveats, acceptably high reliability can be achieved in face-to-face tests, and is often in the .7-.8 Pearson correlation range for well trained raters (see Galaczi 2005 for Cambridge ESOL findings). Acceptable levels of inter-rater correlations, however, require extensive and rigorous rater training and monitoring. For example, the Cambridge ESOL face-to-face speaking tests rely on several thousand examiners who undergo rater induction and

training, followed by annual cycles of standardisation and monitoring in order to control for rater effects such as severity and consistency (see de Velle 2009; Taylor & Galaczi, forthcoming for a description and discussion of Cambridge ESOL practice in terms of rater training and standardisation). This is clearly a complex and large-scale process which needs a well-functioning system for ensuring quality of rater ratings.

In addition to rater effects, interviewer variability is another source of measurement error and threat to the reliability (and validity) of face-to-face speaking tests (Brown 2003). One possible way to control for interviewer reliability and standardise test interactions is with the use of a script which the examiners must follow when conducting the speaking test, as is the practice at Cambridge ESOL (see, for example, the First Certificate in English Handbook for Teachers 2007). Obviously, the challenge is for such scripts to both constrain and allow freedom at the same time, so a balance has to be reached between standardising the interaction and allowing flexibility, an issue which is especially pertinent in oral tests covering a large part of the proficiency spectrum. Such a balance is achieved, for example, with a choice of follow-up questions which interviewers can choose from.

To summarise, CB oral tests have a strong advantage over their face-to-face counterparts in terms of reliability due to the uniformity and consistency of computer-delivered tests and computer-scored performances. Human-scored tests can also have acceptably high reliability coefficients, but an investment needs to be made into a cyclical, rigorous system of examiner training, standardisation and monitoring.

Test impact

Test impact and washback relates to the effect of the test on teaching and learning, and also on society in general. It is generally conceptualised as a continuum, ranging from positive to negative.

The impact and washback of face-to-face speaking assessment is generally regarded as positive, due to the close connection between face-to-face tests, and especially paired and group ones, and communicative classrooms. In the case of CB assessment of speaking, impact can vary depending on the context of use. Such assessment can have positive impact if the alternative is lack of development/assessment of speaking skills. Clearly, any speaking test, even one which lacks interactional features, holds benefits for learning and teaching. The impact of CB tests of speaking can be negative, however, if the test format restricts the focus on interaction in a learning context and prescribes an excessive preoccupation with monologic speech and the limited skills required by some

computer-based tests, e.g. reading texts aloud vs. responding to questions in a simulated role play type situation.

Test-taker perceptions of CB tests have received some attention in the literature as well, as seen in a group of studies on earlier generations of CB speaking tests, where test takers have reported a sense of lack of control and nervousness (Clark 1988; Stansfield 1990). Such test-taker concerns have been addressed to an extent with some newer-generation CB oral tests, which are computer adaptive in nature, such as the COPI test produced by the Centre for Applied Linguistics (Kenyon & Malabonga 2001), and also through the provision of tutorials and practice materials. The COPI includes some positive features which enhance a test's impact, such as the possibility for test takers to select topics, their control over the planning/response time, and the adaptive nature of the test. In fact, Kenyon and Malabonga (2001) reported that on average test takers preferred the COPI to the SOPI because it seemed less difficult, featured a fairer set of questions and situations, made them feel less nervous, had clearer directions, and enabled a more accurate depiction of their strengths, weaknesses, and current abilities to speak in the target language. Such test-taker perceptions are an important example of positive impact. Interestingly, the face-to-face OPI test was perceived by the study participants to be a better measure of real-life speaking skills. In a more recent article, Qian (2009) has reported that although a large proportion of his study participants had no particular preference in terms of direct or semi-direct tests, the number of participants who strongly favoured direct testing far exceeded the number strongly favouring semi-direct testing.

Another positive impact feature of CB oral assessment relates to the potential of automated-scoring speaking tests to provide more detailed feedback information to test takers. The use of technology to provide efficient and instantaneous feedback to test takers holds great promise due to its efficiency and (relative) transparency, as seen, for example with DIALANG in the context of writing, reading and listening, where learners are given feedback about their strengths and weaknesses.

Practicality

Weir (2005: 49) does not include practicality in his framework. He justifies this "heretical view" with the argument that in some cases practicality could affect the quality of the test if it intrudes before sufficient validity evidence is available supporting the interpretation of the test scores. We should not, as Weir (2005: 49) argues, "consider method before trait". Bachman and Palmer (1996) adopt a different view and include practicality in their framework of test usefulness, suggesting that considerations of practicality should follow considerations of the other test qualities. The authors argue that practicality is just as important as the other test qualities and should play a role at every

stage of the cyclical and iterative test development cycle. It is the latter view which is adopted here, and practicality will be discussed in its relation to computer-based and face-to-face speaking tests.

One of the undoubted strengths of computer-based speaking tests is their high practicality, which is manifested in several respects. After the initial resource-intensive set-up, CB speaking tests are cost-effective, since they take away the need for trained raters to be on site and allow for large numbers of test takers to be tested at the same time. They also offer great flexibility in terms of space and time, since online speaking tests can be offered at any time, unlike their face-to-face counterparts, which are constrained by logistical considerations. In contrast, human-delivered and human-scored speaking tests pose a heavier administrative burden, mainly in the development and maintenance of a network of trained and reliable oral examiners, who act as interviewers and raters and need regular training and standardisation, as well as extensive scheduling during exam sessions. It must be noted, however, that the use of technology in examiner training and standardisation does enhance the efficiency of this process, as seen in online examiner training.

A further practical advantage of technology in speaking assessment is that computers can facilitate human interaction among people in the same room, as well as continents apart. For language testers this is an exciting development since technological advances can provide opportunities to connect test takers and examiners remotely and address the 'right-here-right-now' need of face-to-face speaking tests. Direct human-to-human communication (e.g. aided by mobile or VOIP technology) holds great potential for the future in terms of combining the strengths of face-to-face and technology. It is such symbiosis between technology and more traditional approaches to speaking which holds great future potential in terms of truly combining the strengths of the different modes.

Practicality also relates to the logistics of ratings. One of the practical advantages of CB speaking tests is that files can be easily distributed to/accessed by certified raters around the world, making the scoring process much more efficient. In addition, raters are potentially able to listen to candidate performances at any time and anywhere, in any order and combinations. All of these practical strengths of CB speaking tests would potentially increase test reliability and decrease rater effects, since more raters could be used to provide marks for (parts of) the test.

The test security of CB speaking tests is another practicality issue which needs consideration. Roever (2001) discusses the problems of the vulnerability of the internet to computer hackers, as well problems with cheating and item exposure, and as Bernstein et al. (2010: 274) elegantly put it, automated speaking tests "are susceptible to off-construct coaching". In the context of automated-

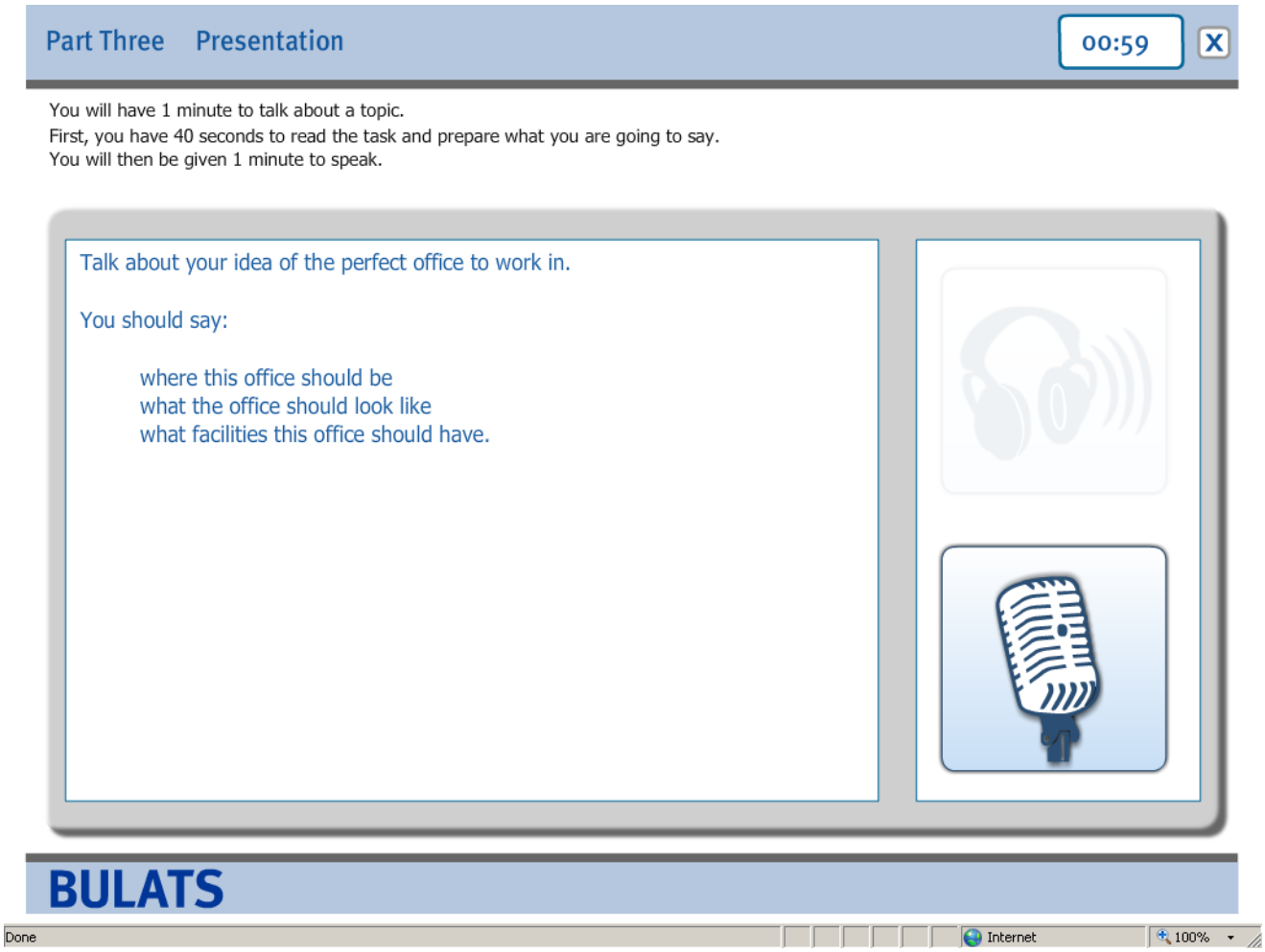
scoring, test takers could try to beat the system by providing language that the algorithm is going to score highly (Bridgeman 2010). Such issues are especially important in high-stakes, large-scales testing, where the security of the test forms part of the validity argument supporting the test.

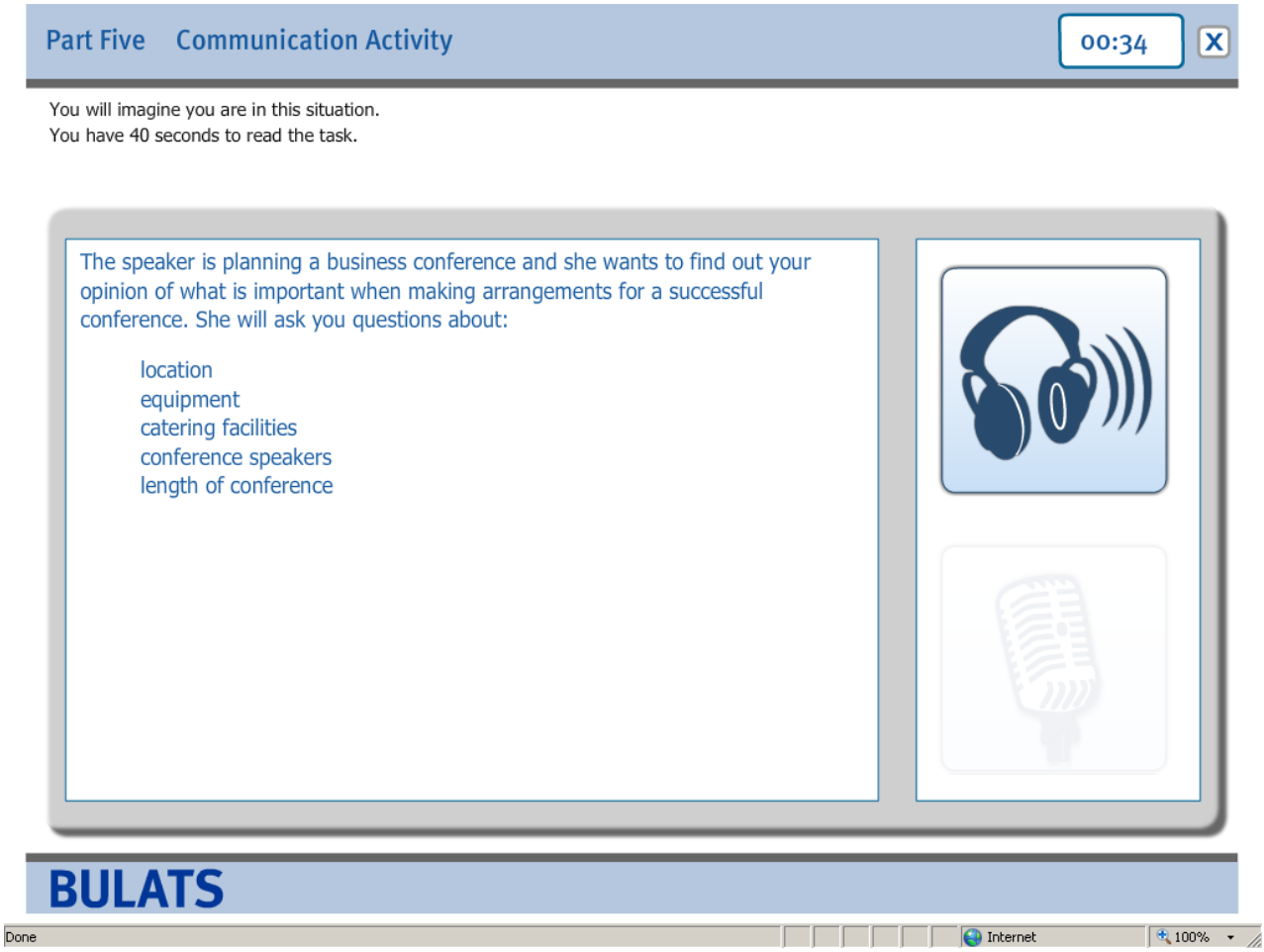
So far we have discussed the implications of test mode on a range of test qualities. The paper now moves to a brief description of the BULATS online speaking test, as an illustration of a computer-delivered test.

The Cambridge ESOL BULATS Online speaking test

The BULATS Online Speaking test is a multi-lingual test for in-company use developed by Cambridge ESOL in response to the need for a reliable, efficient, inexpensive, flexible, and easy to administer method of assessing certain speaking skills. The test (when released) will be taken online in real time at any time convenient to the test taker, and will be accessed online and marked remotely by human raters.

The test comprises five tasks. Part 1 involves an Interview in which test takers respond to 8 questions about him/herself and his/her job; Part 2 involves the repetition of text that might be typically read aloud in a work/business situation; Part 3 comprises a presentation on a work-related topic; Part 4 involves a presentation on a business related topic including comment on a visual such as pie charts or bar charts; Part 5 asks the test takers to imagine that they are in a specific situation with another person and have to respond to questions they may be asked in that situation. For the sake of illustration, two examples are given below from Parts 3 and 5 respectively.


Part Three Presentation
00:59
X
You will have 1 minute to talk about a topic.
First, you have 40 seconds to read the task and prepare what you are going to say.
You will then be given 1 minute to speak.
Talk about your idea of the perfect office to work in.
You should say:
where this office should be
what the office should look like
what facilities this office should have.
BULATS
Done
Internet
100%



A small-scale trial was set up in four countries (UK, Switzerland, Argentina and India) to investigate the technical issues associated with the test, and also to gather test-taker perceptions about the test (Ingham, Chambers & Shortland 2008). The trial findings indicated that candidates were fairly mixed in their response to taking a Speaking Test without an examiner; some preferred a test without an Examiner as they said the presence of another person made them nervous but some liked the support of an Examiner. The timer on the screen was found to be helpful by some but distracting by others. Some examples from the feedback:

- *It's better because you don't know which teacher (ie Examiner) you will get.*
- *I can concentrate more than with an Examiner. If I talk with an Examiner, I'll feel nervous.*
- *When I communicate with an examiner sometimes I feel stresses.*
- *It's a little bit surprising but the software is well done and you feel comfortable quickly.*

But also:

- *It was a bit difficult to manage time, so in a few cases I prepared a shorter answer than required.*
- *It's not a bad idea but the problem is when you don't understand, you can't ask the computer 'Excuse me'.*

- *In my opinion, the timer is not ideal. Because you try to limit your answer and this has a negative influence about the answer itself.*
- *If it was face-to-face, it would be warm and comfortable – less stressful. It's nice to see someone giving reactions to your answers.*

In summary, most candidates liked doing the BULATS test on computer, felt comfortable doing the test without an examiner and agreed the test gave them full opportunity to demonstrate the language skills they needed for their job. The trial also focused on the examiners' experience and found that all examiners commented positively on their experience of assessing on computer. They thought the assessment was more efficient, more reliable, and more convenient and flexible for the examiner than face-to-face assessment, and felt that the mode of delivery had no impact on their judgement

Summary and discussion: The innovation and caveats of computer-based speaking assessment

Finally, as a summary let's consider two simple but key questions: What do machines do better? What do humans do better? As the overview and discussion has indicated so far, the balance and tension between test qualities leads to complex interactions and outcomes which shape the validity and usefulness of a test. The main advantages of computer-delivered and computer-scored speaking tests is their convenience and standardisation of delivery, which enhances their reliability and practicality (Chapelle & Douglas 2006; Douglas and Hegelheimer 2007; Jamieson 2005; Xi 2010). The tradeoffs, on the other hand, relate to the inevitable narrowing of the test construct, since CB speaking tests are limited by the available technology and include constrained tasks which lack an interactional component. In CB speaking tests the construct of communicative language ability is not reflected in its breadth and depth, which creates potential problems for the construct validity of the test.

In contrast, face-to-face speaking tests and the involvement of human interviewers and raters introduces a broader test construct, since interaction becomes an integral part of the test, and so learners' interactional competence can be tapped into. The broader construct, in turn, enhances the validity and authenticity of the test. The caveat with face-to-face speaking tests is the role of interlocutor variability, the potential introduction of construct-irrelevant variance in the testing process, and the role of rater effects, which could impact the reliability of the test. In order to address such issues, careful attention needs to be given to the test design and the use of multiple response formats, which would allow for the benefits of a face-to-face test to be optimised, while controlling for the possible role of variables external to the construct. In addition, resources need to be invested in a rigorous and on-going system of examiner recruitment, training, standardisation and monitoring, which would support the reliability of face-to-face speaking assessment.

It is worth reminding ourselves that there isn't just one way of testing speaking, or one 'best' way to do it. As language testers we can choose from a range of useful formats which aid us in eliciting and assessing speaking skills, from fully automated speaking tests to ones involving human interviewers and raters. All of these formats bring along their strengths and weaknesses. A crucial concept in language assessment is 'fitness for purpose'. Tests are not just valid, they are valid *for* a specific purpose, and as such different test formats have varied applicability for different contexts, age groups, proficiency levels, and score-user requirements. As Wainer et al. (2000: xxi) emphasised a decade ago, "The questions we now must address deal less with 'how to use it?' but more often 'under what circumstances and for what purposes should we use it?'". A decade later, that question still informs current debates on fitness for purpose (e.g. Xi 2010). A CB speaking test would be suitable, for example, for the purpose of providing 'snapshots' of language ability which would be valuable for institutional screening purposes, and where a narrower construct definition is justified, as for example, in a broad survey of linguistic abilities. In contrast, a face-to-face test of speaking would be more suitable in cases where evidence of breath and depth of language is needed.

It is important to view computer-based and face-to-face speaking assessment as *complementary* and not competing perspectives, where technology is seen not as a replacement for traditional methods, but as a new additional possibility. Instead of viewing these different modes as one or another, they could be viewed as one and the other, and used in a way that optimises their strengths and minimises their respective limitations. In a related discussion, Bernstein et al. (2010) make a similar appeal, arguing that automated speaking tests alone should not be used as the sole basis of decision-making, but should be one piece of evidence about a test-takers' ability. The full range of interactive speaking performances that language educators are interested in is currently not adequately elicited in computerized formats; similarly, the complexities of such performances cannot be captured by automated scoring and speech recognition technology. However, the purpose of a test may warrant a CB format, with all the limitations and advantages that brings. As such, we should perhaps use computers for the delivery and (possibly) scoring of short-response constrained production tasks, where the primary interest is assessment of core language elements, and which computers can capture effectively. Placement/screening tests and low-stakes tests, as well as large-scale surveys of linguistic abilities could benefit greatly from such a computer-based approach to speaking assessment. Face-to-face interactive tests could be another 'tier' of speaking assessment, which would elicit complex and richer speech (but at the expense of practicality) and would be more applicable in high-stakes and academic contexts. Such a division of labour would optimise the benefits of both test modes.

Research by Weigle (2010) has again reminded us that computers and humans bring different perspectives to the performance assessment process. In the context of writing assessment, the author reports that human raters and automated rating systems focus on different features, indicating that their evaluations are based on somewhat different constructs. And yet there is enough overlap between the two modes, stemming from reference to the same overarching construct, to provide valuable, albeit differing perspectives. Likewise with speaking tests, computer-delivered/scored tests and human-delivered and scored tests bring unique strengths to the assessment process. The former present a consistent but relatively narrow conceptualisation of the construct of speaking, the latter bring a broader perspective, but with some human inconsistencies. It is in the combination of the two approaches in terms of delivery and scoring that we will see a symbiotic relationship between technology and humans and enhance the assessment of speaking ability, leading to a fairer assessment overall.

References

Alderson, C., & Bachman, L. (2007). Series editors' preface. In C. Alderson & L. Bachman (Eds.), *Assessing language through computer technology* (pp. ix-xi). Cambridge: Cambridge University Pre

Bachman, L and Palmer, A (1996) *Language testing in practice* Oxford: Oxford University Press.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355-377.

Brooks, L (2009) Interacting in pairs in a test of oral proficiency: Co-constructing a better performance *Language Testing 26*(3) 341-366.

Brown, A (2003) Interviewer variation and the co-construction of speaking proficiency *Language Testing 20*(1) 1-25.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Canale, M. (1983). On Some Dimensions of Language Proficiency. In J. W. Oller (Ed.) (pp. 333-342). Rowley, MA.: Newbury House.

Chapelle, C., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Chun, C. (2006). An Analysis of a Language Test for Employment: The Authenticity of the PhonePass Test *Language Assessment Quarterly, 3*(3), 295-306.

Clark, J. L. D. (1988). Validation of a tape-mediated ACTFL/ILR-scale based test of Chinese speaking proficiency. *Language Testing, 5*(2), 197-205.

Cohen, A. D., & Upton, A. T. (2006). *Strategies in responding to the new TOEFL reading tasks*. Princeton: New Jersey.

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

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. In M. McGroarty (Ed.), *Annual Review of Applied Linguistics* (Vol. 27, pp. 115-132). Cambridge: Cambridge University Press.

Ducasse, A M and Brown, A (2009) Assessing paired orals: raters' orientation to interaction *Language Testing 26*(3) 423-443.

Field, J. (forthcoming). Cognitive validity. In L. Taylor (Ed.), *Examining speaking* (Vol. 30). Cambridge: Cambridge University Press.

Fox, J (2004) Biasing for the best in language testing and learning: An interview with Merrill Swain *Language Assessment Quarterly 1*(4) 235-251.

Galaczi, E. D. (2005). Upper Main Suite speaking assessment: towards an understanding of assessment criteria and oral examiner behaviour. *Cambridge ESOL Resarch Notes, 20*(5), 16-19.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: the case of the First Certificate in English examination. *Language Assessment Quarterly, 5*(2), 89-119.

Galaczi, E. D., & Ffrench, A. (forthcoming). Context validity of Cambridge ESOL speaking tests. In L. Taylor (Ed.), *Examining speaking* (Vol. 30). Cambridge: Cambridge University Press.

Hymes, D. (1972). *Directions in Sociolinguistics: The Ethnography of Communication.* New York: Holt, Rinehart and Winston.

Ingham, K., Chambers, L., & Shortland, M. (2008). BULATS Online Speaking Proof of Concept. Unpublished Cambridge ESOL Internal report.

Iwashita, N, Brown, A, McNamara, T and O'Hagan, S (2008) Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics 29*(1) 24-49.

Jamieson, J. (2005). Trends in computer-based second language assessment. In M. McGroarty (Ed.), *Annual review of Applied LInguistics* (Vol. 25, pp. 228-242). Cambridge: Cambridge University Press.

Kenyon, D and Malabonga, V (2001) Comparing Examinee Attitudes Toward Computer-Assisted and Other Proficiency Assessments *Language Learning and Technology 5*(2) 60-83.

Lazaraton, A. (2002). *A Qualitative Approach to the Validation of Oral Language Tests*. Cambridge: Cambridge University Press.

Luoma, S (2004) *Assessing speaking* Cambridge: Cambridge University Press.

Martinez, R (2009) *Towards the inclusion of multiword items in vocabulary assessment* Paper presented at the Language Testing Forum.

May, L. (2009). Co-constructed interaction in a paired speaking test: the rater's perspective. *Language Testing, 26*(3), 397-421.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using Multi-Facet Rasch measurement: Part 1. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using Multi-Facet Rasch measurement: Part 2. *Journal of Applied Measurement, 5*(2), 189-227.

Nakatsuhara, F. (2009). *Conversational styles in group oral tests: How is the conversation constructed*? Unpublished PhD thesis, University of Essex.

O'Loughlin, K (2001) *The equivalence of direct and semi-direct speaking tests* (Vol 13) Cambridge: Cambridge University Press.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency pair-task performance. *Language Testing, 19*(3), 277-295.

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

Qian, D (2009) Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers *Language Assessment Quarterly 6*(2) 113-125.

Riggenbach, H. (1998). Evaluating learner interactional skills: Conversation at the micro level. In R. Young & A. He (Eds.), *Talking and Testing* (pp. 53-67). Amsterdam: Phil.: John Benjamins.

Roever, C. 2001. *Web based language testing.* Language Learning and Technology, Vol 5, No. 2, May 2001, 84 - 94.

Sacks, E. A., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

Savignon, S. (1983). Communicative competence: Theory and classroom practice. Reading, MA: Addison-Wesley

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*, 99-123.

Stansfield, C. (1990). An evaluation of simulated oral proficiency interviews as measures of oral proficiency. In J. E. Alatis (Ed.), *Georgetown University Roundtable of Languages and Linguistics 1990* (pp. 228-234). Washington, D.C.: Georgetown University Press.

Stansfield, C and Kenyon, D (1992) Research on the comparability of the Oral Proficiency Interview and the Simulated Oral Proficiency Interview *System 20*(3) 347-364.

Storch, N. (2002). Patterns of Interaction in ESL Pair Work. *Language Learning, 52*(1), 119-158.

Taylor, L. (2003). The Cambridge approach to speaking assessment. *Cambridge ESOL Research Notes, 13*, 2-4.

Taylor, L. (forthcoming). Introduction. In L. Taylor (Ed.), *Examining Speaking*. Cambridge: Cambridge University Press.

Taylor, L., & Galaczi, E. D. (forthcoming). The scoring validity of Cambridge ESOL speaking tests. In L. Taylor (Ed.), *Examining speaking* (Vol. 30). Cambridge: Cambridge University Press.

Taylor, L and Wigglesworth, G (2009) Are two heads better than one? Pair work in L2 assessment contexts *Language Testing 26*(3) 325-339.

University of Cambridge ESOL Examinations (2007) *First Certificate in English Handbook for Teachers*, Cambridge: UCLES.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., et al. (2000). *Computer adaptive testing: A primer* (2nd edition ed.). Mahwah, NJ: Erlbaum.

Weir, C (2005) *Language testing and validation: An evidence-based approach* Basingstoke: Palgrave Macmillan.

Weir, C & O'Sullivan, B. (forthcoming). In B. O'Sullivan (Ed.), Language Testing: Theories and Practices. Palgrave.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*(3), 291-300.

The author:
Evelina Galaczi
University of Cambridge Examinations
1 Hills Road
Cambridge
United Kingdom
E-mail: Galaczi.E@cambridgeesol.org

Evelina Galaczi is a Research and Validation officer at the "Research and Validation" group at the University of Cambridge ESOL Examinations where she works on the Cambridge ESOL speaking exams and Teaching Awards. She has a background in Applied Linguistics and Language Assessment and holds a Master's and Doctorate degrees from Teachers College, Columbia University (USA). Her research interests include the testing of speaking and writing ability and qualitative approaches to assessment research. She has extensive experience as an EFL/ESL teacher, teacher-trainer and program administrator.

VALIDITY ISSUES IN FACE-TO-FACE VERSUS SEMI-DIRECT TESTS OF SPEAKING

Ildikó Csépes
University of Debrecen, Hungary

Abstract

Threats to test validity have been identified in relation to both face-to-face and semi-direct tests. This presentation is aimed at outlining the most important variables that may influence performance outcomes and discussing some task design considerations that test developers assessing speaking at different levels of proficiency should take into account.

Introduction

Performance on language tests is likely to be affected by both the ability to be measured and the method used in the measurement (Bachman, 1991). Therefore, the measurement of oral proficiency is also affected by the method that is used to elicit a rateable sample from candidates. Oral proficiency seems to have been measured in basically two major modes: in a face-to-face/direct and a tape-mediated/simulated/semi-direct interview format (for short: OPI [Oral Proficiency Interview] and SOPI [Simulated Oral Proficiency Interview]). More recently, however, computer-mediated formats have also been offered as an adaptation of and thus an alternative to the tape-mediated interview format (e.g. the COPI, see Malabonga, Kenyon & Carpenter, 2005). All these formats have one main feature in common: the direction of the exchange is controlled by the examiner-interlocutor or test designer. In the traditional, face-to-face OPI, the conversation unfolds either in a structured or unstructured way between an examiner-interlocutor and a candidate, ensuring that there is some degree of interactiveness. However, in a tape- or computer-mediated interview the interaction is much more tightly controlled since the candidate has to respond only to specific, pre-recorded prompts, as a result of which very little genuine interactiveness, or none at all, characterizes the oral discourse.

In this paper, I would like to argue that some of the threats to test validity in relation to the face-to-face as well as the semi-direct speaking tests can be neutralized by introducing the paired mode in the assessment framework. In order to support this claim, first we will review briefly what the method effects of the OPI/SOPI are and how they may enhance or undermine the validity of oral proficiency

testing. Then we will examine the assumed threats and benefits of the paired exam in the context of face-to-face speaking tests, also discussing the potential of this test mode in computer-mediated assessment of oral abilities. The issue of scoring by human raters versus automated scoring of spontaneous speech is beyond the scope of this paper.

Direct and Semi-direct Individual Test Modes

In the 1980s, the validity of the OPI was beginning to be questioned in proportion to its popularity as a testing method. Shohamy's study (1983) brought to light that the oral interview test lacked stability as it was sensitive to a change in the speech style ("interview" style vs. "reporting" speech style) and a change of the topic of discussion. Hughes (1989, p. 104) pointed out a serious drawback of the interview that was related to the differences between the interviewer and the candidate in terms of their status and role in the interaction: "the candidate speaks as to a superior and is unwilling to take the initiative". Weir (1990, p. 76) criticized the interview format for not being able to "replicate all the features of real life communication such as motivation, purpose and role appropriacy". Similar limitations of the oral proficiency were emphasized by van Lier (1989), who compared OPIs and basic features of conversation, based on the assumption that conversation seems to be the best vehicle to display one's speaking ability in context. His comparison revealed significant differences between these two modes of social interaction. According to van Lier, the basic characteristics of conversation include: "face-to-face interaction, unplannedness, potentially equal distribution of rights and duties in talk, and manifestation of features of reactive and mutual contingency" (p. 495). Van Lier, however, argued that in OPIs emphasis is put on successful elicitation of language and not on successful conversation. As a result, the interaction of the interview is controlled by the interviewer to a large extent. Thus, van Lier questioned the validity of the OPI as an appropriate means to measure conversational ability, and his work has stimulated a number of empirical research studies into the nature of discourse produced in the interview (e.g. Young & Milanovic, 1992; Ross & Berwick, 1992; Young, 1995; Johnson & Tyler, 1998; Kormos, 1999; Ross, 1998). According to He and Young (1998), salient differences between OPIs and ordinary conversation include the topical and turn-taking systems, as well as the speech exchange system and the goal-orientedness.

In the light of the above mentioned limitations of using the OPI as a measurement device to assess candidates' conversational ability, it stands to reason that the simulated oral proficiency interview (SOPI) may be suffering from similar drawbacks in spite of a number of advantages in terms of test reliability, validity and practicality in comparison to the OPI. The latter include the following (Stansfield, 1989):

- the speech sample elicited is longer and therefore, more reliable, also producing greater content validity;
- the format of the questions is invariant, thus reducing the effect of variability of interlocutor behaviour;
- scoring can be carried out by the most reliable rater, ensuring scoring reliability;
- less costly to administer.

The Computerized Oral Proficiency Instrument (COPI) seems to have further strengths as test takers can control the choice of tasks, difficulty level of tasks and their planning and response time, which in turn enhance test takers' favourable perceptions of the test (Malabonga et al., 2005).

The early validation studies related to SOPI all focused on quantitative comparisons between rating outcomes from the SOPI and the OPI. The findings suggested that the SOPI was a valid and reliable surrogate measure in place of the OPI (Stansfield, 1989, 1991; Stansfield & Kenyon, 1992). However, the high correlations gained by these studies on the numerical validity of the SOPI may have been due to the fact that by using a similar rating scale in the two procedures, the final scores only reflected some aspects of the tests but not others. Shohamy (1994) argued that correlations provided necessary but insufficient evidence for test substitution, and therefore she set out to investigate differences between the OPI and the SOPI, using discourse analytic tools. Her findings show, for instance, that in the SOPI there is a sharp shift from one topic to another; a large and varied number of topics are elicited; oral language resembles a monologue; although varied language functions are expected to be used, there is no guarantee that they will in fact be produced; and there are very few paralinguistic and prosodic features. In order to account for the observed differences Shohamy suggested that discourse variation across the two tests could be attributed to the different language elicitation contexts: the lack of the physical presence of a human interlocutor in the SOPI may be responsible for the less conversational and intimate language production.

O'Loughlin (1995) also investigated the comparability of the tape-mediated and the face-to-face test modes in terms of test discourse. In addition to focusing on the degree of lexical density in the two test modes, he also investigated the effects of task type (description, narration, discussion and role-play). The results indicated that there was a difference indeed between the two oral test modes in terms of lexical density: all the tape-based language samples had higher means for lexical density and the lexical density values showed little variation across the four tasks in the tape-mediated mode. However, in the live version the tasks could not be characterized by a similarly low degree of variation. Although the first three tasks showed equally low variation in the live mode, O'Loughlin

found that the lexical density value was clearly much lower in the fourth task (the role play) of the face-to-face test, where the degree of interactiveness was the highest of all the tasks. O'Loughlin's explanation for this finding was that the interlocutors were expected to provide substantial contribution in the live role play and this fact had a strong effect on how the candidate's output was shaped. Therefore, he concluded that "the role of the interlocutor in any given task will strongly influence the degree of lexical density, i.e., the *higher* the level of interaction the *lower* the degree of lexical density in candidate output" (p. 234). This implies that if the interlocutor's contribution is kept at a minimum in a live oral test, lexical density may not be strikingly different from that of the tape-mediated test. Alternatively, if the interaction between interlocutor and candidate is enhanced in the face-to-face test, it is plausible to assume that even more varied language will be tapped by the live version as opposed to the tape-mediated mode.

A study by Luoma (1997), confirming the comparability of scores on a face-to-face test and a tape-mediated test, also investigated whether the features of performance assessors paid attention to were similar or different in the case of the two test types. She found several common features as well as different ones, the latter being primarily related to interactiveness and task structuring. In the tape-mediated test, assessors expected compliance with the instructions, the content of most replies was guided. Therefore, as Luoma points out, the content of the candidates' responses was more predictable than in the face-to-face test. As a consequence, it was not surprising to find that assessors seemed to value creativity in the tape-based version. In the face-to-face test, however, willingness to show ability was expected and valued accordingly. Clearly, the differences in the assessment processes can be attributed to the nature of the tasks and conditions of the tests rather than specific language features in the candidates' output.

The Paired Exam Mode

Having highlighted some empirical findings in relation to differences between direct and semi-direct oral proficiency tests, we need to look for an alternative test mode that can bypass the asymmetrical power relations involved in interlocutor–test taker interactions, but at the same time it can maintain a degree of genuine interactiveness. Such an alternative format seems to be the paired mode of oral proficiency assessment. To date, no such format has been used in semi-direct tests of speaking, but in direct tests this mode seems to have been used more increasingly. The paired mode was introduced by UCLES in their main suite exams in the 1990s, claiming claimed that the paired format had several advantages over the one-to-one interview format (Saville & Hargreaves, 1999). The interaction pattern in the paired format is more varied as more than one interaction pattern is possible: in addition to

providing scope for interaction between the examiner and one candidate, two candidates can get engaged in a conversation with each other and the examiner. When comparing transcripts of paired and one-to-one UCLES exams, Taylor (2000) found that in the CPE paired speaking test the examiner‘s contribution was reduced while the relative contribution of the candidates increased. In addition, the number of turns produced by the candidates increased and the length of their turn varied significantly across different tasks. Based on these findings, Taylor proposed that the paired exam was “capable of generating a richer and more varied sample of spoken language from each candidate” (p.15).

Interaction between candidates in the paired format is also believed to enhance beneficial washback effects on classroom teaching, as the exam is likely to encourage more interaction between learners. A similar concern is echoed by Swain (2001), who claimed that the paired mode of speaking abilities was likely to generate positive washback on teaching and to mirror good language teaching practice. It is important to note that this belief in the potential of the paired exam for positive washback has not been disconfirmed or discarded.

Similarly to other test methods or testing modes, there are potential problems and issues involved in examining candidates in the paired format. However, there is little research evidence that could clarify some of the issues or dispel doubts (Berry, 1997; Iwashita, 1997; O’Sullivan, 2000, 2008; Norton 2005; Csépes, 2009). Among the skeptics, Foot (1999) claims that pairing up candidates entails potential problems of mismatch between them with respect to their proficiency levels and/or personality: if the latter are markedly different, they are likely to affect both the performance and the assessment of the candidates. He also proposes that the paired format is an inappropriate test format with low proficiency-level candidates, but may work well at higher levels of proficiency. Foot, however, failed to support his claims empirically. In contrast to Foot’s unfavourable views about the paired format, there have also been positive views expressed in relation to the beneficial impact (reduced stress levels and greater confidence) of the peer-to-peer examination (Lombardo, 1984; Ikeda, 1998; Együd & Glover, 2001). However, these reports are either related to small-scale investigations or lack adequate empirical support.

The peer interlocutor, as a co-constructor of test performance, can be regarded as a potentially important source of variation that may either positively or negatively affect the discourse outcomes and eventually the assessment of candidates' proficiency. It is assumed that the age, sex, educational level, proficiency and personal qualities of the peer interlocutors as well as the familiarity between them may be significant in influencing the performance outcomes in the paired exam (cf. McNamara, 1996). However, the aforementioned factors are likely to have a variable impact in terms of enhancing

or spoiling candidates' test performance in different contexts. Since the most important aspect of performance outcomes seems to relate to assessment ratings, we will examine how some individual characteristics of the peer interactants influence their test scores.

Among the empirical studies exploring the potential effect of the peer interlocutor's proficiency level on performance outcomes, Iwashita (1997), for example, examined assessment ratings in a small-scale investigation with 17 female learners of Japanese, who represented two markedly different proficiency levels. The high and low proficiency students were asked to perform twice: with a similar and a different proficiency level partner. Iwashita's findings showed that subjects of high proficiency did better when they were paired with subjects of the same proficiency. Subjects of low proficiency, however, did better with subjects of high proficiency. However, given the small sample size and the considerable individual differences in the assessment ratings, the interpretation of Iwashita's findings seems to be rather limited. Norton (2005) also claims in the context of Cambridge ESOL (FCE and CAE) speaking tests that having a higher proficiency level partner can be beneficial for the lower level candidate as the latter can incorporate syntactic as well as lexical structures taken from the peer partner.

Contrary to Iwashita's claims, in Csépes' study (2009) conducted in the context of Hungarian school-leavers aged 18, the peer partner's proficiency failed to impact significantly on candidates' ratings. Thirty candidates took three exams with a different proficiency level partner each time. The scores given by two independent raters suggest that their perceptions of candidates' proficiency were neither positively, nor negatively influenced by the fact that the level of proficiency of the peer partners showed considerable variation. This finding also suggests – at least in the given context – that there is no superior performance condition, which runs counter to Iwashita or Norton's earlier claims.

The impact of interlocutor familiarity in paired-task performance has been investigated by O'Sullivan (2008), who explored how measured performance was affected by the degree of acquaintanceship between the test takers as well as the peer interlocutor's gender. The results of his small-scale investigations in relation to acquaintanceship showed a significant effect: in the Japanese context he found that close acquaintanceship positively affected candidates' ratings as they received higher scores when interacting with friends. The significant impact of familiarity on measured performance was also observed in two large-scale investigations by O'Sullivan (2008). In the Turkish context, he found that acquaintanceship was a significant factor, but its impact worked in the opposite direction: Turkish candidates performed better (i.e. got higher ratings) when they had to interact with a stranger. As a conclusion, he suggested that acquaintanceship was likely to be a culture-specific variable, which

means that its effect may be variable depending on the cultural background of the test takers. In the other study with FCE candidates from Italy, Spain and France, acquaintanceship was found to be significant again. His most important recommendation was that candidates should be allowed to choose their partner.

As has been shown above, there are variables in the paired exam that may impact on performance in predictable ways in specific contexts, but in a different testing context the same variables may turn out to be insignificant or a new pattern of interaction between the variables may be found. The potential for unfair biases or the context sensitive nature of the paired exam warrants further investigation. Validity studies are needed to check how scores are affected by differences between pairs in any new test, including a computer-mediated oral test that aims to incorporate the paired mode into its framework.

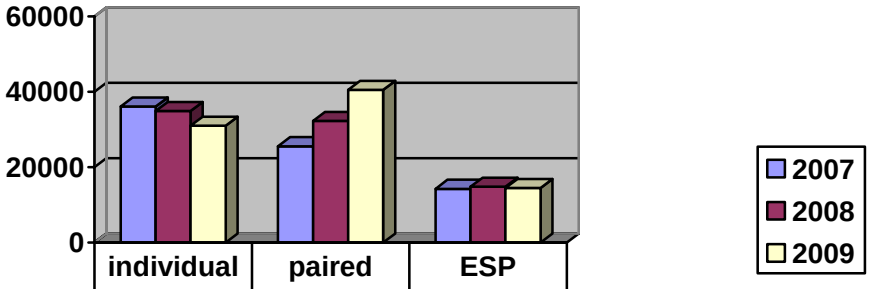
So far little has been said about the design of oral tasks, which is the main source of quality control available to the test designer to ensure that adequate quality and quantity of language can be elicited from the candidates. Since the paired exam is proposed in this paper to compensate for the shortcomings of the interview format, irrespective of the mode of delivery, we will discuss some task design issues in relation to paired tasks. In general, examiner training specifies the form of assistance that the examiner-interlocutor can provide in case of evident imbalance between candidates' contributions or their misunderstanding of the task in the paired exam. Because of the examiner-interlocutor's presence and right for intervention, design faults in pair-tasks can be compensated for in live exams. However, there seem to be specific task characteristics that can enhance candidates' performance and at the same time minimize or exclude any examiner-interlocutor intervention. Based on the lessons learnt from piloting tasks aimed at B1 and B2 of the CEFR within the Hungarian Examination Reform Teacher Support Project (Csépes & Együd, 2004), paired tasks seem to work best if they

- provide candidates with opportunity to bridge an opinion gap between them by asking them to select and/or rank order items in addition to expressing their views on them;
- are symmetric and structured, whereby each candidate is given different prompts in order to create an authentic reason for listening, to raise expectations and to activate prior knowledge and/or experience;
- allow for both agreement and disagreement between the interactants so that they can produce both short and long turns ('allowing for agreement' seems to generate less complex language while more complex and varied language may result from 'allowing for disagreement'),

- are related to *life-like situations* that candidates can identify with;
- require candidates to assume *familiar roles* only.

The above task design considerations seem to apply to the computer-mediated mode as the mode of delivery is unlikely to undermine the reasons or motivation behind candidates' contributions. However, a computer-mediated test using the paired mode cannot rely upon human intervention in the same way as in a direct test. Candidates' contributions can only be controlled through maximizing their response time for a given turn and through appropriate task design.

Candidates' attitudes towards different exam formats or their preference for certain exam formats can also highlight issues of test validity, especially in relation to face validity aspects. To illustrate the status of the paired exam in Hungary, we will draw on the database of the Accreditation Centre for Foreign Language Examinations of the Educational Authority. This database keeps a detailed record of all state accredited foreign language examinations, and the statistics show that the number of successful examinations administered by exam boards that use the paired mode in their speaking test in English at B2 level is clearly on the increase. The increased popularity of the given exams, however, may be attributed to other factors, such as task types used in the other components of the exam. Out of the twenty exam boards five offer their speaking tests in the paired mode (TELC, Cambridge UCLES, ECL, Euro, DExam) and 15 other boards administer speaking tests in the individual mode. However, 10 out of the 15 boards deliver ESP exams in various special fields, such as law, business English, commerce, medicine, etc. The table below shows how the number of successful (passed) exams changed between 2007 and 2009.



	individual	paired	ESP
2007	36109	25544	14280
2008	34904	32305	14811
2009	30987	40541	14491

Table 1 - Number of successful B2 level exams in English between 2007 and 2009

As can also be seen in the chart above, the number of ESP exams is comparable across the years, but the number of exams with the paired oral shows a considerable growth while the one-to-one individual mode can be characterized by a reversed tendency as the numbers reflect a slow, but steady decrease.

Conclusion

In this paper, we have reviewed some major issues with respect to the validity of direct/semi-direct one-to-one oral interviews and the paired mode of direct oral assessments, for which some task design considerations were also proposed. The popularity of the paired-task is also reflected in some of the most widely known performance samples illustrating the levels of the Common European Framework in relation to spoken interaction as they make use of the paired exam mode (e.g. Eurocentres, Cambridge ESOL, CIEP). While several direct tests of speaking seem to advocate the combination the two modes to compensate for the shortcomings of each, computer-mediated tests should similarly seek ways of incorporating interactive features into their design through paired-task performance and utilizing the potential positive washback on teaching and learning.

References

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly*, *25*, 671- 703.

Berry, V. (1997). Ethical considerations when assessing oral proficiency in pairs. In: A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment – Proceedings of LTRC 96* (pp. 107-123). Jyvsäkylä: University of Jyvsäkylä and University of Tampere.

Csépes, I. & Együd, Gy. (2004). *Into Europe: The Speaking Handbook.* Series editor: C. J. Alderson. Budapest: Teleki László Foundation & British Council Hungary.

Csépes, I. (2009). Measuring oral proficiency through paired-task performance. Frankfurt: Peter Lang.

Együd, Gy., & Glover, P. (2001). Oral testing in pairs – a secondary school perspective. *ELT Journal, 55*, 70-76.

Foot, M. C. (1999). Relaxing in pairs. *ELT Journal, 53*, 36-41.

He, A. W., & Young, R. (1998). Language proficiency interviews: a discourse approach. In: R. Young & A. W. He (Eds.), *Talking and testing. Discourse approaches to the assessment of oral proficiency* (pp. 1-26). Amsterdam: John Benjamins Publishing Company.

Hughes, A. (1989). *Testing for language testers*. Cambridge: Cambridge University Press.

Ikeda, K. (1998). The paired learner interview: a preliminary investigation applying Vygotsikan insights. *Language, Culture and Curriculum, 11*, 71-96.

Iwashita, N. (1997). *The validity of the paired interview format in oral performance testing.* Paper presented at the Language Testing Research Colloquium.

Johnson, M., & Tyler, A. (1998). Re-analysing the OPI: How much does it look like natural conversation? In: R. Young & A. W. He (Eds.), *Talking and testing. Discourse approaches to the assessment of oral proficiency* (pp. 27-52). Amsterdam: John Benjamins Publishing Company.

Kormos, J. (1999). Simulating conversations in oral proficiency assessment: a conversation analysis of role play and non-scripted interviews in language exams. *Language Testing, 16*, 163-188.

Lombardo, L. (1984, January). Oral testing: getting a sample of real language. *English Teaching Forum*, pp. 2-6.

Luoma, S. (1997). *Comparability of a tape-mediated and a face-to-face test of speaking*. Unpublished licentiate thesis, University of Jyväskylä, Jyväskylä.

Malabonga, V., Kenyon, D. & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing, 22*, 59-92.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

Norton, J. (2005). The paired format in the Cambridge Speaking Tests. *ELT Journal, 59*, 287-297.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*, 169-192.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System, 28*, 373-386.

O'Sullivan, B. (2008). *Modelling performance in oral language testing*. Frankfurt: Peter Lang.

Ross, S. (1998). Divergent frame interpretations in language proficiency interview interaction. In R. Young & A. Weiyun He (Eds.), *Talking and testing. Discourse approaches to the assessment of oral proficiency* (pp. 1-24). Amsterdam: John Benjamins Publishing Company.

Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*, 159-176.

Saville, N., & Hargreaves, P. (1999). Assessing speaking in the revised FCE. *ELT Journal, 53*, 42-51.

Shohamy, E. (1983). The stability of oral proficiency assessment on the oral interview testing procedures. *Language Learning, 33*, 527-540.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*, 99-123.

Stansfield, C. W. (1989). Simulated Oral Proficiency Interviews. *ERIC Digest*. Washington, DC: ERIC Clearinghouse on Languages and Linguistics.

Stansfield, C. W. (1991). A comparative analysis of simulated and direct proficiency interviews. In: S. Anivan, (Ed.), *Current developments in language testing* (pp. 199-209). Singapore: SEAMEO-RELC.

Stansfield, C. W., & Kenyon, D. M. (1992). The development and validation of a simulated oral proficiency interview. *The Modern Language Journal, 76,* 129-142.

Swain, M. (2001). Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18,* 275-302.

Taylor, L.(2000) Investigating the paired speaking test format. *Research Notes, University of Cambridge Local Examinations Syndicate, 2,* 14-15.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly 23*, 489- 508.

Weir, C. (1990). *Communicative language testing*. Hertfordshire: Prentice Hall International.

Young, R. (1995). Conversational styles in language proficiency interviews. *Language Learning, 45,* 3-42.

Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition, 14,* 403-424.

The author:
Ildikó Csépes
H-4010 Debrecen
P.O.Box 95, Hungary
E-mail: icsepes@delfin.unideb.hu

Ildikó Csépes is a lecturer at the University of Debrecen in Hungary. She has got extensive experience in designing and administering in-house and national level proficiency tests in English. She was involved in the Hungarian Examinations Reform Teacher Support Project of the British Council between 1999 and 2006. Based on the Project's experience in test development and designing an interlocutor/assessor training model, she co-authored *The Speaking Handbook* as part of the INTO EUROPE series. She holds an MA from Lancaster University and a PhD in Language Pedagogy from ELTE, Budapest. Her doctoral research was related to measuring oral proficiency in paired-task performance. Since 2004 she has also been working for the Hungarian Educational Authority as a member of the National Board for Accrediting Foreign Language Examinations. For the last two years she has been acting as the chair of this Board.

THE TESTDAF IMPLEMENTATION OF THE SOPI:

DESIGN, ANALYSIS, AND EVALUATION OF A SEMI-DIRECT SPEAKING TEST

Thomas Eckes, TestDaF Institute, Germany

Abstract

The Test of German as a Foreign Language (*Test Deutsch als Fremdsprache*, TestDaF) is a standardized test designed for foreign learners of German who plan to study in Germany or who require recognized certification of their language ability. In its speaking section, the TestDaF makes use of an adapted version of the Simulated Oral Proficiency Interview (SOPI; Kenyon, 2000; Stansfield & Kenyon, 1988). The present paper discusses the design of the TestDaF speaking instrument and illustrates the measurement approach adopted in order to analyze and evaluate this particular implementation of the SOPI. General requirements for the operational use of the speaking test are delineated and the specific test format is described. The main part deals with the analysis and evaluation of speaking performance ratings obtained from a live TestDaF examination. The paper concludes with perspectives for future test development.

The TestDaF: Purpose and Scope

The TestDaF allows foreign students applying for entry to an institution of higher education in Germany to prove their knowledge of German while still in their home country. Test tasks and items are continuously developed, analyzed, and evaluated at the TestDaF Institute (Hagen, Germany); examinee performance is also centrally scored at this institute (Eckes, 2008a; Eckes et al., 2005; Grotjahn, 2004; see also www.testdaf.de).

The TestDaF measures the four language skills (i.e., reading, listening, writing, and speaking) in separate sections. Examinee performance in each section is related to one of three levels of language ability in the form of band descriptions; these levels (*TestDaF-Niveaustufen*, TestDaF levels, or TDNs for short) are TDN 3, TDN 4, and TDN 5. The TDNs cover the Council of Europe's (2001) Lower Vantage Level (B2.1) to Higher Effective Operational Proficiency (C1.2); that is, the test measures German language ability at an intermediate to high level (see Kecker & Eckes, in press). There is no differentiation among lower ability levels; it is just noted that TDN 3 has not yet been achieved (below TDN 3).

The TestDaF is officially recognized as a language entry exam for students from abroad. Examinees who have achieved at least TDN 4 in each section are eligible for admission to a German institution of higher education (see Eckes et al., 2005).

In April 2001, the TestDaF was administered worldwide for the first time. Until the end of 2009, more than 106,000 students had taken this test. The number of test administrations per year increased from two exams in 2001 to nine exams until present (including three separate exams in the People's Republic of China). Table 1 portrays the growth of the TestDaF candidature from 2001 to 2009, as well as the number of test centers and test countries during this period (see also TestDaF Institute, 2010).

Table 1- *Growth of TestDaF Candidature, Test Centers, and Test Countries*

Year	Test Takers	Test Centers	Test Countries
2001	1,190	81	34
2002	3,582	154	48
2003	7,498	211	65
2004	8,982	261	70
2005	11,052	275	72
2006	13,554	309	74
2007	15,389	318	75
2008	16,882	330	78
2009	18,059	329	77

Speaking Test Requirements

The construct underlying the TestDaF draws on Bachman and Palmer's (1996) model of communicative language ability. More specifically, for each language skill the following areas of language knowledge are taken into account: grammatical knowledge (not assessed separately, but indirectly addressed in each of the TestDaF sections), textual knowledge (coherence/cohesion, rhetorical functions, conversational organization), functional knowledge (ideational, manipulative, heuristic functions), and sociolinguistic knowledge (registers, idiomatic expressions, cultural references). In addition, TestDaF tasks are intended to tap into areas of strategic competence (goal setting, assessment, planning).

Considering the test construct, the TestDaF speaking section is designed as a performance-based instrument, assessing the examinees' ability to communicate appropriately in typical situations of university life. Accordingly, item writers are instructed to produce tasks that elicit language use

relevant to, and characteristic of, this specific context. Beyond this basic, construct-driven requirement for test development, the speaking test has to meet a number of more practical demands. To begin with, trained examiners/interviewers of German as a foreign language are not readily available in many regions of the world. Therefore, the speaking test is to be administered without the use of on-site examiners. Moreover, for reasons of test security as well as cost-effectiveness in terms of test delivery and administration, a large number of examinees are to be tested worldwide on a single test version at the same day.

Further demands relate to issues of standardization, reliability, and validity of the speaking test. Thus, the requirement of standardization says that each examinee receives the same instructions, prompts, and questions as any other examinee taking the test. Ideally, there should be no variation in examiner and/or interviewer input. A high degree of standardization is a prerequisite condition for assuring high reliability of the assessment outcomes. High assessment reliability implies that variation in assessment outcomes that is due to random measurement error is negligibly small. High reliability, in turn, is a necessary but not sufficient condition for high validity. For example, when speaking performance is scored by human raters, differences in rater severity or leniency are generally observed, resulting in variation in assessment outcomes that is not associated with the performance of examinees. As a consequence hereof, examinee speaking ability is not adequately assessed, lowering the validity of the conclusions drawn from the assessment outcomes. It is obvious that in cases like this the assessment instrument has limited fairness as well; that is, some examinees may benefit from being rated by a lenient rater, whereas others may suffer from bad luck in terms of getting a severe rater.

Speaking Test Design

Most of the requirements outlined above are met by the SOPI format. The SOPI is a type of semi-direct speaking test (Luoma, 2004; Qian, 2009), developed in the 1980s at the Center of Applied Linguistics in Washington, DC (for reviews, see Kenyon, 2000; Kuo & Jiang, 1997). This testing format was designed to model the nature of the Oral Proficiency Interview (OPI) used by the American Council on the Teaching of Foreign Languages (ACTFL). Whereas the OPI is a face-to-face interview, the SOPI relies on pre-recorded prompts and a printed test booklet to elicit language from the examinee.

Early research suggested that the SOPI is a reliable and valid technology-based alternative to the OPI (Kenyon, 2000; Stansfield & Kenyon, 1992). For example, Stansfield and Kenyon (1992) performed a number of correlation studies and concluded that “the OPI and the SOPI are close enough in the way they measure general speaking proficiency that they may be viewed as parallel tests delivered in two different formats” (p. 359). However, researchers have also provided evidence that direct and semi-

direct speaking tests may tap different language abilities, in particular, interactive versus monologic speaking ability (see, e.g., O'Loughlin, 2001; Shohamy, 1994; see also Galaczi, this volume). Qian (2009) added another facet to the comparison between direct and semi-direct speaking tests. The author studied affective effects of direct and semi-direct modes for speaking assessment on test-takers and found that "a large proportion of the respondents in the study were quite willing to accept both testing modes" (p. 123).

The typical full-length SOPI, used to assess examinees from the Novice to the Superior level on the ACTFL scale, comprises a total of 15 tasks measuring general speaking ability in a foreign or second language. The SOPI currently in use with the TestDaF is an adapted version consisting of seven speaking tasks (including a warm-up task) tailored to the German academic context. Following this format, the speaking test is administered via audio-recording equipment using prerecorded prompts and printed test booklets. That is, during testing the examinee listens to directions for speaking tasks from a master tape or CD while following along in a test booklet; as the examinee responds to each task, his or her speaking performance is recorded on a separate response tape or CD. Testing time is about 30 minutes. Before being put to operational use, each speaking task is carefully examined in an elaborate evaluation process comprising piloting and trialling stages (see Eckes, 2008a). Table 2 provides an overview of the main features characterizing the TestDaF speaking section.

Table 2 - *Overview of the TestDaF Speaking Assessment Instrument*

Feature	Description
Construct	Ability to communicate appropriately in typical situations of university life
Format	Semi-direct (SOPI); adapted version (German academic context); seven tasks (one warm-up, two tasks each for TDN 3, TDN 4, and TDN 5); testing time is 30 min.
Administration	Tape-mediated or computer-assisted; prompts are pre-recorded and text based (printed test booklet); the examinee speaks into a microphone while the response is recorded
Roles/Situations	Examinees act out themselves in conversations with other students, employees at university, professors, etc.; the conversations are situated in seminars, language courses, cafeteria, etc.
Register	formal, informal, semi-formal
Rating	Experienced and trained raters; analytic rating scheme; top-down rating procedure (performance on top-level TDN 5 tasks rated first)

Speaking tasks are presented to all examinees in the same, fixed order. In the first task, the "warm-up", the examinee is asked to make a simple request; performance on this task is not rated. The other tasks focus on situation-related communication (e.g., obtaining and supplying information), relate to "describing", or deal with "presenting arguments". Two tasks each probe into one of the relevant proficiency levels (i.e., TDN 3, TDN 4, or TDN 5). The test ends with a "wind-down" consisting of a less-challenging task intended to put the examinees at ease before leaving the examination. Table 3 shows how the speaking tasks progress in terms of the challenge (i.e., TDN level) they pose to examinees.

Table 3 - *Progression of Speaking Task Difficulty Level*

	Task No.						
Level	1	2	3	4	5	6	7
TDN 3	x	x					x
TDN 4			x		x		
TDN 5				x		x	

Note. Each "x" means that a given task is presented at the level indicated by the row. Task No. 1 is a warm-up task; performance on this task is not rated.

In each task, examinees are asked to act out themselves in simulated conversations with other students, employees at a university office, lecturers, professors, and so on. These conversations are typically situated in familiar academic settings, including seminars, language courses, and cafeterias. According to the different contexts of language use the relevant registers cover a wide range of formal, informal, and semi-formal varieties. Table 4 shows in a summary fashion which kind of speech act each task requires from the examinees. In the instructions to the test, examinees are explicitly asked to take the specific content of each task into account when responding. Figure 1a and Figure 1b present examples of speaking tasks aiming at proficiency levels TDN 3, TDN 4, and TDN 5, respectively. These tasks are taken from a TestDaF sample speaking test that is available online for test preparation purposes at www.testdaf.de (see the links "Für Teilnehmerinnen und Teilnehmer", "Vorbereitung", "Modellsatz 02").

Table 4 - *Speaking Tasks, TDNs, and Required Speech Acts*

Task No.	TDN	Speech Act
1	3	Asking for information
2	3	Reporting / describing cultural facts
3	4	Describing a diagram or chart
4	5	Commenting on a topic / balancing pros and cons (socio-political/sociocultural area)
5	4	Stating one‘s opinion on a particular topic (personal area)
6	5	Forming / presenting hypotheses based on a diagram
7	3	Giving advice / giving reasons

Note. Task No. 1 is a warm-up task; performance on this task is not rated.

Figure 1a. Sample speaking test. Task 2 (at TDN 3) is shown in the upper half, Task 5 (at TDN 4) in the lower half of page 69.

Figure 1b. Sample speaking test. Task 6 (at TDN 5) is shown in the upper half, the chart that this task refers to in the lower half of page 70.

Ihr Studienfreund Martin möchte aus der Wohnung seiner Eltern ausziehen und sucht deshalb eine neue Wohnung. Er fragt Sie, wie lange die jungen Leute in Ihrem Heimatland bei ihren Eltern leben.

Beschreiben Sie,
– wann junge Menschen in Ihrem Heimatland von zu Hause ausziehen und
– warum sie ihr Elternhaus verlassen.

Sie: Vorbereitungszeit — **1 Minute**

Martin: ...

Sie: Sprechzeit — **1 Minute**

Ihr Freund Steffen muss während seines Studiums ein Praktikum machen. Er hat zwei Möglichkeiten: Steffen kann das Praktikum entweder in der Firma seiner Eltern absolvieren. Oder er macht sein Praktikum in einem anderen Betrieb. Steffen fragt Sie nach Ihrer Meinung.

Sagen Sie Steffen, wozu Sie ihm raten:
– Wägen Sie Vorteile und Nachteile der beiden Möglichkeiten ab.
– Begründen Sie Ihre Meinung.

Sie: Vorbereitungszeit — **2 Minuten**

Steffen: ...

Sie: Sprechzeit — **1 Minute 30 Sekunden**

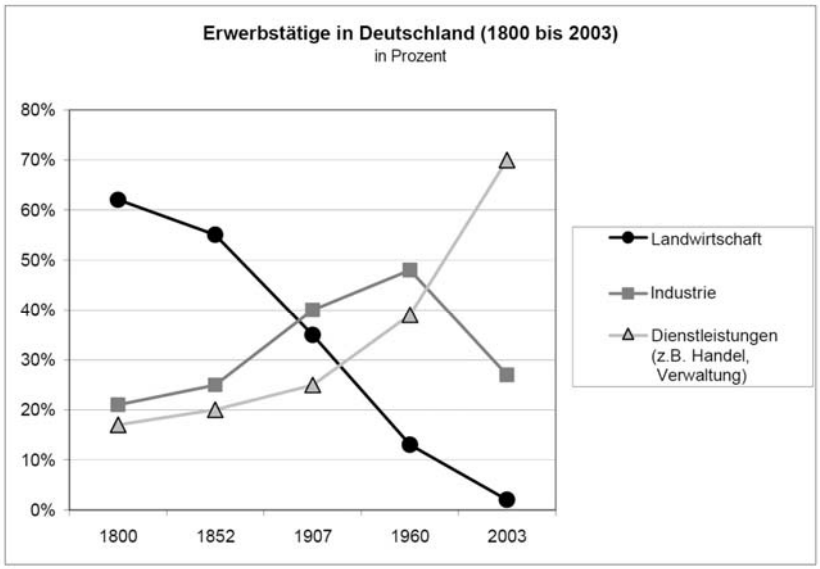
In Ihrem Wirtschaftsseminar geht es heute um die Veränderungen im Bereich Erwerbstätigkeit in Deutschland. Ihre Dozentin, Frau Dr. Maier, hat eine Grafik verteilt, die zeigt, in welchen Wirtschaftsbereichen die Menschen arbeiten. Frau Dr. Maier bittet Sie, Ihre Überlegungen zu Gründen der bisherigen Entwicklung und zur zukünftigen Entwicklung vorzutragen.

Nennen Sie mögliche Gründe für die dargestellte Entwicklung.
Stellen Sie dar, welche Entwicklung Sie für die Zukunft erwarten.
Begründen Sie Ihre Überlegungen anhand der Grafik.

Sie: Vorbereitungszeit — **3 Minuten**

Frau Dr. Maier: ...

Sie: Sprechzeit — **2 Minuten**



Nach: Statistisches Bundesamt, Wiesbaden 2004

All speaking tasks follow the same general design. Thus, each task has two clearly separated parts. In the first part, the communicative situation is described, and the examinee is told what to do. The examinee can read the task along in the test booklet. The booklet also shows the time to plan a response. Depending on the task, planning time ranges from 30s to 180s. During this time the examinee is allowed to take notes. In the second part, the "interlocutor" speaks, and the examinee is asked to listen and to respond after that. The time to respond is shown in the booklet. Again depending on the task, response time ranges from 30s to 120s. The examinee is free to stop responding before the response time is over.

Rating Speaking Test Performance

In direct speaking tests like the OPI, interviewer behavior is likely to exert considerable influence on the assessment outcomes, thus contributing to construct-irrelevant variance in examinee scores (see, e.g., Bachman, 2002; Brown, 2005; O'Sullivan, 2008). Due to its format, the SOPI eliminates any examiner or interviewer variability. Another important construct-irrelevant source of variability, however, remains largely unaffected by the SOPI format. This source relates to characteristics of the raters evaluating the quality of examinee responses after the test is completed.

Raters involved in rating examinee performance on TestDaF speaking tasks are subjected to elaborate training and monitoring procedures. Yet, extensive rater training sessions and detailed rater monitoring and feedback practices notwithstanding, rater variability generally remains a major unwanted component of examinee score variance (see, e.g., Hoyt & Kerns, 1999; Knoch, Read & von Randow, 2007; Lumley & McNamara, 1995; O'Sullivan & Rignall, 2007; Weigle, 1998). One reason why it is so difficult, if not impossible, to reduce rater variability to an acceptably low level in most assessment contexts is that this variability can manifest itself in a number of ways. For example, rater variability may take the form of (a) differences in rater severity or leniency, (b) differences in raters' understanding and use of rating scale categories, (c) differences in the kind of performance features raters attend to, (d) differences in the way raters interpret and use scoring criteria, or (e) various interactions of raters with examinees, tasks, criteria, and other facets of the assessment situation (see, e.g., Brown, 2005; Eckes, 2008b, 2009b; McNamara, 1996; Lumley, 2005).

The usual, or standard, approach to resolving the rater variability problem, especially in high-stakes assessments, consists of three components: rater training, independent ratings of the same performance by two or more raters (repeated ratings), and establishing interrater reliability. Since this approach has been shown to encounter difficulties of the kind mentioned above, another approach that has attracted a great deal of attention lately is to eliminate rater variability altogether by replacing human rating with

fully automated scoring (see, e.g., Chapelle & Chung, 2010; Shermis & Burstein, 2003; Williamson, Bejar & Mislevy, 2006).

For assessing writing performance, automated scoring systems have been in place for quite a while. Recent examples include *e-rater*, a system operationally used with the TOEFL iBT writing section (Attali, 2007; Weigle, 2010), the *Intelligent Essay Assessor* (IEA), used with the Pearson Test of English (PTE) Academic (Pearson, 2009), or *IntelliMetric* (Elliot, 2003; Wang & Brown, 2007). With some delay, automated scoring systems for assessing speaking performance have followed suit, such as *SpeechRater* (Xi, Higgins, Zechner & Williamson, 2008; Zechner, Higgins & Xi & Williamson, 2009), or the PTE Academic speaking test, which makes use of Ordinate technology (Bernstein & Cheng, 2007; see also van Moere, this volume).

The approach adopted within the context of the TestDaF speaking test is to continue employing human raters, but to compensate for differences in rater severity by means of a measurement approach that is based on the many-facet Rasch model (Linacre, 1989; Linacre & Wright, 2002). This approach is discussed in some detail next.

Speaking Test Analysis and Evaluation

The many-facet Rasch measurement (MFRM) model allows the researcher to examine more variables (or "facets") than the two that are typically included in a paper-and-pencil testing situation (i.e., examinees and items). Thus, in speaking performance assessments, additional facets that may be of particular interest refer to raters, speaking tasks, and scoring criteria. Within each facet, each element (i.e., each individual examinee, rater, task, or criterion) is represented by a parameter. These parameters denote distinct attributes of the facets involved, such as proficiency (for examinees), severity (for raters), and difficulty (for tasks or criteria).

Viewed from a measurement perspective, an appropriate approach to the analysis of many-facet data would involve three general steps. Step 1 refers to a careful inspection of the assessment design; that is, relevant issues to be considered at this stage concern the sample of examinees at which the assessment is targeted, the selection of raters to be used in the assessment, the nature of the speaking tasks, and many others. Step 2 concerns the specification of an appropriate measurement model; for example, determining the facets to be examined, or defining the structure of the rating scale. Step 3 calls for implementing that model in order to provide a fine-grained analysis and evaluation of the functioning of each of the facets under consideration (for a detailed discussion, see Eckes, 2009a, in press).

In what follows, I illustrate relevant features of the many-facet Rasch analysis routinely applied to the TestDaF rater-mediated system of speaking performance assessment (see also Eckes, 2005). The

database consisted of ratings of examinee performance on a speaking test as part of a live exam that took place in July 2008.

Examinees

The speaking test was administered to 1,771 participants (1,072 females, 699 males). Participants' mean age was 24.62 years (*SD* = 5.02); 87.0% of participants were aged between 18 and 30 years.

There were 152 TestDaF test centers involved in this administration (104 centers in Germany, 48 centers in 33 foreign countries). In terms of the number of examinees, the following five national groups ranked highest (percentage in parentheses): People's Republic of China (9.8%), Russia (9.4%), Ukraine (6.3%), Turkey (6.1%), Poland (4.6%).

Raters

Thirty-seven raters participated in the scoring of examinee speaking performance. Raters were all experienced teachers and specialists in the field of German as a foreign language, and were systematically trained and monitored to comply with scoring guidelines.

Procedure

Ratings of examinee speaking performance were carried out according to a detailed catalogue of performance aspects comprising eight criteria. The first two criteria (*comprehensibility, content*) were more holistic in nature, referring to the *overall impression* upon first listening to the oral performance, whereas the others were more of an analytic kind, referring to various aspects of *linguistic realization* (*vocabulary*, *correctness*, *adequacy*) and *task fulfillment* (*completeness*, *argumentation*, *standpoint*). On each criterion, examinee performance was scored using the four-point TDN scale (with categories *below TDN 3*, *TDN 3*, *TDN 4*, *TDN 5*).

Ratings were provided according to a top-down procedure. Thus, raters started with rating examinee performance on the two most challenging tasks (i.e., TDN 5 tasks). If the examinee was clearly a top-level speaker, then it was not necessary for the rater to listen to the examinee's performances on any of the lower-level tasks. Otherwise, the performance ratings were continued at TDN 4. If the examinee was clearly a speaker at that level, then the rating was finished; if not, examinee performances on the least-challenging tasks were also rated. In general, this procedure serves to increase rater efficiency and to prevent rater fatigue or waning of rater attentiveness. It is particularly efficient when used with digitally recorded speaking performances saved on CD, allowing the raters to skip forward and back within the audio file to quickly locate the next performance to be rated.

Each examinee performance was rated by a single rater. Such a rating design calls for measures to satisfy the precondition of connectivity of the resulting sparse data matrix. That is, all raters, examinees, tasks, and criteria were to be connected in the design such that they could be placed in a common frame of reference (Linacre & Wright, 2002). To generate a connective data matrix, each rater had to provide ratings for the same set of performances, in addition to his or her normal workload. The additional performances, representing the range of TDN levels, had been pre-selected from a larger set of examinee performances in a previous trialling of the respective writing task.

Data analysis

The rating data were analyzed by means of the computer program FACETS (Version 3.66; Linacre, 2010). The program used the ratings that raters awarded to examinees to estimate individual examinee proficiencies, rater severities, task and criterion difficulties, respectively, and scale category difficulties. FACETS calibrated the examinees, raters, tasks, criteria, and the rating scale onto the same equal-interval scale (i.e., the logit scale), creating a single frame of reference for interpreting the results of the analysis (for an introductory overview of MFRM, see Eckes, 2009a, in press).

Variable map

Figure 2 displays the variable map representing the calibrations of the examinees, raters, tasks, criteria, and the four-point TDN rating scale as raters used it to rate examinee speaking performance.

The variability across raters in their level of severity was substantial. Thus, the rater severity measures showed a 2.70-logit spread, which was more than a fifth (21.7%) of the logit spread observed for examinee proficiency measures (12.46 logits). In other words, differences in rater severity were far from being negligible. This was consistently revealed by rater separation statistics: (a) the fixed chi-square value was highly significant, indicating that at least two raters did not share the same parameter (after allowing for measurement error), (b) the rater separation index showed that within the present group of raters there were about 19 statistically distinct strata of severity (to illustrate, when raters exercised a similar level of severity, an index value close to 1 would be expected), and (c) the rater separation reliability was close to unity, attesting to a very high amount of rater disagreement.

Figure 2. Variable map from the FACETS analysis of TestDaF speaking performance data. Each star in the second column represents 17 examinees, and a dot represents fewer than 17 examinees. The horizontal dashed lines in the last two columns indicate the category threshold measures for the four-category TDN scale (for tasks at TDN 5; i.e., Task 4 and Task 6) and for the three-category TDN scale

(for tasks at TDN 4; i.e., Task 2 and Task 5), respectively; the thresholds for the two-category scale coincide with the difficulty measures of tasks at TDN 3 (i.e., Task 2 and Task 7).

Logit	Examinee	Rater	Task	Criterion	TDN Scales	
	High	***Severe***	***Difficult***	***Hard***	**(TDN 5)**	**(TDN 4)**
7	.					
	.					
	.					
6	.					
	.					
	.					
5	.					
	*.					
	**.					
4	**.					
	****.					
	****.					
3	***.					
	*****.					
	****.				**-----**	
2	*******.					
	*********					
	**********	**	**6**			**-----**
1	*********.	*	**4**		**4**	
	*******.	*****		**standpoint**		
	*****.	******	**3**	**argument. correctness**		
0	***.	**********		**content completeness**	**-----**	**3**
	***.	*****	**5**	**adequacy**		
	**.	*****		**comprehens. vocabulary**		
-1	**.	*	**2**		**3**	
	*.	**				**-----**
	*.		**7**			
-2	*.					
	.				**-----**	
	.					
-3	.					
	.					
	.					
-4	.					
	.					
	.					
-5	.					
	.					
-6	.					
	Low	***Lenient***	***Easy***	***Easy***	**(below 3)**	**(below 3)**

Compensating for rater severity differences

Expressing the rater severity differences in the metric of the TDN scale showed that the most severe rater provided ratings that were, on average, 0.95 raw-score points lower than those provided by the most lenient rater. That is, the severity difference between these two raters was almost one TDN level. Obviously, then, rater severity differences in the order revealed here can have important consequences for examinees.

A case in point is Examinee 561 (see Table 5). Based on the MFRM model, this examinee had an estimated proficiency of 1.99 logits (*SE* = 0.27); the observed average was 3.27. Rounding the observed average to the next TDN level, the final level awarded would have been TDN 3. By contrast, expressing the examinee's proficiency measure in terms of the TDN metric yielded an average of 3.64. This so-called *fair average* is higher than the observed average because it resulted from compensating for the severity of Rater 15 (severity measure = 0.81 logits, *SE* = 0.04), who had happened to rate this examinee's speaking performance. Again rounding to the next TDN level, the examinee would have been awarded the final level TDN 4, making him or her eligible for university admission. Thus, in the case of Examinee 561, applying the many-facet Rasch analysis would have led to an *upward adjustment* of this examinee's result on the speaking test. The same kind of adjustment would have occurred with Examinee 1170, though one TDN level up the scale.

Conversely, Examinee 335 would have received a *downward adjustment* based on the fair average (2.39; below TDN 3), as opposed to the observed average (2.71; TDN 3). In fact, this examinee's performance had been rated by the most lenient rater in the group (Rater 27, severity = –1.30 logits, *SE* = 0.05). Hence, as compared to the other raters in the group, Rater 27 overestimated the proficiency of Examinee 335. This overestimation was corrected by the proficiency measure (or fair average) provided by the MFRM analysis. Table 5 also shows two examinees (i.e., 515, 1631) whose TDN assignments remained unaffected by the score adjustment.

Table 5 - *Examinee Measurement Results (Illustrative Examples)*

Examinee	Proficiency Measures	*SE*	*N* Ratings	Observed Average	Fair Average
515	3.92	0.56	16	4.75	4.84
1170	3.70	0.44	16	4.44	4.81
561	1.99	0.27	48	3.27	3.64
335	-1.87	0.25	48	2.71	2.39
1631	-1.89	0.44	48	2.12	2.38

Note. Proficiency measures are shown in units of the logit scale. *SE* = Standard error. Fair Averages present examinee proficiency measures in units of the TDN scale (with scores from "2" for the lowest category to "5" for the highest category).

Further findings from the MFRM analysis

The MFRM approach makes available a wide variety of analytic procedures and statistical indicators that help to evaluate the quality of speaking performance ratings in almost any desired detail. Probing into the degree of rater variability and compensating for differences in rater severity illustrate some of the practical benefits that may accrue from using MFRM models – clearly important benefits from the point of view of examinees and other stakeholders. Of course, there is much more to be learned from a MFRM analysis of performance ratings. Due to space restrictions, I only briefly outline some further results of the MFRM analysis relevant for an evaluation of the TestDaF speaking test.

Important information on the overall functioning of the speaking performance assessment is provided by the examinee separation reliability statistic. This statistic indicates how well one can differentiate among the examinees in terms of their levels of proficiency. Usually, performance assessment aims to differentiate among examinees in terms of their proficiency as well as possible. Hence, high examinee separation reliability is the desired goal. For the present data, the MFRM analysis showed that this goal had been achieved (reliability = .96). A formally related, practically useful statistic is the examinee separation, or number of examinee strata, index. This index gives the number of measurably different levels of examinee proficiency. In our sample of examinees, the separation index was 6.83, which suggested that among the 1,771 examinees included in the analysis, there were almost seven statistically distinct classes of examinee proficiency. Note that this finding nicely related to the four-category TestDaF scale. That is, the measurement system worked to produce at least as much reliably different levels of examinee proficiency as the TestDaF speaking section was supposed to differentiate.

For each element of each facet, a MFRM analysis provides *fit indices* showing the degree to which observed ratings match the expected ratings that are generated by the model. Regarding the rater facet, fit indices provide estimates of the consistency with which each individual rater made use of the scale categories across examinees, tasks, and criteria. In the present analysis, rater fit indices showed that the vast majority of raters provided highly consistent ratings. Raters exhibiting a tendency toward inconsistency can be subjected to specific rater training in order to reduce this kind of variability (see, e.g., Elder, Knoch, Barkhuizen & von Randow, 2005; Weigle, 1998; Wigglesworth, 1993).

Another point of interest concerns the relative difficulty of the individual speaking tasks. Considering the design of the TestDaF speaking test, Tasks 2 and 7 are supposed to aim at proficiency level TDN 3, Tasks 3 and 5 at TDN 4, and Tasks 4 and 6 at TDN 5. The task measurement results indicated that these three groups of tasks were nicely ordered from less difficult to highly difficult. At the same time, however, difficulty measures *within* two of these groups (i.e., TDN 3 and TDN 4 tasks, respectively) differed significantly by about half a logít (see also Figure 2). Ideally, there should be no significant

difference between tasks aiming at the same TDN. Hence, a finding such as this one needs to be addressed in discussions with item writers in order to come closer to the intended difficulty distribution.

The measurement results for the criterion facet showed that *standpoint* was the most difficult criterion; that is, examinees were less likely to receive a high rating on this criterion than on any of the other criteria; at the opposite end of the difficulty scale were *comprehensibility* and *vocabulary*, which proved to be the easiest ones (see also Figure 2). The criterion difficulty measures showed a 1.23-logit spread, which was quite in line with what could be expected in the present assessment context. Importantly, fit indices for each of the eight criteria stayed well within even narrow quality control limits, attesting to psychometric unidimensionality of this set of criteria (Henning, 1992; McNamara, 1996). That is, all criteria seemed to relate to the same latent dimension, as assumed by the Rasch model used here.

Since the input data to the MFRM analysis were ratings provided on an ordinal scale, the question arises as to how well the categories on the TDN scale, that is, the scores awarded to examinees, are separated from one another. The analysis typically provides a number of useful indices for studying the functioning of rating scales. For example, for each rating scale category, the average of the examinee proficiency measures that went into the calculation of the category calibration measure should advance monotonically with categories. When this pattern is borne out in the data, the results suggest that examinees with higher ratings are indeed exhibiting “more” of the variable that is being measured than examinees with lower ratings. In the present analysis, the findings strongly confirmed that the TDN rating scale categories were properly ordered and working as intended.

More Validity Evidence

As shown above, the overall functioning of the present speaking performance assessment as assessed by means of the examinee separation reliability was highly satisfactory. Computing this statistic across all TestDaF speaking tests administered so far revealed that in not a single case the examinee separation reliability fell below .94. Corroborating this finding, the number of strata index computed on the same data basis ranged from 5.5 to 6.5. Hence, each and every live exam reliably differentiated at least five classes of examinees in terms of their level of speaking proficiency – a clear indication scoring validity (Weir, 2005).

Additional evidence of validity was provided by a benchmarking study (Kecker & Eckes, in press) that followed the empirical approach as suggested by the CEFR manual (Council of Europe, 2003). In this study, experts rated each of nine spoken production samples taken from a live TestDaF exam on a nine-category CEFR scale covering the levels A1 to C2 (including plus levels). Ratings were analyzed

using the FACETS program. The results confirmed that the pre-assigned TDN levels and the intended CEFR levels were in close agreement, except for two samples that appeared to be placed too low on the CEFR scale.

In a related validation study, Kecker and Eckes (in press) examined the extent to which TDN levels that examinees achieved in a live TestDaF speaking test corresponded to the same examinees' CEFR levels awarded to them by their teachers using the global CEFR scale. Cross-tabulation of teacher-assigned CEFR levels and TDN levels revealed that in 72.4% of the cases the levels were as predicted; that is, almost three out of four examinees received TDN levels that were in line with the expected B2–C1 range along the CEFR scale.

Summary and Conclusion

Since its inception in 2001, the TestDaF adaptation of the SOPI to the German academic context has been successful in terms of acceptance by stakeholders, ease and efficiency of administration, and scoring validity. Over the years, tape-mediated test delivery has been increasingly replaced by computer-assisted test administration. This technological advance has contributed to the spreading of the TestDaF in the world, and it has also contributed (in combination with the top-down rating procedure) to the improvement of scoring efficiency.

Concomitant analysis and evaluation of the ratings of examinee speaking performance has shown that, overall, within-rater consistency is sufficiently high. Hence, ongoing rater training and rater monitoring activities have worked out in this respect. However, as evidenced on every occasion by many-facet Rasch analysis of the rating data, between-rater differences in severity have remained at a level that is much too high to be ignored. In order to compensate for severity differences, final TDN levels are assigned to examinees based on the computation of fair averages.

Currently, detailed qualitative and quantitative feedback from examinees, language teachers, and exam officers has stimulated thinking about further improvements that may be achieved in the future. Some issues that figure prominently in this process concern the following questions: (a) Are the fixed planning and response times adequately designed? (b) Are the required speech acts sufficiently distinct and relevant to contemporary academic context? (c) Are the scoring criteria and performance-level descriptors clearly defined and well-differentiated from one another? (d) Are the less-challenging tasks that aim at level TDN 3 really located at the intended CEFR level (i.e., B2)?

Recently, the Center for Applied Linguistics has developed an approach to assessing speaking proficiency that is more flexible than the SOPI. This approach utilizes the Computerized Oral Proficiency Instrument (COPI; Malabonga, Kenyon & Carpenter, 2005; see also Kenyon & Malone, this volume). The COPI allows examinee control over several aspects of test administration, including

control of the time they take to prepare for and respond to a COPI task. Therefore, at least the first question mentioned above may be readily addressed by following the lines laid out by the COPI.

Acknowledgements

I would like to thank my colleagues at the TestDaF Institute for many stimulating discussions on issues concerning the evaluation of the TestDaF speaking test. Special thanks go to Gabriele Kecker for helpful comments on an earlier version of this paper.

References

Attali, Y. (2007). *Construct validity of* e-rater *in scoring TOEFL essays* (Research Report, RR-07-21). Princeton, NJ: Educational Testing Service.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*, 453–476.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bernstein, J., & Cheng, J. (2007). Logic and validation of fully automatic spoken English test. In M. Holland & F. P. Fisher (Eds.), *The path of speech technologies in computer assisted language learning: From research toward practice* (pp. 174–194). Florence, KY: Routledge.

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt, Germany: Lang.

Chapelle, C. A., & Chung, Y.-R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*. Advance online publication. doi: 10.1177/0265532210364405

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2003). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEF)*. Manual (preliminary pilot version). Strasbourg: Language Policy Division.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*, 197–221.

Eckes, T. (2008a). Assuring the quality of TestDaF examinations: A psychometric modeling approach. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference May 2005* (pp. 157–178). Cambridge, UK: Cambridge University Press.

Eckes, T. (2008b). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155–185.

Eckes, T. (2009a). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the*

manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H). Strasbourg, France: Council of Europe/Language Policy Division. Retrieved from http://www.coe.int/t/dg4/linguistic/manuel1_EN.asp?#P19_2121

Eckes, T. (2009b). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment: Proceedings of the 28th Language Testing Research Colloquium* (pp. 43–73). Frankfurt/Main, Germany: Lang.

Eckes, T. (in press). *Many-facet Rasch measurement: An introduction*. Frankfurt/Main, Germany: Lang.

Eckes, T., Ellis, M., Kalnberzina, V., Pižorn, K., Springer, C., Szollás, K., & Tsagari, C. (2005). Progress and problems in reforming public language examinations in Europe: Cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France, and Germany. *Language Testing*, *22*, 355–377.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, *2*, 175–196.

Elliot, S. (2003). IntelliMetric™: From here to validity. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Mahwah, NJ: Erlbaum.

Grotjahn, R. (2004). TestDaF: Theoretical basis and empirical research. In M. Milanovic & C. J. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference July 2001* (pp. 189–203). Cambridge, UK: Cambridge University Press.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*, 1–11.

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*, 403–424.

Kecker, G., & Eckes, T. (in press). Putting the Manual to the test: The TestDaF–CEFR linking project. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft Manual.* Cambridge, UK: Cambridge University Press.

Kenyon, D. M. (2000). Tape-mediated oral proficiency testing: Considerations in developing Simulated Oral Proficiency Interviews (SOPIs). In S. Bolton (Ed.), *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests* [TESTDAF: Foundations of developing a new language test] (pp. 87–106). München, Germany: Goethe-Institut.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26–43.

Kuo, J., & Jiang, X. (1997). Assessing the assessments: The OPI and the SOPI. *Foreign Language*

Annals, 30, 503–512.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (2010). *Facets Rasch model computer program* [Software manual]. Chicago: Winsteps.com.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3,* 484–509.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective.* Frankfurt, Germany: Lang.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12,* 54–71.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing,* 22, 59–92.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

O'Loughlin, K. J. (2001). *The equivalence of direct and semi-direct speaking tests*. Cambridge, UK: Cambridge University Press.

O'Sullivan, B. (2008). *Modelling performance in tests of spoken language.* Frankfurt, Germany: Lang.

O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS Writing Module. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 446–478). Cambridge, UK: Cambridge University Press.

Pearson. (2009). *PTE Academic automated scoring.* Retrieved from http://www.pearsonpte.com/SiteCollectionDocuments/AutomatedScoringUK.pdf

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly, 6,* 113–125.

Shermis, M. D., & Burstein, J. C. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11,* 99–123.

Stansfield, C. W., & Kenyon, D. M. (1988). *Development of the Portuguese speaking test*. Washington, DC: Center for Applied Linguistics.

Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20,* 347–364.

TestDaF Institute. (2010). *Jahresbericht 2008/09* [Annual report 2008/09]. Hagen, Germany: Author.

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment, 6,* 4–28.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263–287.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*. Advance online publication. doi: 10.1177/0265532210364406

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Basingstoke, UK: Palgrave Macmillan.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*, 305–335.

Williamson, D. M., Mislevy, R. J., & Bejar I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Erlbaum.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater*^{SM} *v1.0* (Research Report, RR-08-62). Princeton, NJ: Educational Testing Service.

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

The author:
Thomas Eckes
TestDaF Institute
Feithstraße 188
D - 58084 Hagen
E-mail: thomas.eckes@testdaf.de

Thomas Eckes has been at the TestDaF Institute, Hagen, Germany, since 2001. He is currently Deputy Director and Head of the Language Testing Methodology, Research, and Validation unit. He has extensive teaching experience in educational measurement and statistics, as well as in cognitive and social psychology. He has published numerous articles in edited volumes and peer-review journals, including *Language Testing*, *Language Assessment Quarterly*, *Diagnostica*, *Journal of Classification*, *Multivariate Behavioral Research*, and *Journal of Personality and Social Psychology*. Recently, he has contributed a chapter on many-facet Rasch measurement to the CEFR Manual Reference Supplement, available online at www.coe.int/lang. His research interests include: rater cognition and rater behaviour; rater effects; polytomous IRT models; construct validation of C-tests; standard setting; computerized item-banking; Internet-delivered testing.

AUTOMATED SPOKEN LANGUAGE TESTING: TEST CONSTRUCTION AND SCORING MODEL DEVELOPMENT

Alistair Van Moere
Knowledge Technologies, Pearson

Abstract

This paper proposes solutions for using computers to assess spoken competencies in second and third languages. It is argued that if we would like to compare the oral skills of learners of L2 and L3 in different regions, then we need to apply a set of tasks which meet certain criteria – the tasks should be: cognitively simple, performance-rich, yet practical to administer and score. A fully automated test is described which meets these criteria. The tasks are simple and consist of repeating sentences, responding to questions, and re-telling stories. These elicit speech acts of varying lengths, from a few words to a few sentences. The test is practical in the sense that it is both administered and scored by computer, thereby ensuring standardized test conditions and reliable, consistent scoring. Further, when using speech processing technology, the learner's performances are seen to be far richer than many people realize. The speech processors analyse features of the learner's responses into dozens of component parts to provide subskill scores such as sentence mastery, vocabulary, fluency, and pronunciation.

1. Introduction

Fully automated testing of spoken language has been in use for over a decade. The term "fully automated" is used, rather than "computer assisted", because in this approach there is no human intervention from the time the test-taker sits down to take the test, to the time the test scores are retrieved. Scoring is conducted by means of speech recognition software and algorithms which assign scores to certain features of speech (Balogh and Bernstein, 2006).

There are numerous benefits to adopting an automated approach. First, automated tests can be taken on demand, anytime and anywhere, without the logistical difficulties of scheduling trained examiners into exam rooms with test-takers. Second, automated tests are easy to deliver and can be taken from virtually any landline telephone or computer with minimum hardware requirements. Third, the test presentation is identical regardless of the test-taker's location because test questions are pre-recorded and so are delivered to the test-taker with standardized speed, clarity, and word choice. Fourth, the scoring is computerized and scores are returned to the test-taker within minutes. Fifth, the speech processors and scoring algorithms produce scores that are practically the same as expert human

evaluations of a test-taker's responses (Pearson, 2009a, 2009b). And finally, the automated approach is very reliable; the machine will always give the same score to the same performance, will not be biased by accent or irrelevant variables, and does not become prone to fatigue, distraction, or idiosyncratic interpretation of a rating criteria. Although every test context is unique and some situations will require human-administered face-to-face testing, the many advantages of the automated approach certainly make it suitable for general L2 or L3 spoken proficiency testing, including language surveys.

This paper continues in Section 2 by giving a brief overview of the contexts in which automated scoring is currently used, the degree to which automated scores correlate with scores assigned by trained human examiners, and how they are benchmarked to the CEFR. Section 3 describes the test administrators' toolkit, including online score access and student record. Section 4 describes the system for speech and language processing, and why even one-sentence spoken responses from test-takers can be considered so data-rich. Section 5 addresses what some critics argue is a limitation to the approach, which is the construct that is assessed. Although some argue that automated tests are not sufficiently communicative, section 5 describes how tasks in the automated approach tap the very same constructs that underlie spoken communication. Section 6 describes the tasks that might be useful for a language survey, as well as the subscores and feedback which could be generated from an automated test.

2. Automated tests in use

Fully automated testing is the most recent among three main innovations in the evolution of spoken language testing (see Table 1). The first innovation came in the 1950's, when the Foreign Service Institute (FSI) formalized a procedure and rating scale for oral proficiency interviews (OPI's). Then in the 1980's, the American Council on the Teaching of Foreign Languages (ACTFL) introduced semi-direct oral proficiency interviews (SOPI's) which were administered using tape recorder or computer, and where candidates' performances were recorded for later human scoring. Most recently, in the 1990's, Ordinate Corporation (now part of Pearson) developed fully automated tests which were administered by computer and also scored by computer, using speech recognition and speech processing technology. Of the three approaches, the automated approach would appear to be the most exciting in terms of its potential and future development.

Table 1 - Three innovations spoken language testing (Bernstein et al, 2008)

	OPI	**SOPI**	**Automated**
Administration	human	machine	computer
Scoring	human	human	computer
Year	1950's	1980's	1990's

A number of automated tests have been developed and are currently available globally for a variety of purposes. Table 2 lists the tests available and the high-stakes contexts in which the tests are used, including use by governments, universities, and multinational corporations. Some tests are for general purpose proficiency evaluation and others target a specific domain of use such as university entrance testing (e.g. the PTE Academic) or radiotelephony communication (e.g. the Aviation English test). In addition to those listed in Table 2, Versant tests for spoken proficiency in Chinese and French languages are currently in development.

The correlation coefficients in the middle column of Table 2 reveal the strong relationship between test scores as generated by the machine and test scores as provided by expert human raters. The coefficients are drawn from a set of validation tests; that is, after the test is developed, tests from 100 to 160 test-takers are sent through the automated scoring and are also given to expert human raters to evaluate. Then the two sets of scores, machine and human, are correlated to find the relationship between them. The coefficients are all above r=0.93, showing that machine and human scores and highly comparable (where zero represents no relationship and one represents a perfect relationship).

Table 2 - Automated tests in use and the correlation coefficients between test scores as provided by the machine and test scores as provided by expert human raters.

Automated Test	**Correlation (r) between machine and human scores**	**Primary Users**
Spanish	0.97 (n=100)	US Government including Department of Homeland Security and US Dept of Defense
Dutch	0.93 (n=139)	Dutch Government as part of immigration and naturalization procedure
Arabic	0.98 (n=134)	US Defense Language Institute in the Arabic training program
English	0.97 (n=150)	School districts and universities, companies including AT&T, Dell, IBM, LG, Accenture
Aviation English	0.94 (n=140)	Boeing, Emirates Airlines, Belgian Government, Indian Government, Air Asia
PTE Academic	0.97 (n=158)	Students for university entrance (the test is recognized by 1,000+ institutions worldwide)

The automated tests in various languages are known as Versant tests. Benchmarking exercises linking the test scores to CEFR levels have been conducted for most of these tests. Linking Versant scores to the CEFR bands is established by judging test-taker performance: a panel of expert judges evaluates test-taker performances as they respond to open-ended questions or perform tasks such as listening to stories and retelling them. The CEFR judgments are then aligned with test-taker scores. Experiments have shown (Bernstein and De Jong, 2001) that these tests can reliably separate candidates into CEFR Levels ranging from below A1 to C1 and above. This means, as illustrated in Figure 1, that a student may for example take a Spanish test and receive immediate feedback that they have spoken Spanish equivalent to B2, and then take a Dutch test to find that they are proficient in spoken Dutch at A2, all within the same thirty-minute period.

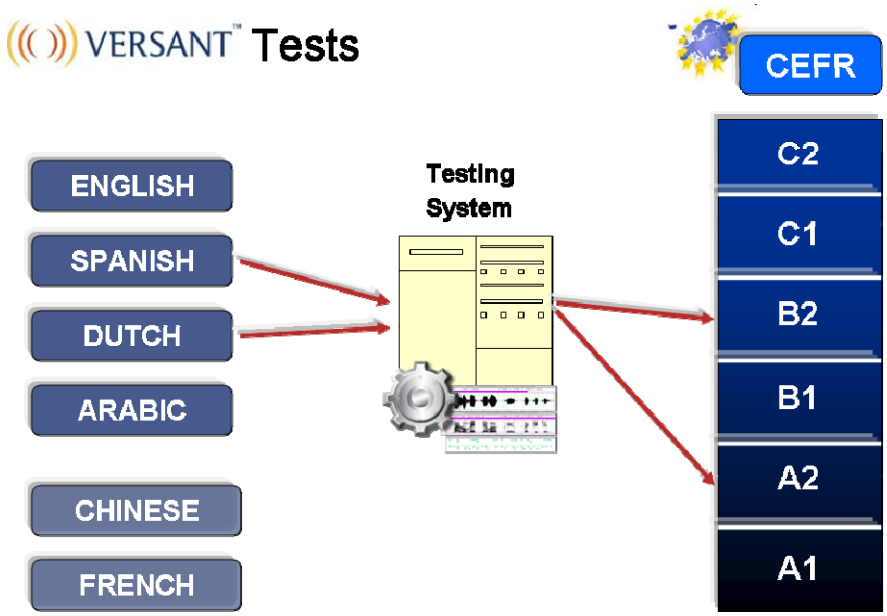


Figure 1 - Illustration of immediate Versant test score feedback on full range of CEFR Levels (Chinese and French tests are due to be released in 2011).

3. Test administration

The administration of the automated test is simple, as follows:

Step 1. Students are given a unique Test Identification Number by their teacher or administrator. They can take the test via telephone or computer; students either dial a local number on their telephones or they login to a platform which is easily downloadable onto any computer.

Step 2. Dialing or login connects the student to remote test delivery servers. The servers select pre-recorded test questions from item banks and present them to the student, one at a time. The student listens and responds.

Step 3. The student's responses are captured as sound files and sent over the internet to scoring servers, which analyse the acoustic signal.

Step 4. The test scores are posted to a secure website. The administrator can track which students have completed the test and download a complete set of scores into a spreadsheet. If the administrator allows it, students can also access their scores online using their Test Identification Numbers.

The entire process usually takes 10-20 minutes from beginning to end, but longer or shorter tests may be constructed depending on the needs of the user. For example, a child's reading test may take only two minutes (Van Moere and Downey, 2010), but a high-stakes test for airplane pilots may take up to 30 minutes (Van Moere et al, 2009). The test items can either be randomly selected on-the-fly from large item banks or presented in fixed-form, depending on the design specifications. The actual speech analysis and scoring ordinarily takes several minutes. Note that the 4-step process described above is appropriate for low-stakes or institutional testing. In high-stakes contexts, such as for the Dutch immigration test (De Jong et al, 2009), the candidates' identity is verified prior to proctored exams being administered in secure test environments.

During the test, instructions are spoken by an examiner voice. The test questions (or prompts) are spoken by numerous different speakers, thereby testing the candidate on a range of different speech speeds and styles. This is in contrast to traditional face-to-face tests in which the test-taker interacts with only one interlocutor. The system is intelligent in that when the candidate finishes responding to one prompt, this is detected and the next prompt is presented.

A test score management system is available to administrators. This management system is accessible via secure website and contains a record not only of test scores, but also a selection of the candidate's responses which are stored as audio files. Depending on the type of test and stakes involved, there may be several test questions which are presented to the test-taker but not scored. The administrator may listen to these responses by simply clicking on an icon next to each candidate's set of test scores.

These audio files may serve as a quick way to check the candidate's identity, but are more commonly used for further research and evaluation by decision-makers, researchers, test developers and other authorized users (see Figure 2). Recorded test responses may also be collected to comprise a language portfolio for candidate's to monitor their own learning. If access is enabled for them, candidates may listen to their selected item performances following each test in order to understand how they sounded and detect any areas for improvement, and also periodically return to the audio files to track their own learning trajectory.

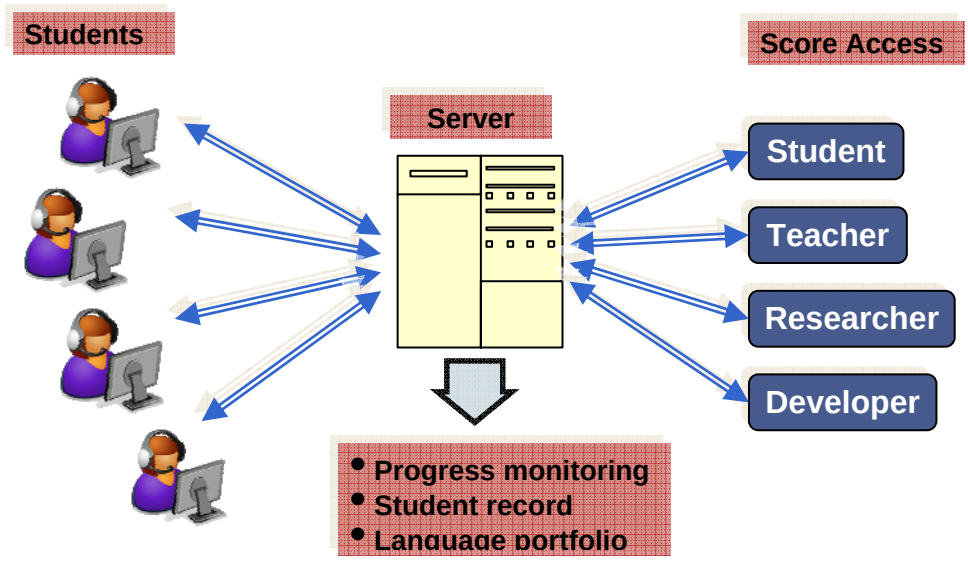


Figure 2 - Illustration of test score management system which also stores audio recording spoken responses to selected test questions.

4. Automated speech and language processing

The goal of the Versant testing system is to provide an accurate and reliable measure of the candidate's spoken language proficiency. Once a candidate has provided a response, two aspects of the speech are evaluated:

1. Content: Did the candidate answer correctly with the expected words in a correct sequence?
2. Manner: Did the candidate respond intelligibly and with an appropriate pace, rhythm and pausing?

The ways in which these measures are extracted can be shown through the following example. Figure 3 shows an analysis of the utterance "My roommate never cleans the kitchen." The first mapping consists of a waveform representation of the spoken sentence; this consists of a time-series plot of the energy of sounds produced by the test taker. The second mapping is a spectrogram that represents the spectral composition the response, where darker shading represents more energy in a range of

frequencies during the production of a phoneme. The third plot presents the words and their component sounds which were actually "understood" by the speech recognizer ("My roommate never cleans the kitchen"). From this, the Versant speech processors can extract a number of features of the spoken responses, such as syllables, stress, clarity of phone production, duration, and hesitations. In a sentence-long utterance such as given in Figures 3 and 4, there are between 30 and 50 individual pieces of information that can be extracted and fed into the scoring models. Over a 15-minute test consisting of 60 responses, over 2,000 pieces of information may be analyzed.

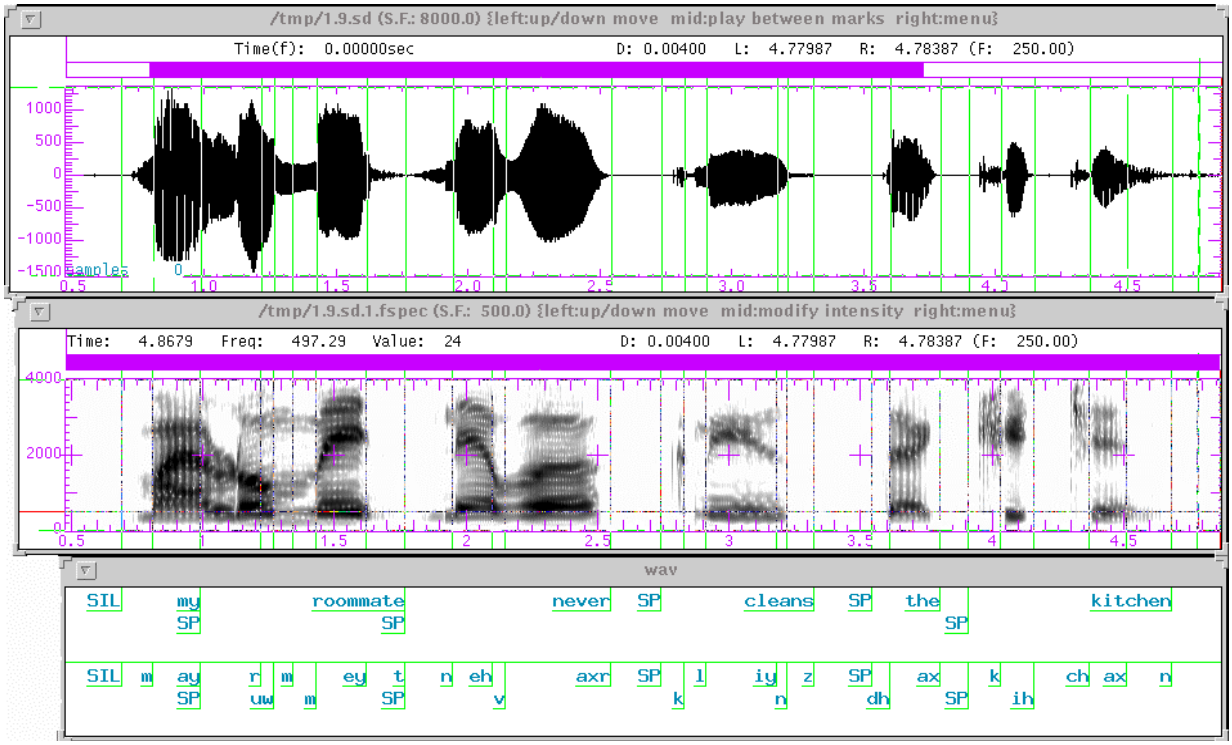


Figure 3 - Waveform, spectrogram, and phone sounds for the utterance "My roommate never cleans the kitchen".

Further exemplification is provided in Figure 4, which reveals how the different features of the response are used to inform different subscores (reflecting different dimensions of language proficiency). This example is drawn from the PTE Academic and shows waveforms of two speakers, one L1 speaker of English and one learner, who have heard the prompt "New York City is famous for its ethnic diversity" and have then attempted to repeat the sentence word-for-word. The test-taker whose L1 is English is able to repeat the sentence verbatim, with natural pace. The learner, however, gives a performance with numerous infelicities, of which four will be explained here.

First, compared to the pace established by a sample of L1 speakers, the learner's response time is unusually slow: 5.5 seconds as opposed to an average of 3.0. This reveals something about the learner's psycholinguistic processing capabilities, namely that he required greater resources than the

L1 speaker in order to decode the message that was heard and then encode the word-string in the response (see Section 5 for a discussion of this). Second, the learner makes several pronunciation errors. For example, the waveform shows that the learner stresses the second syllable in "City" which as an incorrect lexical stress would be reflected in the pronunciation subscore. Another error which the speech processors analysed but which is not visible in the waveform is the learner's mispronunciation of "diversity" which was spoken as "divershity". Third, the learner's utterance exhibits pauses in mid-sentence which were not present in the native speech, or at least were outside of the normal parameters of high-proficiency speech. These unnatural pauses and other duration measures are reflected in the fluency subscore. For example, every L1 speaker in the sample linked the words "City␣is" together, and gave a natural, very short silence (50-100 milliseconds at most) as they transitioned between "is" and "famous" (i.e. "New␣York City␣is famous"). The learner, on the other hand, did not link "City" and "is" but actually gave a long (300 millisecond) pause between the two words (i.e. "New␣York City is famous"), which stands out as being uncharacteristic of proficient sentence processing. Fourth, the learner repeated the word "ethnic" twice. This kind of accuracy error – when learners omit, insert, or substitute other words – also reveals the limits of their psycholinguistic processing. Proficient or native speakers can usually repeat most sentences (up to 9-words long) effortlessly as they have automatized the language and need not pay attention to the linguistic code. Learners, however, have finite cognitive resources and must allocate their resources variously to comprehension, response formulation, linguistic encoding, articulatory encoding and monitoring. Repetitions or insertions in this kind of task, as well as the slow pace of speech mentioned above, are evidence of learners struggling to cope with the demands that are routine when communicating in another language.

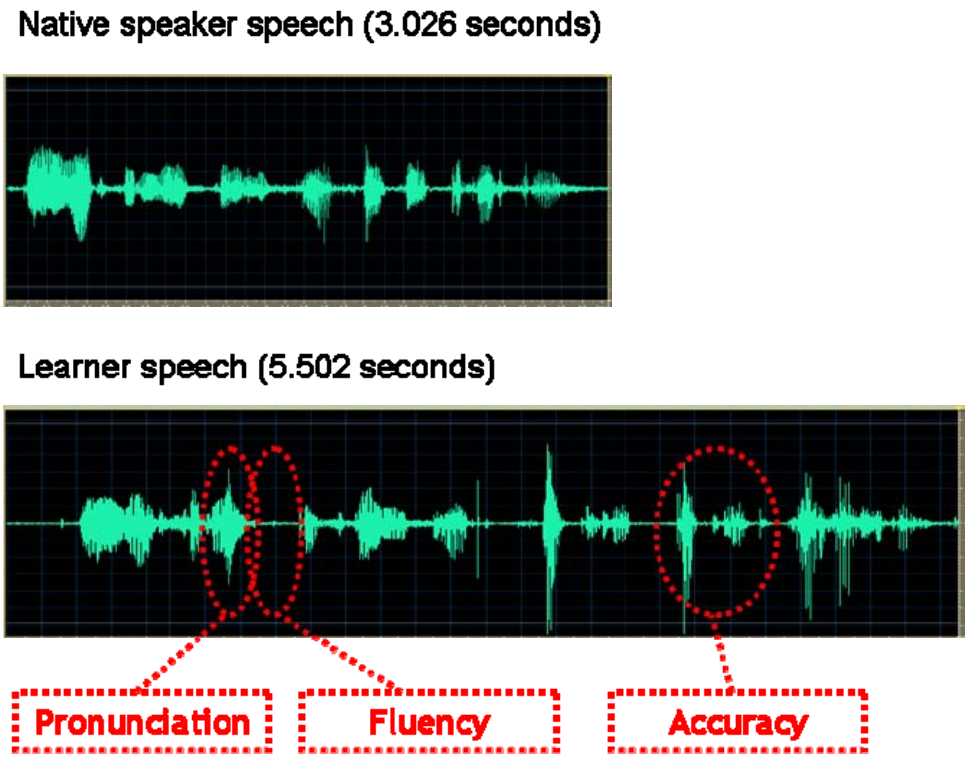


Figure 4 - Comparison of L1 speaker performance and learner performance on a Repeat Sentence task where the prompt was “New York City is famous for its ethnic diversity”

A question that is frequently asked is: how well does the recognizer cope with different accents? The Versant system has been developed and trained to ensure that a wide range of acceptable accents and speech styles are treated equally. The acoustic models for the speech recognizer (models of each sound in the language) have been trained on speech data from a diverse sample of L1 and learner speakers. In this way, the speech recognizer is optimized for many varieties of accented speech patterns. Further, the speech recognizer is also trained on language models that represent not only the correct answers, but also the errors and disfluencies that are common when learners respond to the items. The machine generally recognizes response words as well as or better than a naïve listener, but does not generally recognize as accurately as a trained listener who knows the language content. Since many native speakers were used to model answers, a range of acceptable parameters exist which allow test-takers with naturally varying rates of speech to score highly, as long as expert humans would rate those responses highly.

5. The test construct

The automated approach assesses a well-defined set of linguistic constructs consisting of lexical knowledge, together with knowledge of phonology, morphology and syntax, which all operate in combination with the automaticity with which these types of knowledge can be applied (see below). However, this discussion will first address the two main constraints to the automated approach. First, it

requires contact with a computer rather than a live person, and so provides limited opportunity for communicative and social interaction. Second, the speech recognition operates more accurately when the kinds of responses that the test-takers give are more constrained rather than completely open-ended, and so certain functional interactions are difficult to elicit in an automated test. It is sometimes thought that these two limitations narrow the construct that is assessed. However, to state that automated tests have a narrow construct is an over-simplification and is in some respects misleading. This section will show how the construct of these automated tests underlie spoken communication, and how it offers a psycholinguistic construct which is under-represented by communicative testing approaches.

Spoken language constructs may be divided into two broad categories, as shown in Figure 5. On the left of the figure are the *mechanical* aspects of the language: vocabulary, grammar, pronunciation, fluency. These are the core language skills which make possible communication that is coherent, meaningful, and intelligible. Without these, it would be impossible to give directions to the train station, express needs and wants, hypothesize, or argue an opinion. But if these mechanical aspects are mastered, then all of these communicative acts would be technically possible. On the right of the figure are the *social* aspects of the language, which include but are not limited to: nuance, politeness, turn-taking, negotiation. Competency in the social dimension allows the speaker to use language appropriately with interlocutors of different social standing, obey the conventions of turn-taking, and argue an opinion in such a way as not to cause offence, for example.

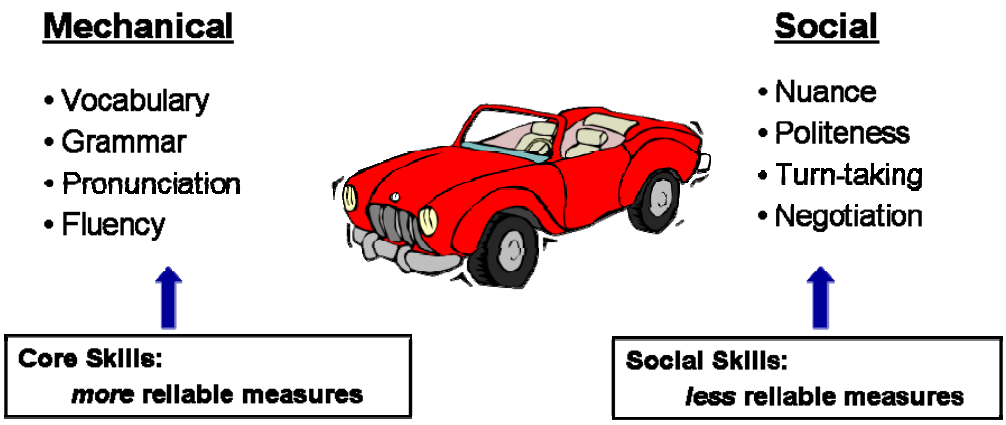


Figure 5 - Oral language skills as seen through the analogy of a car, where mechanical knowledge is likened to the engine of communication and social knowledge is likened to the comfort and design features of the car.

Two important areas in which mechanical and social skills differ are (1) the reliability with which they can be measured in an oral test, and (2) the fairness and standardization with which they can be elicited. To explain these, the analogy of the car in Figure 5 is useful. Using an automated approach, the mechanical language skills can be accurately and consistently measured with known metrics. Mechanical skills are akin to the quantifiable metrics that are used to evaluate cars: horsepower, fuel consumption, physical dimensions, how many seconds it takes to go from zero to sixty miles per hour, and so on. For example, when a Sentence Repeat item is presented to test-takers, that item has already been presented to numerous L1 speakers and learners during field testing. The precise features of the item that make it a unique spoken form are known, its difficulty in relation to all the other items in the test is known, and the kinds of errors that learners make while attempting to respond are already anticipated by the scoring models. This means that, as with the mechanical components of a car, assessment of mechanical language skills can be quantifiable and derived from empirical, data-driven scoring models.

The social language skills, on the other hand, are judgments of a performance using a given criteria. These are fuzzier, less reliable measures which may be considered similar to how the car feels to the driver when taking a corner, the comfort of the interior, whether the design of the exterior is pleasing, and so on. Although social communication is the goal of learning to speak a language, the pitfalls of over-emphasizing the social construct in an oral assessment must be acknowledged. Social skills mediate communication but they are not the building blocks of communication; it is the core skills that make communication possible (and the mechanics that makes the car go). Further, it is very difficult to elicit social skills in a standardized way and score them fairly in a 10 or 15 minute oral performance test. This is because test-takers will exhibit different behaviours under different test conditions (e.g. interlocutors and tasks). Moreover, it is difficult for raters to disambiguate social skills from the core building blocks on which the communication depends, e.g. grammar, vocabulary, and pronunciation.

Leaving the car analogy behind, the automated approach further describes a psycholinguistic construct which is, unfortunately, unaccounted for in traditional interview and human-rated oral proficiency tests, and that is *automaticity*. Bernstein, Van Moere and Cheng (2010) point out that the speed and accuracy of psycholinguistic processing is a reliable predictor of proficiency. It is such a good predictor that we need not assess learners on higher-order cognitive functions and low-frequency lexis in order to establish their proficiency. Rather, we can present test-takers with quite simple tasks involving everyday language structures, and by presenting this language at a conversational pace we can then analyse a test-taker's ability to respond at a conversational pace. L1 speakers and high-proficiency speakers comprehend automatically and respond by easily selecting lexical and language

structures, and articulating them effortlessly and fluidly. On the other hand, learners must pay attention to the linguistic code at every step in the process, from decoding the meaning of words, to formulating a response, to making the sounds of the words. Bernstein et al (2010) show that the constructs thus assessed in automated tests yield score data that correlate with interview tests from beginner through to advanced levels.

This view of testing is compatible with the view of language put forward by Hulstijn (2010), who argues that core language skills which are common to all L1 speakers are linguistic knowledge (vocabulary, grammar, pronunciation) and automaticity (speed of processing). Hulstijn explains this in terms of Basic and Higher Language Cognition. He posits that language proficiency is essentially made up of the phonetic, phonological, morphonological, morphosyntactic, and lexical domains. Organizational and higher-order skills, on the other hand, relate to different constructs that are dependant on intellect, education, professional careers and leisure time activities. According to this theory, native speakers all share the same basic linguistic cognition, but vary markedly in their use of higher linguistic cognition, as shown in Figure 6.

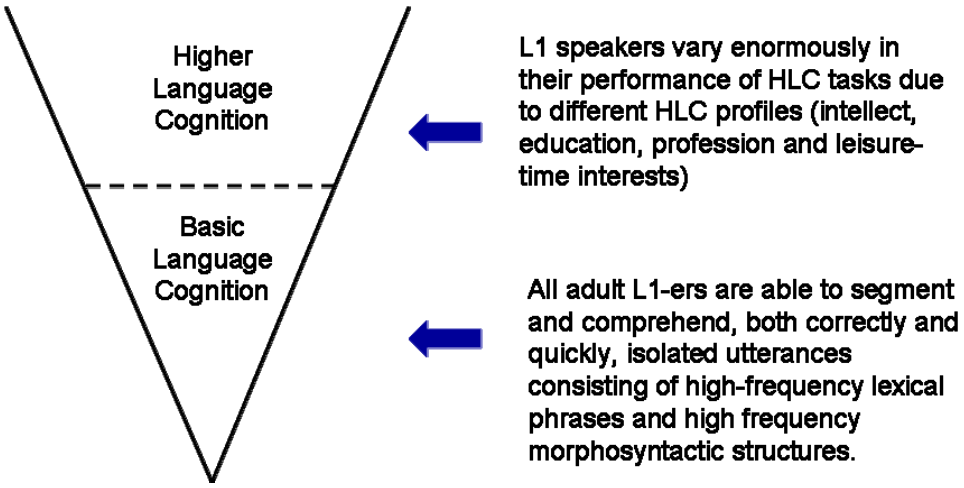


Figure 6 - Basic Language Cognition seen as underlying Higher Language Cognition (adapated from Hulstijn, 2010)

Basic Language Cognition (BLC) encompasses knowledge of high-frequency words and structures that may occur in any communicative situation, *in combination with* the automaticity with which these types of knowledge can be processed (Hulstijn, 2010). Higher Language Cognition (HLC) consists of intellect and experience-driven skills which vary according to the speaker's profile. For example,

Hulstijn argues that many L1 speakers are unable to perform at CEFR Overall Oral Production B2 and above:

> **B2:** Can give clear, systematically developed descriptions and presentations, with appropriate highlighting of significant points, and relevant supporting detail
>
> **C1:** Can give clear, detailed descriptions and presentations on complex subjects, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion.
>
> (Council of Europe, 2004:58)

After acknowledging its limitations, this section has explained the advantages of automated scoring in terms of its construct. The next section proposes a test design for a language survey which maximizes the practicalities of computerized test delivery and accuracy of automated scoring.

6. Test tasks

Surveying spoken language proficiency in L2 and L3 in different countries requires standardized indicators such as given in the descriptors of the CEFR scales, as well as test tasks that are cognitively simple, performance-rich, and practical to administer and score. Sentence-level responses to items with established performance properties are most likely to meet these criteria, and are useful for comparing levels of achievement across languages. Figure 7 gives some of the tasks that might be included in such a test. A multi-task, multi-trait approach ensures that most traits are informed by performance on more than one kind of task. The main features of each task are as follows:

- Describe a picture or graph:

  This task elicits quite long (e.g. 40 second) turns and is beneficial because the input is non-linguistic, thereby ensuring the test-takers use their own words to construct the response.

- Listen to a story and then retell it:

  This provides a good assessment of comprehension and expression, and also allows the test designer control over the language of the input.

- Listen to a question and give a short (one word or phrase) answer:

  The task assesses productive vocabulary and real-time lexical access (for example: “What is frozen water called?). As each item takes just six seconds approximately, a reliable estimate of vocabulary can be achieved in just two or three minutes.

- Listen to a sentence and repeat it, or listen to a sentence with incorrect word order and arrange it correctly:

  These are tests of general speaking proficiency and psycholinguistic competence; they are data-rich and provide highly reliable measures. Although it is sometimes misconstrued as a memory

test, research has convincingly shown that it sentence repetition is a measure of language skills and involves understanding the sentence and then reconstructing it (Vinther, 2002)

- Read a sentence or passage aloud:

  The task provides measures in pronunciation, fluency, and expressive reading. It is also useful as it is comparatively easy to establish national norms and compare performance across different languages with measures such as reading rate.

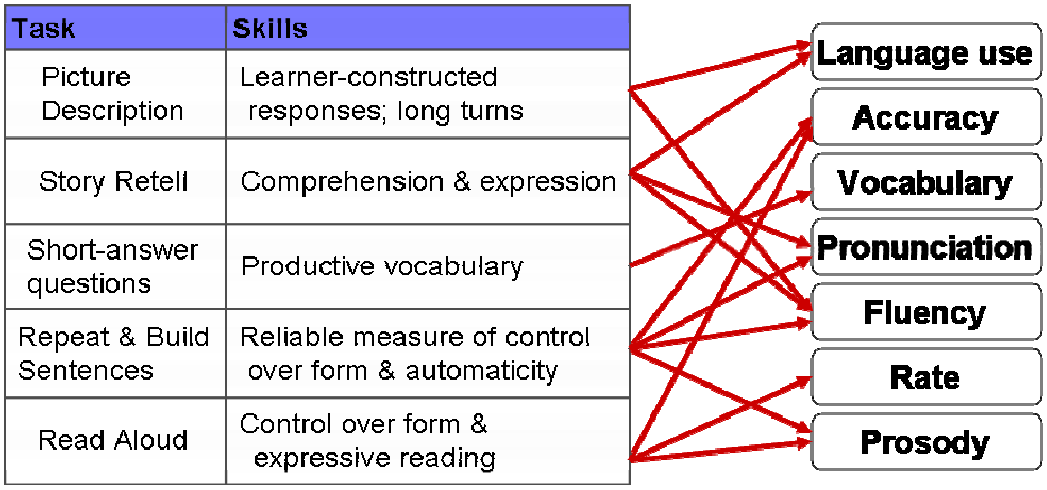


Figure 7 - A sample test design and subscores that could be produced when adopting an automated approach

A good test design such as in Figure 7 includes a balance of tasks for eliciting various speech acts, including long and short turns, thereby providing face validity and positive washback. Face validity can be further enhanced by including extra tasks in the test which need not be scored by machine. For example, questions which elicit functions such as hypothesizing or giving opinions may be added to the test, and the test-takers' responses can be stored for further research or for the language portfolio mentioned in Figure 2.

This paper has shown why automated spoken language testing is valid for certain assessment contexts, with the important practical advantage that the tests can be administered by telephone or computer on demand. Because such testing does away with the logistical problems and costs of training and retaining human examiners and scorers, automated testing is resource-efficient, while being remarkably consistent across time and location. Moreover, the approach is at least as reliable as the best human-scored tests and has a well-defined construct for performance in listening and speaking. For language surveys, automated tests offer the capability of analyzing every utterance to fine-grained

detail in order to produce a rich set of performance-based measures that are linked to the CEFR, but provide so much more detail than a CEFR level alone. Although the technology is not yet capable of assessing social skills or nuance of meaning, it does assess core language skills such as knowledge of grammar and lexis coupled with automaticity. These are the skills which underlie social interaction and are generally predictive of success in communication.

Acknowledgement and disclaimer:

The author gratefully acknowledges Jared Bernstein and Brent Townsend for developing the technology and shaping the theory behind the automated tests described in this paper. The author works for Pearson, the provider of the Versant automated tests.

References

Balogh, J. & Bernstein, J. (2006). Workable models of standard performance in English and Spanish. In Y. Matsumoto, D.Y. Oshima, O.R. Robinson, & P. Sells (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., Pado, U., Suzuki, M., & Van Moere, A. (2008). Developing and Validating an Automated Test of Spoken Modern Standard Arabic. Paper presented at the East Coast Organization of Language Testers (ECOLT), November 8, 2009: Washington DC.

Bernstein, J. & De Jong, J. H.A.L. (2001). An experiment in predicting proficiency within the Common Europe Framework Level Descriptors. In Y.N. Leung et al. (Eds.), *Selected Papers from the Tenth International Symposium on English Teaching* (pp. 8-14). Taipei, ROC: The Crane Publishing.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27(3), 355-378.

Council of Europe (2004). A Common European Framework of Reference for Languages: Learning, teaching, assessment. Available online at http://www.coe.int/T/DG4/Portfolio/documents/Framework_EN.pdf (accessed August 2010)

De Jong, J. H. A. L., Lennig, M., Kerkhoff, A., Poelmans, P (2009). Development of a Test of Spoken Dutch for Prospective Immigrants. *Language Assessment Quarterly*, 6(1), 41-60.

Hulstijn, J. (2010). Harmony or conflict? Language activities, linguistic competencies, and intellectual functioning in the Common European Framework of Reference for Languages. Paper presented at the European Association of Language Testing and Assessment (EALTA), May 29, 2010: The Hague.

Pearson (2009a). *Versant English Test: Test description and validation summary*. Pearson Knowledge Technologies, Palo Alto, California. Available online at http://www.ordinate.com/technology/VersantEnglishTestValidation.pdf (accessed August 2010).

Pearson (2009b) *Versant Spanish Test: Test description and validation summary*. Pearson Knowledge Technologies, Palo Alto, California. Available online at http://www.ordinate.com/technology/VersantSpanishTestValidation.pdf (accessed August 2010).

Van Moere, A. and Downey, R. (2010). Usability testing with stakeholders in the development of a formative assessment of oral reading fluency. Paper presented at the European Association of Language Testing and Assessment (EALTA), May 29, 2010: The Hague.

Van Moere, A., Suzuki, M., Downey, R., and Chang, J. (2009).Implementing ICAO Language Proficiency Requirements in the Versant Aviation English Test. *Australian Review of Applied Linguistics*, 32 (3).

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

The author:
Alistair Van Moere
Knowledge Technologies
299 S. California Avenue, suite 300
Palo Alto, CA 94306
E-mail: avanmoere@pearsonkt.com

Alistair Van Moere is Director for Test Development at Pearson Knowledge Technologies, where he oversees numerous automated language tests used globally in high-stakes contexts. He has a Ph.D. in Applied Linguistics from Lancaster University and 15 years experience working in English language education and assessments.

COMPUTER-BASED TESTING OF SECOND LANGUAGE PRAGMATICS

Carsten Roever
Linguistics & Applied Linguistics, The University of Melbourne

Abstract

Computer-based testing of pragmatics is a relatively small area but holds a great deal of promise. Some tests already exist that deliver items via the computer but all current productive testing instruments in interlanguage pragmatics underrepresent the construct of pragmatic ability and need to integrate extended monologic and dialogic discourse. This is a difficult task for large-scale survey tests because they must balance adequate construct representation with optimum use of resources. Computer-based testing already offers useful options for item delivery and rating but has not been used for automated scoring of pragmatic performance. Some item types will be suggested that might allow scoring via an automated speech recognition engine.

Testing of second language (L2) pragmatics is a relatively recent development, and computer-based assessment is a fairly small part of this growing research area. In this paper, I will first give an overview of pragmatics tests that have been developed to date with a particular focus on computer-based tests. I will then outline some possibilities for large-scale computer-based survey testing of second language pragmatics.

What is pragmatics?

Pragmatics is generally viewed as the study of how the external social context of language use affects what speakers say and how hearers interpret it (Leech, 1983; Levinson, 1983). Language users adapt their use of language to a range of facets of the communicative situation, including the interlocutor, the physical setting, the large speech event, the channel of communication and others (Hymes, 1972; Coulmas, 1981). To take a simple example, people speak differently in a job interview then they would while chatting with friends in a pub.

Competent language users know the rules and norms of the target speech community ("sociopragmatics"; Leech, 1983) and have the linguistic tools to display and implement their knowledge for communicative purposes ("pragmalinguistics"; Leech, 1983). For example, many European languages make a difference between a close form (tu / du) and a distant form (vous / Sie). A

pragmatically competent language user knows whether the target culture would characterize his or her relationship with an interlocutor as close or distant, and also knows what form to use to address the interlocutor accordingly.

Pragmatic competence also includes knowledge of how to imbue a message with meaning beyond its basic propositional content. For example, the tone of voice in a statement like “oh great” makes all the difference as to how it will be interpreted. Speakers can convey to the interlocutor a wide range of displays of self that involve an intricate interplay between pragmalinguistic and sociopragmatic knowledge, i.e., they can convey that they are angry, respectful, enthusiastic, dignified, friendly etc. Such displays depend crucially on language but also on facial expression, prosody, gesture, and a range of other cues that allow the interlocutor to interpret what the speaker means, and which have become known as contextualization cues (Gumperz, 1982).

A central area of research that brings together sociopragmatic and pragmalinguistic knowledge is the use of polite language. Competent language users know how to produce speech acts (Austin, 1962; Searle, 1968, 1975) such as requests, apologies, refusals, suggestions, compliments etc. in a way that minimizes the risk of face loss and avoids rupturing the relationship with the interlocutor (Brown & Levinson, 1987). They do so by means of politeness, which allows them to soften requests (“I was wondering if you could…”), make advice less aggressive (“Maybe you could do it that way”) or tone down refusals (“I’d really like to be help but I’m on my way out, you see…”). In a highly influential book, Brown and Levinson (1987) identified three context factors that they claim influence the degree of politeness of a speech act: relative Power, Social Distance, and Degree of Imposition 1. Relative power concerns the power differential between the interlocutors and is equal when neither interlocutor has more power than the other (friends, housemates, family members) but unequal when one interlocutor is more powerful (judge-accused, boss-employee, professor-student). Brown and Levinson emphasize that Power depends on the situation, not the absolute social rank: a student asking a professor for an extension on an assignment is less powerful than the professor but a student sitting on a hiring committee and interviewing a professor for a position is more powerful than the professor in that situation. Social distance concerns the degree of acquaintanceship or the interlocutors' membership in the same social group. For example, housemates, friends, work colleagues or a supervisor with whom you work closely have low social distance settings but a random stranger on the street, a new customer or a fellow employee with whom you have never had contact before have high

1 The names of the context factors are conventionally spelled with capital letters when they refer to Brown and Levinson's concepts rather than just generic notions.

social distance settings. Just like Power, Social Distance spans a continuum and is situation dependent. Interlocutors who might feel that they are "in the same boat" due to shared social background variables (same gender, same age group, same social class, same nationality) generally have lower social distance even as strangers than interlocutors between whom there are differences in these variables. Finally, imposition refers to the "cost" in time, money or inconvenience for the interlocutor of doing what the speaker requests, or the "damage" caused by an action for which the speaker is apologizing. For example, asking to borrow a pen from a sales clerk to sign a credit card slip is a very low imposition request but asking to borrow someone's house for a raucous all-weekend party is a high imposition request. Again, imposition is not absolute and depends on the context: asking a professor for an extension might be a low imposition request if the professor has previously indicated that she is quite happy to grant extensions. However, it might be high imposition if the professor has made it clear that she abhors giving extensions and almost never grants them.

Given the amount of social and linguistic knowledge required for adequate pragmatic performance, it is easy to see how speakers from two different cultures could talk at cross purposes (“crosstalk” in Gumperz, 1982) and misunderstand each other. Different assumptions about social rules and norms, different analyses of what constitutes high or low Power, Distance or Imposition, or less-than-perfect ability to show politeness or respect can lead to offense and breakdowns in communication. For example, some cultures have more direct norms for making requests of intimates (family members or close friends) than others, where even such requests need to be softened by politeness. Interlocutors from such contradictory cultures may constantly feel offended by each other: one because s/he thinks that the other is bossy and rude, the other because s/he feels that her friend is keeping him/her at a distance by constantly being polite and treating him/her like a stranger.

In a somewhat different vein, misunderstandings can also be caused by different ways of structuring discourse. For example, Young (1994) has shown that Chinese and English speakers place the main point of an extended discourse contribution differently: in English, the main point is put first and then supported with reasons or examples. In Chinese, it is withheld until the end of the contribution and reasons or examples precede it and lead to it. Unexpected discourse structure can cause confusion for the listeners who may take the first statement as the main point if s/he is used to English discourse structure and then feel lost in what appears to be a series of unrelated arguments. Similarly, Holmes (2003) gives examples of differences in Anglo-Maori story telling, where the Anglo listeners expected an explicit closing (coda) to the story but Maori speakers did not provide such explicit closings, leading to confusion on the part of the listener.

These are just some examples of misunderstandings that can occur due to differences in cultural norms and pragmalinguistic abilities. Language users are generally unaware of these underlying rule systems but keenly feel their violation as “improper” or “rude” behavior or an indication that the interlocutor is an ineffective communicator. Such impressions can lead to serious interpersonal conflict, and it is therefore important to assess what learners know about the pragmatics of the target speech community so that teaching interventions or consciousness raising activities can be designed that help them avoid pragmatic pitfalls.

Tests of L2 pragmatics

Even though pragmatic competence is part of the major models of communicative competence (Bachman, 1990; Bachman & Palmer, 1996; Canale & Swain, 1980), the first test battery of second language pragmatics was not developed until 1995 when Hudson, Detmer and Brown (1995) pioneered a multi-method battery for Japanese learners of English. They focused on learners’ sociopragmatic ability (Leech, 1983) to produce and recognize situationally appropriate and polite language following Brown and Levinson’s (1987) framework. As their instruments, Hudson et al. employed three different types of discourse completion test (oral, written and multiple choice), role plays, and two types of self-assessment. Discourse completion tests (DCTs) are a frequently used instrument in pragmatics research (Kasper, 2000) and usually consist of two parts: a situation description, which explains the relationship with the imaginary interlocutor and the context of communication, and a gap for the test taker's response. Figure 1 shows an example of a DCT item:

Figure 1: DCT item

You are traveling overseas and arranged for a friend to take you to the airport, which is 45 minutes away, but at the last minute, she called and cancelled. Your only option is to ask your housemate Jack to give you a lift. Jack is in the living room reading the paper.

You say: _____

Hudson et al. also used multiple-choice DCT items, which contained three response options. The test elicited requests, apologies and refusals, and varied Power, Distance, and Degree of Imposition. Three raters rated test taker performance with regard to use of the correct speech act, idiomaticity of formulaic expressions, amount of information provided, and degrees of formality, directness and politeness. Hudson et al. piloted their instrument with a small, homogenous group of Japanese learners

of English. Hudson (2001) reports acceptable rating reliabilities ranging from .75 to .86 for the productive sections.

Hudson et al.'s test was used by Yoshitake (1997) with Japanese EFL learners and adapted for Japanese as a second language by Yamashita (1996) and for Korean as a foreign language by Ahn (2005). Brown (2001, 2008) shows that most components of the test battery function reliably except the multiple-choice DCT, which had disappointingly low reliabilities in Yamashita's (1996) and Yoshitake's (1997) studies. This was seen as unfortunate as the multiple-choice DCT is the most practical (i.e., least resource-intensive) part of the battery, and the only one that could conceivably be machine scored.

In response to this keenly felt lack of a practical test of L2 pragmatics, Liu (2006) claims to have developed a reliable multiple-choice DCT. He developed his instrument bottom-up, collecting productive DCT data and generating distracters from inappropriate non-native speaker responses while employing native-speaker responses for the correct response options. However, his instrument was limited to Mandarin-speaking learners of English, and McNamara and Roever (2006) raise concerns about the validity of inferences which might be more appropriately related to recognition of idiomaticity than general pragmatic competence.

The second large-scale test development and validation project was Roever's (2001, 2005, 2006) web-based test for learners of English. Roever focused on pragmalinguistic knowledge and assessed learners' ability to interpret implicature following Bouton (1988, 1994, 1999), recognize routine formulae and produce the speech acts request, apology and refusal. The implicature and routines sections consisted of 12 four-option multiple-choice items each, and the speech act section of 12 brief response items. Roever's test had high overall reliability and the implicature and speech act sections distinguished well between learners at different proficiency levels whereas the routines section was more sensitive to exposure differences (Roever, 2010b). The test also had high practicality due to its web-delivered format, which allowed random ordering of items and sections, self-scoring of the multiple-choice sections with immediate post-test delivery of scores, and easy rating of the speech act section. Unlike Hudson, Detmer and Brown's (1995) or Liu's (2006) test, Roever's instrument was not designed for learners with a specific native language and Roever (2007, in press) used differential item analysis (DIF) to show that most items did not advantage test takers of a specific first language background.

Based on Roever's (2005) test, Roever, Elder, Harding, Knoch, McNamara, Ryan, and Wigglesworth (2009) developed and piloted an implicature and speech act section for a larger testing instrument. This development, together with Roever's (2010a) implicature section for a university screening test, illustrate the only occasions where pragmatics testing instruments were deployed as part of a larger proficiency battery.

Besides the major assessment projects undertaken by Hudson, Detmer and Brown (1995) and Roever (2005), some smaller projects exist and a number of testing instruments have been developed for research purposes. Tada (2005) developed a computer-based test using video prompts to contextualize oral and multiple-choice DCT items to assess EFL learners' knowledge of the speech acts request and apology. He obtained satisfactory reliabilities for both item types. Walters (2004, 2007, 2009) based his instrument on conversation analysis and used a listening test, role play and DCT to assess ESL learners' comprehension and production of responses to compliments, assessments, and pre-sequences. Possibly due to the highly homogenous group of test takers, his instrument had very low reliabilities but he achieved acceptable inter-rater reliability for the role play.

Some tests were also specifically developed as part of acquisitionally oriented research projects. Most recently, Takimoto (2009) used role plays, DCTs, and an aural and a written appropriateness judgment task to ascertain the success of a teaching intervention and obtained high reliabilities in the .9 region. The written judgment task was computer-delivered. Other speech act oriented projects included Bardovi-Harlig and Dörnyei's (1998) investigation of pragmatic awareness in second and foreign language learning settings, for which they developed a video-prompt based judgment task of pragmatic appropriateness and severity. Niezgoda and Roever (2001) and Schauer (2006) used the same instrument with different populations. Matsumura (2001, 2003) deployed a multiple-choice DCT to trace learning of sociopragmatic norms for giving advice by Japanese study-abroad students compared to non-study abroad students. Cook (2001) employed a listening comprehension task to investigate Japanese as a foreign language learners' ability to distinguish between situationally appropriate and inappropriate speech styles in Japanese. Finally, in the area of research into comprehension of implicature, Bouton (1988, 1994, 1999) developed a multiple-choice test of implicature, and Taguchi (2005, 2007, 2008) undertook a series of studies using computer-based tests to investigate correctness and reaction time in learners' comprehension of L2 implicature.

Validity, practicality and computer-based testing of pragmatics

Validity is a central consideration in the assessment of test quality (AERA, 1999), and concerns the degree to which defensible inferences can be drawn from test scores (Messick, 1989). In all major

models of validity (Kane, 1992, 2001; Messick, 1989; Mislevy, Steinberg & Almond, 2003) extrapolation of scores to the non-test environment is the crucial and most complex step because it essentially involves a prediction of how test takers will perform in real-world settings. The more comprehensive the construct under assessment, the more sweeping the predictions that can be made about non-test performance, and the more informative the test results are for test users. However, productive tests of L2 pragmatics have so far only examined a very limited part of the overall construct of pragmatic competence.

The majority of tests developed so far, and all productive instruments, have been developed within the theoretical framework of speech act theory (Austin, 1962; Searle, 1969, 1975) and politeness theory (Brown & Levinson, 1987) with an analytical approach usually following the categories developed in the cross-cultural speech act realization project (CCSARP) (Blum-Kulka, House & Kasper, 1989). The construct under assessment is learners' ability to use language with a sufficient degree of politeness given a situational context, which is usually defined by the power differential and social distance (degree of acquaintanceship and commonality) with the interlocutor and the degree of imposition of the speech act, nearly always a request, apology or refusal (Kasper & Rose, 2002).

Besides its overriding focus on politeness and the small range of specific speech acts investigated, the first generation of pragmatics tests also suffered from a strong tendency to atomize individual speech acts and assess them outside a conversational context. This was aided and abetted by the use of DCTs, which by their very nature can only elicit a single utterance (for rare exceptions of sequential DCTs, see Kuha, 1999; Chiu, Liou & Yeh, 2007). It was therefore impossible to assess learners' ability to participate in extended interaction, and it was equally impossible to take the co-constructed nature of interaction into account. Finally, and possibly most damningly, DCTs produce inauthentic data, overrepresenting pragmatic strategies that occur rarely or not at all in reality while underrepresenting some that occur frequently (Golato, 2003).

It is therefore important to assess the construct of L2 pragmatic competence more broadly with instruments that go beyond assessing just the politeness of isolated speech acts. The most important innovation to achieve broader construct coverage in L2 pragmatics assessment would be the integration of extended discourse in social settings, involving both monologic and dialogic discourse. Research has shown that producing both types of discourse can be problematic for second language users and requires learning and development just as much as other parts of learners' second language system (Al-Gahtani & Roever, 2010; Cook, 2001; Hassall, 2001; Ishida, 2009; Kim, 2009; Young, 1994).

However, practicality is a constraint on broadening the construct in the setting of large-scale computer-based survey testing. How can tests that offer a more comprehensive construct representation be designed so that they are minimally resource intensive? This is especially difficult as dialogic interaction is generally tested with a human interlocutor, e.g., in the role plays used by Hudson et al. (1995), the Oral Proficiency Interview (Breiner-Sanders, Lowe, Miles, & Swender, 2000), or the speaking components of IELTS. Needless to say, interaction also needs a human scorer, exacerbating the resources issue. In the following, I will outline some ideas for computer-based instruments that might enable large-scale testing under a comprehensive construct of pragmatic competence.

Computer-based testing can refer to some or all of a variety of features: computer-based delivery of test tasks, computer-based capture of test-taker responses, or computer-based scoring of responses. For tests of pragmatics that involve dialogic interaction, computer delivery can also include producing a computer-generated response to test-taker input, spoken or written. Computer-based delivery of a pragmatics test and capture of responses has been done for productive items by Roever (2005, 2006), though only in writing, Roever et al. (2009), with voice recording, Tada (2005), with video support and voice recording, and Takimoto (2009) for acceptability judgments. No test so far has scored productive test taker responses automatically but a couple of tests have used computer-generated responses to test taker input (Chiu, Liou & Yeh, 2007; Kuha, 1999), though the input was not spoken.

The ultimate goal for a computer-based test of L2 pragmatics would be a fully integrated system that delivers items supported by sound or video, captures spoken test taker responses, analyses and scores the responses, and generates an adequate next-turn response in a test involving dialogic interaction. Needless to say, analysis of spoken test taker responses and generation of a tailor-made next turn are not feasible at this time because they require highly sophisticated automatic speech recognition, which has not yet been developed for high-entropy, low-predictability tasks like free conversation (Zechner, Higgins, Xi, & Williamson, 2009).

At present and in the near future, computer-based delivery may not lead to a revolution in pragmatics testing but it is an area that can be refined and the logistical advantages of computers can be exploited for this purpose. For monologic tasks, the computer can be used to establish the social context in which the monolog is to be delivered, e.g., leaving a voicemail for your boss, thanking a committee for an honor you have received, making a sales pitch to a customer or giving a self-introduction as part of a job interview. A video clip of an interview panel inviting the test taker to introduce himself/herself would go a long way towards establishing adequate situational context, and a job interview via the web is not a far-fetched scenario nowadays. Another option is to embed monologic sections in a longer

conversation ("pseudo-interaction"), similar to the Australian TV show "Thank god you're here", where a participant is prompted to produce a number of longer contributions in an extended interaction through questions like "what do you think about this?", or "can you explain this to them?". While any language produced and recorded by the test taker would need to be scored by human raters, scoring could at least be facilitated by delivering speech samples to raters online. Scorers could assess features like comprehensible and target-like discourse organization, situationally appropriate speech style, and adequate and comprehensible use of contextualization cues to convey a message and display the self effectively.

There are also some task types that might be low-entropy and predictable enough to lend themselves to computer-based scoring. One possibility is productive tasks eliciting routine formulae and other routinized expressions. Roever (2005) tested routine formulae by means of recognition but given the highly constrained nature of possible responses, it might be possible to train an automatic speech recognition engine to identify whether the formula produced matches the target. The contextualization could be achieved through video-based scenarios that depict the situational context and contain interaction but omit the focal utterance which participants are then asked to supply. Routinized expressions are short, so a fairly large item set could be tested, and research indicates that they are strongly dependent on exposure to the target language community rather than general proficiency (House, 1996; Roever, 2005, 2010b) so they cover an aspect of pragmatic competence that is not already subsumed by the usual proficiency constructs assessed in most tests.

Another possible approach is to try a pragmatic cloze, i.e., a conversation with gaps. This might be particularly feasible for routine institutional talk that runs a highly predictable course, e.g., booking a table at a restaurant, making small talk at a cocktail party, or clearing customs. The test taker could watch the whole interaction with one party's contributions deleted, and then be shown a transcript where they insert the missing turns by speaking them into a microphone. They can also later be allowed to listen to the conversation with their contributions inserted and make changes if desired. The range of possible test taker utterances is constrained through the following turn, for which the utterance must be appropriately recipient designed. Given the small universe of possible responses, training a speech recognition engine might be feasible for this item type.

Needless to say, automatic scoring of even short, highly constrained utterances requires a great deal of research but it might be the most feasible option for large-scale use of pragmatic items in survey instruments.

Conclusion

Testing of second language pragmatics is a growing area that adds an important component to measurement of second language proficiency. Previous work has focused on the measurement of learners' appropriate and polite use of speech acts but future assessment work should go beyond that and expand the construct. This is a challenging task to implement in large-scale survey assessments where practicality is an overriding concern. However, computer-based tasks can be used to facilitate elicitation and scoring of performance data, and it may be possible to develop some highly constrained, low-entropy task types that allow automatic scoring of spoken test taker production. A great deal of research is needed to enable the design of a broad yet practical test of second language pragmatics.

References

American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Ahn, R.C. (2005) *Five measures of interlanguage pragmatics in KFL (Korean as a foreign language) learners*. Unpublished PhD thesis, University of Hawaii at Manoa.

Al-Gahtani, S. & Roever, C. (2010). *Role-playing L2 requests: proficiency and discursive development.* Unpublished manuscript, The University of Melbourne.

Austin, J.L. (1962). *How to do things with words.* Oxford: Oxford University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L.F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bardovi-Harlig, K. & Dörnyei, Z. (1998). Do language learners recognize pragmatic violations? Pragmatic vs. grammatical awareness in instructed L2 learning. *TESOL Quarterly, 32*, 233-259.

Blum-Kulka, S., House, J. & Kasper, G. (Eds.) (1989). *Cross-cultural pragmatics: requests and apologies*. Norwood, N.J.: Ablex.

Bouton, L. (1988). A cross-cultural study of ability to interpret implicatures in English. *World Englishes, 17,* 183-196.

Bouton, L.F. (1994). Conversational implicature in the second language: Learned slowly when not deliberately taught. *Journal of Pragmatics, 22,* 157-167.

Bouton. L.F. (1999). Developing non-native speaker skills in interpreting conversational implicatures in English: explicit teaching can ease the process. In E. Hinkel (Ed.), *Culture in second language teaching and learning* (pp. 47-70). Cambridge: Cambridge University Press.

Breiner-Sanders, K.E., Lowe, P., Miles, J. & Swender, E. (2000). ACTFL proficiency guidelines speaking. Revised 1999. *Foreign Language Annals, 33* (1), 13-18.

Brown, J. D. (2001). Six types of pragmatics tests in two different contexts. In K. Rose, & G. Kasper (Eds.) *Pragmatics in language teaching* (pp. 301-325). New York: Cambridge University Press.

Brown, J. D. (2008). Raters, functions, item types and the dependability of L2 pragmatics tests. In E. Alcón Soler, & A. Martínez-Flor (Eds.), *Investigating pragmatics in foreign language learning, teaching and testing* (pp. 224-248). Clevedon: Multilingual Matters.

Brown, P., & Levinson, S. D. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguisitics, 1,* 1-47.

Chiu, T. L., Liou, H.C., & Yeh, Y. (2007). A study of web-based oral activities enhanced by automatic speech recognition for EFL college learning. *Computer Assisted Language Learning, 20,* 3, 209-233.

Cook, H. M. (2001). Why can't learners of JFL distinguish polite from impolite speech styles? In K. Rose, & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 80-102). Cambridge: Cambridge University Press.

Coulmas, F. (1981). Introduction: Conversational Routine. In F. Coulmas (Ed.), *Conversational Routine* (pp. 1-20). The Hague: Mouton.

Golato, A. (2003). Studying compliment responses: A comparison of DCTs and recordings of naturally occurring talk. *Applied Linguistics*, *24*, 1, 90-121.

Gumperz, J. (1982). *Discourse strategies.* Cambridge: Cambridge University Press.

Hassall, T. (2001). Modifying requests in a second language. *IRAL,* 39 (4), 259-283.

Holmes, J. (2003). 'I couldn't follow her story…': Ethnic differences in New Zealand narratives. In J. House, G. Kasper, & S. Ross (Eds.), *Misunderstanding in social life* (pp. 173-198). Harlow: Pearson Education.

House, J. (1996). Developing pragmatic fluency in English as a foreign language: Routines and metapragmatic awareness. *Studies in Second Language Acquisition*, *18*, 225-252.

Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics* (Technical Report #7). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Hymes, D. (1972) On communicative competence. In J. Pride, & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269-293). Harmondsworth: Penguin.

Ishida, M. (2009). Development of interactional competence: Changes in the use of ne during Japanese study abroad. In H. thi Nguyen & G. Kasper (Eds.), *Talk-in-interaction: multilingual perspectives* (pp. 351-385). Honolulu, HI: National Foreign Language Resource Center, University of Hawai'i.

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

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kasper, G. (2000). Data collection in pragmatics research. In H. Spencer-Oatey (Ed.), *Culturally speaking* (pp. 316-341). London: Continuum.

Kasper, G., & Rose, K. R. (2002), *Pragmatic Development in a Second Language.* Oxford: Basil Blackwell.

Kim, Younhee, 2009. The Korean discourse markers –untey and kuntey in native-nonnative conversation: an acquisitional perspective. In: Nguyen, H., & Kasper, G. (Eds.), *Talk-in-interaction: multilingual perspectives* (pp. 317–350). Honolulu, HI: National Foreign Language Resource Center.

Kuha, M. (1999). *The influence of interaction and instructions on speech act data*. Unpublished doctoral dissertation, Indiana University.

Leech, G. (1983). *Principles of pragmatics.* London: Longman.

Levinson, S. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Liu, J. (2006). *Measuring interlanguage pragmatic knowledge of EFL learners*. Frankfurt: Peter Lang.

Matsumura, S. (2001). Learning the rules for offering advice: A quantitative approach to second language socialization. *Language Learning, 51*(4), 635-679.

Matsumura, S. (2003). Modelling the relationships among interlanguage pragmatic development, L2 proficiency, and exposure to L2. *Applied Linguistics, 24*(4), 465-491.

McNamara, T.F., & Roever, C. (2006). *Language Testing: The social dimension.* Oxford: Basil Blackwell.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). New York: American Council on Education & Macmillan.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of assessment arguments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Nguyen, H.T. (2008). Sequence organization as local and longitudinal achievement. *Text & Talk, 28*(4), 501-528.

Niezgoda, K., & Roever, C. (2001). Grammatical and pragmatic awareness: A function of the learning environment? In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 63-79). Cambridge: Cambridge University Press.

Roever (in press). *Effects of native language in a test of ESL pragmatics*: *A DIF approach*. In G.Kasper (Ed.), Pragmatics & Language Learning, Vol. 12. Honolulu, HI: National Foreign Language Resource Center.

Roever, C. (2005). *Testing ESL pragmatics*. Frankfurt: Peter Lang.

Roever, C. (2006). Validation of a web-based test of ESL pragmalinguistics. *Language Testing, 23* (2), 229-256.

Roever, C. (2007). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly, 4*, 2, 165-189.

Roever, C. (2010a). *Testing implicature under operational conditions*. Unpublished manuscript, The University of Melbourne.

Roever, C. (2010b). *What learners get for free: learning of routine formulae in ESL and EFL environments*. Unpublished manuscript, The University of Melbourne.

Roever, C., Elder, C., Harding, L.W., Knoch, U., McNamara, T.F., Ryan, K., & Wigglesworth, G. (2009). *Social language tasks: speech acts and implicatures*. Unpublished manuscript, University of Melbourne.

Schauer, G. A. (2006). Pragmatic awareness in ESL and EFL contexts: Contrast and development. *Language Learning, 56*(2), 269-318.

Searle, J. (1969). *Speech acts*. Cambridge, UK: Cambridge University Press.

Searle, J. (1975). Indirect speech acts. In P. Cole, & J.L. Morgan (eds.), *Syntax and Semantics, 3: Speech Acts* (pp. 59–82). New York: Academic Press.

Tada, M. (2005). *Assessment of EFL pragmatic production and perception using video prompts.* Unpublished doctoral dissertation, Temple University.

Taguchi, N. (2005). Comprehending implied meaning in English as a foreign language. *The Modern Language Journal, 89*(4), 543-562.

Taguchi, N. (2007). Development of speed and accuracy in pragmatic comprehension in English as a foreign language. *TESOL Quarterly, 41*(2), 313-338.

Taguchi, N. (2008). Pragmatic comprehension in Japanese as a foreign language. *The Modern Language Journal, 92*(4), 558-576.

Takimoto, M. (2009). Exploring the effects of input-based treatment and test on the development of learners' pragmatic proficiency. *Journal of Pragmatics 41*, 1029-1046.

Walters, F. S. (2004). *An application of conversation analysis to the development of a test of second language pragmatic competence*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Walters, F.S. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing, 24*(2), 155-183.

Walters, F.S. (2009). A conversation analysis-informed test of L2 aural pragmatic comprehensions. *TESOL Quarterly, 43* (1), 29-54.

Yamashita, S.O. (1996). *Six measures of JSL pragmatics* (Technical Report #14). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Young, L.W.L. (1994). *Crosstalk and culture in Sino-American communication.* Cambridge: Cambridge University Press.

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

The author:
Carsten Roever
The University of Melbourne
Parkville, 3010
Australia
E-mail: carsten@unimelb.edu.au

Carsten Roever is a lecturer in Applied Linguistics at the University of Melbourne. He holds an MA from the University of Duisburg and a PhD from the University of Hawai'i. His research interests are second language acquisition, interlanguage pragmatics and language testing. He is the editor of the Australian Review of Applied Linguistics.

European Commission

EUR 24558 EN – Joint Research Centre – Institute for the Protection and Security of the Citizen
Title: Computer-based Assessment of Foreign Language Speaking Skills
Editor: Luísa Araújo
Luxembourg: Publications Office of the European Union
2010 – 114 pp. – 21 x 29.70 cm
EUR – Scientific and Technical Research series – ISSN 1018-5593
ISBN 978-92-79-17173-4
doi:10.2788/30519

Abstract

The goal of the conference organized by the Centre for Research on Lifelong Learning and the Directorate General of Education and Culture was to discuss the implementation of computer-based speaking skills assessment to measure foreign language proficiency. In particular, the conference aimed at addressing questions pertaining to the validity of such assessments, to identify factors that influence performance outcomes, to present existing speech eliciting formats, and to explore new domains of language assessment. Various international experts in the area of foreign language assessment presented papers related to these topics and, accordingly, the conference proceedings now available in this publication reflect state of the art research in the field of computer-based assessment (CBA) of foreign language oral skills.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

JRC
EUROPEAN COMMISSION

LB-NA-24558-EN-C

Publications Office

ISBN 978-92-79-17173-4