SOAS Working Papers in Linguistics Vol. 12 (2002): 335-362

Oral proficiency and its assessment revisited

Noriko Inagaki
ni2@soas.ac.uk

1 Introduction

1.1 The Context of the Problem

Oral proficiency testing has gained in popularity and now has a central place in language testing (LT) fields today. In the post-modern era of testing,[1] there is a widespread practice of oral proficiency testing, yet so far there has been no agreed theoretical model of communicative proficiency. This imbalance between theory and practice arose because the actual practice of oral proficiency testing predated any formulation of theoretical models of 'communicative competence' or 'communicative language ability'. The practice of 'direct' measuring of oral proficiency started in the early 1950s, when the U.S government Foreign Service Institute (FSI) initiated the project of creating a proficiency scale, and an interview method for the purpose of testing the speaking skills of their officials. As developers in FSI considered themselves pioneers in the field, they did not request any help or advice from outside of their institution (Spolsky 1995:98). The FSI project continued in collaboration with the Peace Corps, the Educational Testing Service (ETS), the U.S. government's Interagency Language Roundtable (ILR) and the American Council on the Teaching of Foreign Language (ACTFL).[2] The ACTFL Oral Proficiency Interview (OPI) was adapted from the FSI system for the use of academic institutions. The active promotion by the ACTFL with a massive number of teaching workshops and the training of foreign language professionals using 'proficiency' as an organizing principle, set off 'the Proficiency Movement', which has had a tremendous impact on all foreign language education in the US and elsewhere. The ACTFL Proficiency

[1] Spolsky (1977, 1981 quoted in 1995:2), in his review of the history of development of language testing, identified three major eras 1) 'pre-scientific (or traditional) era' (prior to the early 1950s), 2) 'psychometric-structuralist (or modern) era' (early 1950s through about the late 1960s) and 3) 'psycholinguistic-sociolinguistic (or post-modern) era' (late 1960s and following).

[2] Confidence in the FSI testing system led the Defense Language Institute and the Central Intelligence Agency (CIA) and the Peace Corps to adopt the same testing methods. In 1968, these agencies came together to produce a standardised version of the testing system, which is now known as Interagency Language Roundtable (ILR). The FSI system eventually spread outside of the US government. In the 1970s, the FSI testing system was adopted in many universities and states for the purpose of bilingual teacher certification. As the FSI system was adopted in universities, colleges and language schools, a new problem arose that many students could not score above 2/2+ level on the FSI rating scale (1 to 6). One of the recommendations made in 1979 was to set up national criteria and assessment programmes to develop language tests. This task was given to the ACTFL. ACTFL added a number of bands (or levels) at the lower end of the scale. ACTFL published provisional Guidelines in 1982 and the actual Guidelines in 1986. (Fulcher 1997:76-77)

Guidelines (rating scale) also influenced many subsequent rating scales.[3]

The ACTFL Proficiency Guidelines were released to the public in 1986. Many LT researchers[4] questioned not only the criteria used in the guidelines but also the whole validity of the FSI/IRT/ACTFL testing system. The absence of any theoretical foundation for the test system and for the rating scale was repeatedly criticised. However, the proponents of ACTFL OPI affirmed that their Proficiency Guidelines were "experientially, rather than theoretically based" (Omaggio 1983:331) and the practice of the OPI therefore continued. Yet, the ACTFL OPI remains the most widely used method of measuring oral proficiency in the US and worldwide until today.[5] ACTFL released its revised ACTFL Proficiency Guidelines—Speaking (1999) as well as revising its ACTFL OPI Tester Training Manual (1999) in 1999.

As theoretical investigations of oral proficiency had fallen behind, researchers have endeavoured to catch up. Canale and Swain (1980), Bachman (1990) and Bachman and Palmer (1996), building on Hymes's notion of 'communicative competence' (1972), proposed different theoretical models of communicative proficiency which may be applicable to language testing. Many researchers in applied linguistics (AL) and in LT today while considering Bachman's (1990) and Bachman and Palmer's (1996) models as the most current, still find them unsatisfactory. LT researchers have also investigated theoretical issues related to oral proficiency and its assessment, as well as criticising the ACTFL OPI and Guidelines.

As investigations into the nature of proficiency progressed, many researchers in AL and LT developed a growing awareness that language proficiency is multi-faceted and complex in nature. Consequently, the methodological perspectives have been broadened. There has been much collaboration between LT professionals and researchers in other related disciplines such as applied linguistics (Second Language Acquisition (SLA)[6], Interlanguage studies), pragmatics (Speech Acts[7], discourse analysis, conversation analysis), psycholinguistics, psychometrics and sociolinguistics (ethnography of communication[8]). Having gained insights from diverse approaches from various disciplines, it seems extremely important for us to re-examine the practice of current oral proficiency assessment and get ourselves reoriented before

[3] Australian Second Language Proficiency Rating (ASLPR) mirrors much of the language and philosophy of the ACTFL scale (Fulcher 1997:78). The FSI scale has also been transported to Europe and interpreted in various forms there (Spolsky 1995: 350).
[4] e.g. Bachman and Palmer 1981; Bachman and Savignon 1986; Lantolf and Frawley 1985, 1988; Bachman 1988; Kramsch 1986; Raffaldini, 1988; Valdman 1988; Shohamy 1990
[5] Since 1982, OPI Workshops have been offered at over 125 teaching institutions throughout the United States, Europe, Central and South America, and Asia. At present, OPIs are being conducted in 37 languages. For details, see Testing For Proficiency section at http://www.actfl.org/
[6] See Bachman and Cohen 1998.
[7] cf. Austin 1962. Searle 1965; 1981.
[8] Rivera ed. 1983. cf. Saville-Troike 1989; Gumperz and Hymes 1962;

launching into future studies of oral proficiency and its assessment.

1.2 Purpose of the Study

In this study, I will endeavour to answer the following two main questions related to oral proficiency.

The first regards Description of Levels of Oral Proficiency. How should we describe levels of oral proficiency? In other words, how should we approach the task of formulating a proficiency scale? What is the relationship between a proficiency scale and an appropriate model of communicative proficiency?

The second relates to Assessment Methods of Oral Proficiency. How should we assess oral proficiency? The investigations include a review of various criticisms of the assessment methods created by ACTFL. What alternative methods of assessing oral proficiency might be constructed?

The methods I employ seek to reinvestigate current discussion on these issues, to get them into perspective and to suggest some future directions. The study is partly informative, covering various issues on the subject and the arguments in the study are theoretical in nature. Therefore, the study does not intend to provide any empirical evidence on particular aspects of oral proficiency.

Section 1 introduces the problem by providing the context and the statement of purpose of the study. Section 2 investigates the method of describing proficiency (rating scales) and assessment methods of oral proficiency. The study includes a case study of the ACTFL Proficiency Guidelines—Speaking (revised 1999). Section 3 summarises the discussion of the preceding sections

2 Assessment of Oral Proficiency

This section deals with two issues related to the assessment of oral proficiency: (a) describing oral proficiency levels (a rating scale) and (b) actually assessing oral proficiency. After a brief explanation of performance assessment (2.1) and a rating scale (2.2.1), I will examine the FSI/ILR/ETS/ACTFL scales, which were prototypes for many subsequent scales (2.2.2), including a case study of the revised ACTFL Proficiency Guideline (1999) (2.2.3) and then present basic requirements and considerations for developing a future rating scale (2.2.4). In 2.3, I will investigate assessment methods of oral proficiency, which include a review of the criticisms of the ACTFL OPI (2.3.1) and then evaluate two recent new developments in oral assessment methods (2.3.2). Lastly, as a summary of the discussion, I provide a visual representation of the process of oral language testing, from test construction to assessment and discuss the problem of variables in assessment (2.4).

2.1 Performance Assessment vs. Traditional Assessment

Oral proficiency is one of those areas which are difficult to assess using *indirect* measures. The widespread practice of *indirect* discrete-point testing in the 1950s[9] could not meet the need to assess productive skills, particularly speaking skills. The need to assess the oral skills of US foreign officials led the Foreign Services Institute (FSI) to develop a *direct* measure of oral proficiency. The FSI interview and accompanying rating scales are often seen as the beginning of *performance assessment* in second language testing.

According to McNamara (1996:6), a defining characteristic of 'performance assessment' is that "actual performances of relevant tasks are required of candidates, rather than more abstract demonstrations of knowledge, often by means of pencil-and-paper tests". 'Performance assessment' has been used in different fields. McNamara (1996:8) claims that "second language performance assessment is distinguished from performance assessment in other contexts because of the simultaneous role of language as a medium or vehicle of performance, and as a potential target of assessment itself". The following shows the features of a performance-based assessment in comparison with a traditional fixed response (pencil-and-paper) assessment.

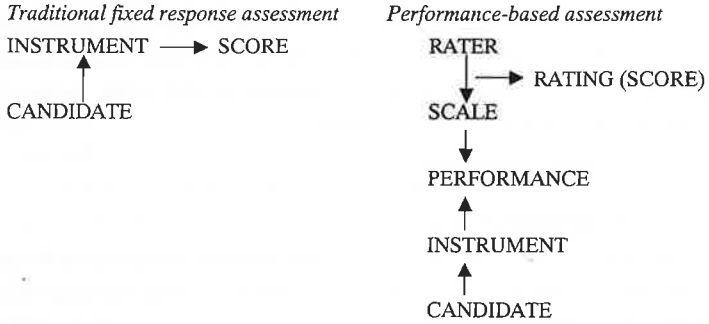

Figure 2.1 The characteristics of performed assessment (McNamara 1996: 9)

[9] The model of proficiency based on this approach is called 'the Skills and Component Model', because language proficiency is considered to be divisible into skills (listening, speaking, reading and writing) and components (phonology/ orthography, morphology, syntax and lexicon) (Carroll 1968:54-55). Testing linguistic discrete points was popular for the following reasons: they yield data which are easily quantifiable; they allow a wide coverage of items; tests which focus on 'discrete' linguistic items are efficient and have the reliability of marking associated with objectively scored tests. The weakness of the 'discrete-point' approach is that measurement depends on only that part of proficiency, which is neatly quantifiable (Weir 1988:2). Some examples of typical discrete-point testing are the *Michigan Test of English Language Proficiency* and the early version of the *Test of English as a Foreign Language* (Kunnan 1999:708).

A performance-based assessment is more complicated than traditional fixed response assessment because of the two additional stages in the process. One is *performance* by the candidate, which is, in many cases, elicited through an instrument such as an interview and the other is a *rating process* in which raters[10] judge the quality of the performance against rating scales. McNamara (1997:134) points out that the richness of performance assessment also introduces some new problems; the complexity and potential variability of the performance assessment setting can easily jeopardize fairness and influence the candidate's final score. Further discussion on variables in performance assessment will be discussed in Section 2.4.

2.2 Description of Proficiency Levels: Rating Scale

2.2.1 What is a rating scale (or proficiency scale)?

As performance assessment involves a rater's judgment of the candidate's performance against some rating scale, the quality of the rating scale has tremendous impact on the quality of the performance. McNamara (1997:135) describes the importance of a rating scale as follows: "At the heart of the construct validity[11] of many performance assessments is a rating scale... as this offers an operational definition of the construct being measured."

What is a rating scale? It is "the description of language proficiency consisting of a series of constructed levels against which the candidate's performance is judged (Davies, et al. 1999:153)". A rating scale provides an operational definition of proficiency and therefore is sometimes referred to as a 'proficiency scale'. A proficiency scale typically consists of sub-scales for the skills of speaking, reading, writing and listening (ibid. 154). Scales usually range from zero through to native-like proficiency (Stern 1983:341).[12] At each level or band, statements describe the level of performance required of the candidate, which are often called 'descriptors' (Davies, et al. 1999:43). A rating scale is not only important for rating procedure but also indispensable for the construction of performance tests.[13]

Despite the crucial place that rating scales occupy in performance assessment, until recently "most scales of language proficiency have been produced to appeal to intuition and to those scales which already exist rather than to theories of linguistic

[10] *Rater* refers to a "judge or observer who operates a rating scale in the measurement of oral and written proficiency (Davies, et al. 1999:161)".
[11] Construct validity of a language test refers to "an indication of how representative it is of an underlying theory of language learning. Construct validation involves an investigations of the qualities that a test measures, thus providing a basis for the rational of a test" (Davies, et al. 1999:33).
[12] A series of studies show that 'native-like proficiency' is a very ambivalent notion, because the levels of performance of native speakers can and do vary. See McNamara 1996:182-198.
[13] There are different purposes for using proficiency scales, which will be further discussed in 2.2.4. North and Schneider (1998:219) list the purposes of using scales other than for test development.

description or of measurement (North and Schneider 1998:217)". Most of these scales directly or indirectly relate back to the US FSI scale, which later evolved into other scales such as the ILR (Interagency Language Roundtable) scale and the ACTFL (American Council on the Teaching of Foreign Language) Proficiency Guidelines.

2.2.2 The FSI/ILR/ETS/ACTFL proficiency scales

The original FSI proficiency scale, developed in 1958, consisted of five-paragraph-length descriptions of general speaking performance, ranging from Elementary Proficiency (Level 1) up to Native or Bilingual Proficiency (Level 5). The descriptors are created, based on typical language use requirements and degrees of language performance by FSI graduates in actual foreign service (Clark and Clifford 1988:130; See Appendix 1 for the verbal descriptions of the original FSI proficiency levels). The criteria are "accent, comprehension, fluency, grammar and vocabulary (Fulcher 1997:76)." The FSI scale has a range of 11 possible scores (0, 0+, 1, 1+, 2, 2+, 3, 3+, 4, 4+, and 5) (Clark and Clifford 1988:131). Raters are not expected to use the scale for rating, but for checking if they have paid attention to each of the factors listed in their holistic marking (Fulcher 1997:76). The FSI proficiency scale "was something of a revolution (ibid.)", because it was the first rating scale provided with verbal descriptions.

As other US government agencies, such as the Defense Language Institute (DLI), the Language school of the Central Intelligence Agency (CIA) and the U.S. Peace Corps started to use the FSI style interview and the proficiency scale, variants of FSI scale were developed. In 1968, these different agencies came together to produce a standardized version of scale, which is known as the ILR (Interagency Language Roundtable) scale. In the early 1980s, 'ACTFL/ETS Speaking Proficiency Guidelines' was produced in cooperation with ETS and ACTFL for the use in academic contexts.[14] The ACTFL/ETS scale makes no distinction above FSI level 3, because students do not normally score higher than level 3 and measurement discrimination above level 3 is not usually necessary in regular academic contexts (Clark and Clifford, 1988:132-33). (See Figure 2.2.)

[14] The details of the development of these ratings scales can be found in Liskin-Gasparro 1984; Barnwell 1987, 1996;

ACTFL/ETS	FSI	ILR
	0	0 no practical proficiency
Novice-Low		
Novice-Mid		
Novice-High		0+
Intermediate-Low	1	1 survival proficiency
Intermediate-Mid		
Intermediate-High	1+	1+
Advanced	2	2 limited working proficiency
Advanced-Plus	2+	
Superior	3	3 professional working proficiency
	3+	3+
	4	4 distinguished proficiency
	4+	4+
	5	5 native or bilingual proficiency

Figure 2.2 Relationship of ACTFL/ETS to FSI scale and ILR scale
(Cf. Clark and Clifford 1988:133; Buck ed. 1989)

The ACTFL/ETS guidelines consist of verbal descriptions of four major levels (Novice, Intermediate, Advanced and Superior) and their sublevels (Total of 9 levels).[15] The ACTFL/ETS guidelines for speaking were further refined and the proficiency descriptions for all four skills came out as the 'ACTFL Proficiency Guidelines' (provisional guidelines in 1982; final guidelines in 1986)[16].

Many researchers have criticised the ACTFL scales mostly for their lack of

[15] Verbal descriptions for major levels of the ACTFL/ETS guidelines are as follows: "**NOVICE:** The novice level is characterized by an ability to communicate minimally with learned material"; "**INTERMEDIATE:** The intermediate level is characterized by an ability to create with the language by combining and recombining learned elements, though primarily in a reactive mode; initiate, minimally sustain, and close in a simple way basic communicative tasks; and ask and answer questions"; "**ADVANCED:** The advanced level is characterized by an ability to converse in a clearly participatory fashion; initiate, sustain, and bring to closure a wide variety of communicative tasks, including those that require an increased ability to convey meaning with diverse language strategies due to a complication or an unforeseen turn of events; satisfy the requirements of school and work situations; and narrate and describe with paragraph-length connected discourse"; "**SUPERIOR:** The superior level is characterized by an ability to participate effectively in most formal and informal conversations on practical, social, professional and abstract topics; and support opinions and hypothesize using nativelike discourse strategies." (Clark and Clifford 1988:133-135).

[16] The ACTFL Proficiency Guidelines also consist of 9 levels. The 1986 version of the ACTFL Guidelines can be found in Henning 1992:370-371.

"empirical underpinnings (Fulcher 1996:164)".[17] In response to such criticisms, Henning (1990), Dandonoli and Henning (1990) and Henning (1992) attempted to present empirical evidence for the validity of the ACTFL scale.[18] Henning (1992:369) claims that these studies managed to provide "some limited research evidence related to the reliability, construct and criterion-related validity[19], scalability, and generalizability obtained according to the ACTFL scale". However, Fulcher (1996) later re-evaluated these studies and invalidated their validity claims.

Another common criticism is that the ACTFL approach is only experientially based and that its "theoretical underpinnings are shaky (Valdman 1988:121)". ACTFL has never claimed that the Guidelines reflect a particular SLA model (Dandonoli and Henning 1990:12) or any model of communicative proficiency or of measurement. Omaggio (1983:331) sees the experiential base of the Guideline as a strength rather than a weakness. He states that the Guidelines are "experientially, rather than theoretically, based; that is they *describe* the way language learners and acquirers typically function along the whole range of possible levels of competence, rather than *prescribe* the way any given theorist *thinks* learners ought to function (emphasis in original)".[20] However, the ordering of skills and function could vary, depending on the learner's purpose or needs in learning the language and on the learning method. Spolsky (1989:65) claims that "a single set of guidelines or a single scale could only be justified if there were evidence of an empirically provable necessary learning order". However, no evidence of any particular SLA research has been provided to support the ordering in the Guidelines.

The ACTFL Guidelines have also been criticised for their 'behavioural' or 'real-life (RL)' approach (Bachman 1990:325-330).[21] The RL approach "defines language proficiency as the ability to perform language tasks in non-test situations, and authenticity as the extent to which test tasks replicate 'real-life' language use tasks (ibid. 307)." The proficiency scale in the RL approach seeks to describe globally or holistically what the candidate can do at a particular level in terms of features of

[17] e.g. Lantolf and Frawley 1985, 1988; Pienemann, Johnson and Brindley 1988.

[18] The aims of the studies include the following: 1. to investigate the claims that traits (or constructs) are confounded with test methods in the ACTFL scale (in response to Bachman and Savignon 1986; Bachman 1988), 2. to support the difficulty hierarchy represented by the Guidelines (scalability), 3. to examine whether the scale is generalisable across languages (generalisability), and 4. to discover to what extent naïve (untrained) native raters were capable of interpreting descriptors. See Henning (1992).

[19] Criterion-related validity refers to indicating how close a test is to its criterion. It is established statistically using correlation (Davies, et al.1999:39).

[20] Liskin-Gasparro (1984:37) also states that the ACTFL Guidelines are "descriptive rather than prescriptive, based on experience rather than theory".

[21] Bachman (1990) observes that there are two approaches to defining authenticity: the 'real-life (RL)' approach and the 'interactive-ability' approach. Bachman's understanding of authenticity is further developed in Bachman and Palmer 1996:23-29.

'real-life' performance. One problem with this approach is that "it treats the behavioural manifestation of an ability as the trait itself (ibid. 309)". Bachman claims that the identification of trait with performance "does not permit us to make inferences beyond the testing context, and thus severely limits both the interpretation and use of test results and the type of evidence that can be brought forth in support of that interpretation or use (ibid.)".

Despite the lack of theoretical or empirical underpinnings, the ACTFL level descriptions seem to be interpreted as "learner norms" or "a picture of universal developmental patterns" (Brindley 1998:116). This is probably because ACTFL actively promoted the Oral Proficiency Interview (OPI) and the Guidelines and also conducted massive training sessions of foreign language professionals using proficiency as an organizing principle, which came to be known as the Proficiency Movement. De Jong (1990:72 quoted in North and Schneider 1998:221) mentions the case that the acceptability of a scale relies "primarily on the authority of the scholars involved in their definition, or on the political status of the bodies that control and promote them". His description seems to portray the situation of the ACTFL scale rather accurately.

2.2.3 Evaluation of the ACTFL Proficiency Guidelines (revised 1999)

Having received various criticisms, ACTFL recently released new ACTFL Proficiency Guidelines (revised 1999).[22] The ACTFL claims that the 1999 revision is made as a result of re-evaluation and refinement of the 1986 version after "additional years of oral testing and of interpretation of the Guidelines, as well as numerous research projects, scholarly articles and debates (Breiner-Sanders, et al. 2000:13)".

The new 1999 version is characterised by an increased amount of information: the actual descriptors are much longer than those in the 1986 edition in *the ACTFL OPI Interview Tester Training Manual* (1999), the Guidelines are further accompanied by "Explanatory Notes" in order to "provide insights and definitions to clarify the terminology and assumptions of the descriptions (Swender, et al. eds. 1999:81)" and by a "Chart of Summary Highlights"[23] in order "to alert the reader to the major features of the levels and to serve as a quick reference (Breiner-Sanders, et al. 2000:14)". Such an elaboration of the verbal descriptors may actually have made it harder for raters to pay attention to all the descriptions of each level.

A noticeable difference in format is that the 1999 version are in descending

[22] ACTFL Proficiency Guidelines—Speaking (revised 1999) is downloadable from ACTFL web site. http://www.actfl.org/index.cfm?weburl=/public/articles/index.cfm?cat=28

[23] A 'Chart of Summary Highlights' provides general description of characteristics of speakers at major levels (Superior, Advanced, Intermediate and Novice).

order rather than in ascending order.[24] Another change was the division of Advanced level into High, Mid, and Low.[25] However, no striking differences are found in the descriptions of the main characteristics of major levels, probably because there are no major changes in assessment criteria.[26] Salaberry (2000:297), through critical analysis of the revised ACTFL-OPI Tester Training Manual (ACTFL-TTM) (Swender, et al. eds. 1999), concludes that "the 1999 revision of the ACTFL-TTM has not introduced substantial changes to the 1986 edition".

The 1999 revision of the ACTFL Proficiency Guidelines indicates that there has not been any major change in their approach. In fact, the revision only served to update the information, clarify some issues and correct some misunderstandings.[27] Perhaps there will have to be a drastic change in their overall philosophy of test development, such as renouncing their long-practiced experiential approach, before we can expect any major changes in their actual Guidelines.

2.2.4 Future directions in the formulation of rating scales

Empirically based and theoretically based approach

These criticisms of the intuitive approach of the FSI/ILR/ETS/ACTFL scales led researchers to attempt new approaches in formulating rating scales. Although many researchers agree that proficiency scales should be empirically based and theoretically based, the problem is that we have yet to develop a satisfactory model of proficiency. Brindley (1989:56 quoted in North and Schneider 1998:242) states that "we cannot wait for the emergence of empirically validated models of proficiency in order to build up criteria for assessing learners' second language performance."

The attempts to formulate empirically based scales of proficiency only began recently and are still at a trial and error stage. North (1995) and North and Schneider (1998) report a project of developing a scale of language proficiency "in the form of descriptor bank".[28] These studies are part of the Council of Europe project for

[24] Breiner-Sanders, et al (2000:14) explain two advantages of this: "First, it emphasizes that the High levels are more closely related to the level above than to the one below, and represents a considerable step towards accomplishing the functions at the level above, not just excellence in the functions of the level itself. Second it allows for fewer negatives and less redundancy in the descriptions when they refer, as they must, to the inability of a speaker to function consistently at a higher level."

[25] This makes the 1999 guidelines consist of ten levels instead of nine. This change was made to meet the growing needs in the academic and commercial communities (Breiner-Sanders, et. al. 2000:14).

[26] The assessment criteria, which give brief descriptions of performance at each level in the following five areas: 1. global tasks/functions, 2. context, 3. content, 4. accuracy and 5. text type, remain the same (Swender et al. ed. 1999).

[27] The purposes of the revision are to "to make the document more accessible to those who have not received recent training in ACTFL oral proficiency testing, to clarify the issues that have divided testers and teacher, and to provide a corrective to what the committee perceived to have been possible misinterpretations of the descriptions provided in earlier versions of the Guidelines (Breiner-Sanders, et al. 2000:14)".

[28] The studies were conducted by the Swiss National Science Research Council. The project team first created a descriptor pool through the analysis of forty-one existing proficiency scales. Then the descriptor pool went through a qualitative validation process through wide consultation with foreign-language teacher representatives. Then quantitative validation was undertaken by analysing the collected data using multi-faceted Rasch measurement (see footnote 53).

creating a *Common European Framework of Reference (for Language Learning and Teaching)*[29], which is also connected with the *European Language Portfolio* project.[30] In the Portfolio, "information about achievement in different educational contexts will be presented according to the levels of the *Common European Framework of Reference*, in order to allow trans-national comparability of qualifications (ECC2a)". The Council of Europe has rigorously investigated the issues related to the formulation of the common framework scale. A document entitled "Modern Languages: Learning, Teaching Assessment. A Common European Framework of Reference (ECC2a)[31]" presented at the conference of the Council of Europe in Strasbourg in 1997, shows the thoroughness of their research. This Council of Europe project is promising for two reasons: firstly, because this is the first thorough large-scale attempt to formulate an empirically- and theoretically-based scale, and secondly, because of their philosophy of test development. They have made their studies accountable outside of their organization by making the paper open to scholarly discussion.

We are still looking for a proper procedure for developing empirically based rating scales. North and Schneider's (1998) project demonstrate one model procedure.[32] Brindley (1998:135) evaluates these recent attempts and claims that through such efforts, "it might eventually be possible not only to develop a better understanding of task performance in context but also to gain deeper insights into the complex linguistic and cognitive skills which go to make up language proficiency".

[29] The original proposal was made in 1991. The revised draft framework is currently in use, which was submitted in 1996 and 1997.) See the Europe of Cultural Co-operation web page at http://culture2.coe.int/portfolio/documents_intro/common_framework.html (ECC2a).

[30] *Portfolio* consists of three parts: "a passport recording formal qualification in an internationally transparent manner, language biography describing language knowledge and learning experiences, and a dossier in which the learners' own work can be included" http://culture2.coe.int/portfolio/inc.asp?L=E&M=$t/208-1-0-1/main_pages/contents_portfolio.html (ECC 1).

[31] Available from http://culture.coe.fr/lang/eng/eedu2.4.html.

[32] The following is the procedure which was used in North and Schneider's (1998:242-243) project: 1) comprehensive documentation of experience and consensus in the field of proficiency scales; 2) classification of descriptors to a taxonomy informed by theoretical models; 3) pre-testing of categories; 4) formulations and translations to ensure that the descriptors represent clear, useful, relevant, accessible, stand-alone criterion statements; 5) scaling of the descriptors with a measurement model; replication of the scale values.
Turner and Upshur (1995) attempt another procedure. They derive scale descriptors from analysis of learner performance, using a series of binary choices concerning key features distinguishing between score models.

Some basic requirements and considerations for developing future rating scales

Experience with the first proficiency scales —the FSI/ILR/ETS/ACTFL scales —gave much insight to future scale developers. Once proficiency scales such as these have been widely used in significant contexts, it may not be easy to make any major change because of the possible impact it may have on individuals and institutions involved. Therefore developers of rating scales should keep in mind the seriousness of the task. The following are some basic requirements and consideration for the development of future rating scales.

First, a proficiency scale should reflect current theoretical models of proficiency. The model represents a particular view of knowledge of language and language use, so this should be the starting point and the operational model and the proficiency scale should be derived from it. Yet this must not be a one-way process. The insights into the nature of proficiency obtained by using the actual scale should be fed back in to amend the theoretical model.

Secondly, a proficiency scale should not be made by appeal to intuition but be empirically validated based on current models of measurement. Once a provisional pool of descriptors is formed, these descriptors should be put to the test for validation. Recently a combination of qualitative and quantitative methods has been employed for this purpose.[33] The empirical validation of any scale is indispensable in order that that scale may be objectively based.[34]

Thirdly, the purposes of a proficiency scale should be clear and the content of the descriptors should match these purposes. Alderson (1991:71-76) uses a three-way classification according to the purposes for which scales are written.

The first are *assessor-oriented scales*, "with the function of guiding the rating process (ibid 73)" The descriptors of such scales should contain "aspects of the quality of the performance expected"[35]. As the assessors are usually expected to familiarise themselves with the scale before using it, the descriptors can include technical terms.[36]

The second are *user-oriented scales*, "with a reporting function (Alderson 1991:72)", which help "test users —employers and admission officers— to interpret test results (ibid.)", The descriptors must use non-technical language and should have

[33] For further information on qualitative approach, see Banerjee and Luona1997. The quantitative method involves statistical processes using various test theories such as Classical Test Theory Analysis and/or Item Response theory. For the details of quantitative analysis, see Bachman and Eignor 1997.

[34] Some studies demonstrate that diverse rater groups tend to have different perceptions of the same speech performance (Barnwell 1989; Hadden 1991;Chalhoub-Deville 1995). Therefore it is important to derive the criteria/dimension salient to diverse rater groups using empirical validation processes.

[35] http://www.rimini.com/provveditorato/didattica/saperi/linguestraniere/links/quadro_europeo/inglese/paragrafi/eedu2_4i.htm (ECC2b).

[36] Barnwell (1989:155) comments that raters learn to use their scale in a particular way through a process of shared experience and socialization.

a positive tone even at the lower level(ECC2b).[37]

The third are *constructor-oriented* scales, "with the function of guiding the construction of test appropriate levels (Alderson 1991:74)". Such scales should contain "specific communicative tasks which the learner might be asked to perform in tests (ECC2b)".

Alderson (op.cit., 74) claims that problems occur when the three functions are confused. I concur with his view. The rating scale that is designed to serve many different purposes is likely to prove less useful for any one given purpose.

Fourthly, the generalisability of the descriptors should be determined depending upon the context in which a scale may be used. If the scale is meant to be a common scale for diverse contexts, the descriptors need to be sufficiently general to accommodate all such contexts. If the scale is meant to assess language proficiency for a specific context, the descriptors need to be geared toward that specific context. For example, the descriptors for the use of assessing the English proficiency of health professionals need to express the quality of the language performance expected in that specific context.

No matter how generalisable the descriptors may be, however, they are not automatically transferable to other language contexts, because socio-cultural expectations usually vary from culture to culture. For instance, Japanese is characterised by its highly complex honorific system and proficient speakers in Japanese are expected to handle its intricacies successfully. There is, however, little consideration of such criteria included in the ACTFL Guidelines for Japanese language, because they are simply a direct translation of the English language Guidelines.[38] Having recognised some inadequacy in Japanese testing situations, interviewers of Japanese language are advised to give two role-play tasks (one requiring use of honorifics and the other requiring use of casual/informal speech) at Superior levels in order to test if candidates can handle these language-specific devices. Such devices, however, are expected not only at Superior levels, but at all levels. Also since such elements regarding politeness of language are not articulated in the Guidelines, raters tend to employ their own subjective judgement. The scale developers should not assume that one common, generalisable or culture-free scale will work for languages in all cultures but must take such culture-relevant or culture-specific features into consideration.[39]

[37] http://www.rimini.com/provveditorato/didattica/saperi/linguestraniere/links/quadro_europeo/inglese/paragrafi/eedu2_4i.htm (ECC2b).

[38] See 'ACTFL Gengo Unyoo Nooryoku Kijun—Wa-ginoo 1999 Revised' (ACTFL Proficiency Guidelines—Speaking (revised1999)) in ACTFL Oral Proficiency Interview Tester Training Manual, 1999:120-125.

[39] Ideally LT researchers who use the target language as L1 should develop the descriptors that contain such culture-specific aspects of proficiency.

Fifthly, the number of levels of the scale should "be adequate to show progression in different sectors, but in any particular context, should not exceed the number of levels people are capable of making reasonably consistent distinctions between (North 1995:447)". ACTFL has modified the number of levels for practical reasons. For example, 11 possible scores (0, 0+, 1, 1+, 2, 2+, 3, 3+, 4, 4+, and 5) in the original FSI/ILR scale were not appropriate for the assessment of the academic contexts of ACTFL/ETS; therefore was adjusted into the 9 levels, which was adopted in the ACTFL 1986 Guidelines. Further adjustment was made in the 1999 revision by dividing Advanced level into High, Mid, and Low (a total of 10 levels). In 1996, ACTFL developed a Standard Speaking Test (SST) and the accompanying SST level descriptors (9 levels). (See Figure 2.3)

ACTFL Level	SST Level
Advanced	9
Intermediate-High	8
Intermediate-Mid	7
	6
Intermediate-Low	5
	4
Novice-High	3
Novice-Mid	2
Novice-Low	1

Figure 2.3 Relationship of ACTFL Level to SST level
(Source: *Standard Speaking Test Manual.* ACTFL-ALC Press. 1996:30)

SST is modelled after OPI and designed to measure English speaking proficiency primarily from Novice-Low through to Intermediate-High Level. SST does not discriminate higher than Intermediate-High Level, but includes finer sublevels for Intermediate-Mid and Intermediate-Low Levels (ACTFL-ALC 1996:30). These finer levels were created out of the practical need to discriminate between the levels of the majority of ESL learners in Japan, who often fall into either Intermediate-Mid or Intermediate-Low in OPI level. It is, however, doubtful that ACTFL can provide statistical evidence to justify such finer sublevels being employed.

The sixth and last consideration relates to the format of the rating feedback to candidates. The descriptors of the ACTFL Guidelines seem to suggest that there is an underlying assumption that communicative proficiency consists of multiple components.[40] Yet the final rating given is a unidimensional score. Spolsky

[40] For example, one of the assessment criteria "accuracy" is judged by categories such as fluency, grammar, pragmatic competence, pronunciation, sociolinguistic competence and vocabulary. (Buck et al. eds. 1989: 3-1.)

(1995:358) argues that "it is...a mistake to assume that knowledge of a language is best considered as the possession of something external and automatically measurable, like money". He claims that it is more like "knowing a friend" and that it is absurd to think that such knowledge can be squeezed into a single number or single point on a unidimensional scale (ibid). One possible modification that I suggest is to combine a single score feedback with some diagnostic descriptions on particular features of a candidate's performance. This kind of feedback may still not be sufficient, but at least is more consistent with the current view of the multi-componential nature of proficiency.

2.3 Assessment methods of Oral Proficiency

2.3.1 Criticisms of the ACTFL OPI

The FSI/ILR/ACTFL Oral Proficiency Interview, which is generally considered the first "direct" measure of oral proficiency, is an interactive face-to-face interview. The interviewer leads a candidate through various communication activities in order to get a rateable speech sample. The interview lasts from ten to thirty minutes depending on the proficiency level of the candidate. The OPI includes the following four phrases: *warm-up*, *level checks*, *probes* and *wind-down*. The interviewer repeats level checks and probes until the candidate's *performance floor* and *performance ceiling* is established.[41]

Early criticisms of the OPI have been "focused on various kinds of validity—construct, content[42], concurrent[43]—and reliability, rating procedures and rating criteria (Lazaraton 1992:373)".[44] Its authenticity has also been called into question. For example, Spolsky (1985:34) points out that the OPI interview is inauthentic in that the interviewer manipulates the conversation to allow the candidate to demonstrate the full range of his or her abilities in performing various artificial tasks.[45] Despite the criticism, there seems to be widespread agreement that the oral

[41] *Performance ceiling* refers to "the limitation of a performance defined by the operations which are beyond the interviewee's ability to perform well". *Performance floor* refers to "the linguistic operations the interviewee can perform with consistent success and accuracy" (Buck et al. eds. 1989: G-4).
[42] 'Content validity' refers to "a conceptual or non-statistical validity based on a systematic analysis of the test content to determine whether it includes an adequate sample of the target domain to be measured (Davies, et al. 1999:34)".
[43] 'Concurrent validity' refers to "a type of validity which is concerned with the relationship between what is measured by a test (usually a newly developed test) and another existing criterion measure (ibid. 30)".
[44] See Bachman and Palmer 1981; Bachman and Savignon 1986; Bachman 1988; Kramsch 1986; Lantolf and Frawley 1985, 1988.
[45] The authenticity issue will also be discussed in the next section. Detailed treatment of the authenticity debate is beyond the scope of this study. See Stevenson 1985; Shohamy and Reves 1985;

interview is the most appropriate vehicle for oral proficiency assessment (Lazaraton 1992:373). Recently researchers have started to look into the actual interaction between an interviewer and a candidate elicited in the OPI to determine the quality of the interaction. For example, Van Lier's (1989) studies shows that the interaction in the OPI is characterised by asymmetrical contingency, while everyday conversation is based on mutual contingency with equal distribution of interaction. Johnson and Tyler's (1998:47) study shows that the interaction of OPI does not demonstrate "the salient features of conversation involved in turn-taking and negotiation of topic". Such qualitative studies of the OPI from a discourse/conversation analytic perspective have given new insights into the nature of the oral proficiency interview process.[46]

2.3.2 Some new developments in the assessment of oral proficiency

The first strand: Development of variations of the OPI: SOPI and COPI

After four decades of the practice of OPI, there are at least two new developments exploring other methods of assessing methods of oral proficiency. The first strand is found in the development of two variations of OPI, produced by the Center for Applied Linguistics (CAL). One is the Simulated Oral Proficiency Interview (SOPI).[47] It is a semi-direct (tape-mediated) version of OPI. The SOPI was developed for practical reasons –for providing proficiency testing in situations where it is difficult to give an oral interview due to the unavailability of trained testers or lack of time and budget for such testing. The SOPI follows the OPI process as closely as possible and is assessed globally using the ACTFL Proficiency Guidelines. The SOPI is administered to a candidate using a test booklet and a master tape. The candidate can take a test individually using two tape recorders, one for listening to the instruction and the other for recording the candidate's response to each task given (CAL2).

Stansfield and Kenyon (1992), focusing on test scores, claim that the OPI and SOPI are highly comparable as measures of oral proficiency. Shohamy's (1994) discourse analytic approach of the study, however, shows that the discourses elicited in the SOPI and OPI "differed in rhetorical functions, structures, genre, communicative properties, discourse strategies, prosodic paralinguistic features, speech functions and discourse markers (Shohamy 1998:162)". Kuo and Xixiang's (1997:510) study shows that even though both OPI and SOPI are assessed using the

Seliger 1985; Spolsky 1985. "Authenticity" in language testing has been debated since the mid 1960s. A summary of the discussions until 2000 can be found in Lewkowicz 2000.

[46] Also see Lazaraton 1992; Ross and Berwick 1992; Ross 1992; Kormos 1999; Young and Milanovic 1992. Young 1995; Young and He eds. 1998 contains other recent articles on discourse approaches to oral proficiency assessment.

[47] SOPIs are available for Arabic, Chinese, French, German, Japanese, Hebrew, Hausa, Indonesian, Portuguese and Spanish as of January 2000. http://www.cal.org/public/FLTests.htm (CAL2).

ACTFL guidelines, the assessment procedures of the two tests tend to differ. For example, the rating of SOPI tends to be a "segment rating" rather than a "global rating".[48]

The other variation of OPI is the Computerized Oral Proficiency Interview (COPI). The goal of COPI is "to use the advantages of computer technology to improve the SOPI, by giving examinees more control of various aspects of the testing situation and increasing raters' efficiency in scoring the test (CAL1)". The steps of COPI "include a self-assessment of the examinee's proficiency level; a practice task in which the examinee will be given the choice of starting at an easier or more difficult level; several picture-based, topic-based and situation-based tasks; and lastly feedback on the level of tasks the examinee took (CAL1)". Empirical studies of the COPI have just begun.[49] It is too early to make any systematic comments on this new test.

The development of the COPI and SOPI needs careful evaluation. Practicality is one important factor in developing a test along with all the other factors such as validity, reliability, and authenticity,[50] but should never be placed before the other factors. Test users should weigh the advantages and disadvantages of using SOPI or COPI. COPI or SOPI can never replace the face-to-face personal interaction of OPI. Affective factors in using these alternative tests must be carefully examined as well as the linguistic features of the elicited speech of the candidate.

The second strand: Alternative assessment

Another strand in the recent oral proficiency assessment comes from a search for authenticity. Spolsky (1985:39) states that "any language test is by its very nature inauthentic, abnormal language behaviour" and that the "only full solution is the ethnographic testing, which is the long, patient and sympathetic observation by observers who care to help".[51] Authentic assessment, then, seems possible only through a non-test, 'alternative assessment', which can be defined as "an ongoing process involving the student and teacher in making judgments about the student's progress in language using non-conventional strategies (Hancock 1994)". Portfolio assessment, self-assessment, self-monitoring, and project assessment, are some examples of alternative methods of assessment.

[48] *Segment rating* refers to an evaluation of a response to a particular test question, and not of the speaker as a whole. For example, there are fifteen tasks in a SOPI and therefore fifteen segment ratings. Each response to a task is independently rated using the ACTFL level descriptors. At the end, the fifteen segment ratings form the basis for the *global rating* through a rigorous algorithm (Kuo and Xixian 1997:510).

[49] See Malabonga.2000

[50] Nevo and Shohamy (1986) point out that utility, feasibility and fairness need to be examined, as well as validity and reliability in test development (Reve 1991:183; Shohamy 1988:177).

[51] Shohamy and Reves (1985:58) also claim that we need an ethnographic approach in order to elicit not authentic test language but authentic real-life language.

In a conventional test setting, there is a clear role distinction between an assessor and a candidate, whereas in alternative assessment, 'assessment' is seen as "an interactive process that engages both teacher and student in monitoring the student's performance (Hancock 1994)". Alternative assessment has resulted in a paradigm shift in assessment methodology. Investigation of various possible alternative assessment methods in oral proficiency should be encouraged in the future.

2.4 The Dynamic Process of Oral Proficiency Assessment

As a summary of what we have discussed so far, I provide a visual representation of the process of oral proficiency assessment in Figure 2.4.

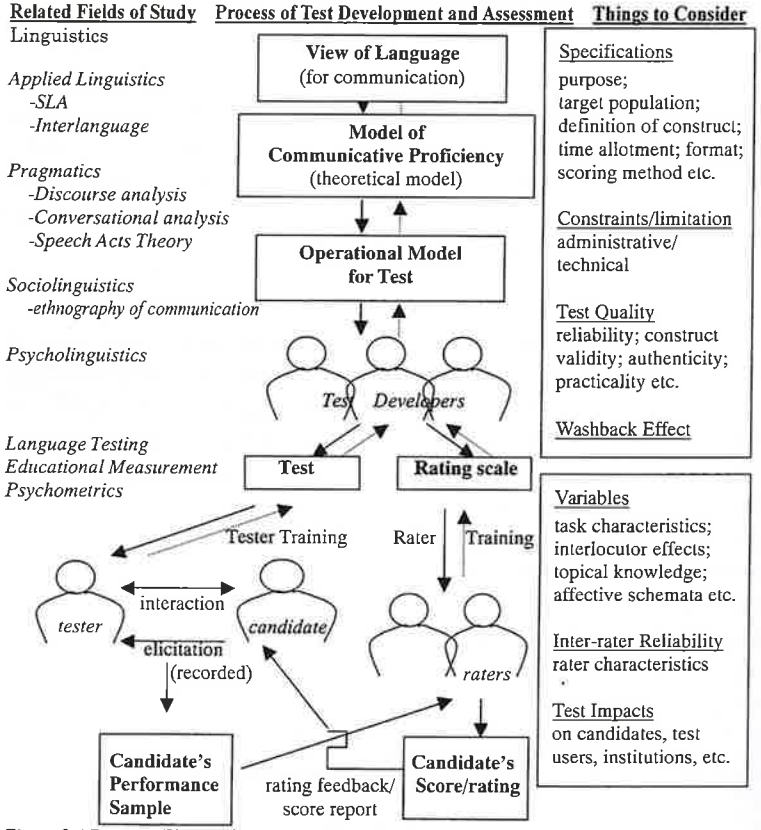

Figure 2.4 Dynamic Flow in the process of Oral Proficiency Assessment

The flowchart shows the process of test development (the upper half) and of the actual performance testing (the lower half). The dotted arrows going in a reverse direction show that findings in the actual practice of performance assessment may be reflected back, making amendments to theoretical models. The right column shows "the things to consider" at different stages in the process. The left column shows "the related fields of study" which contribute to the process of assessment. Assessing oral proficiency through performance testing is considered to be a "direct" measure of oral proficiency. However this chart reveals that the whole process is extremely complicated and that things can go wrong at almost any stage. For example, task choice or task processing conditions during the interview and/or interlocutor effects could have an influence on the test outcome.[52] Similarly, in the rating process, raters may vary in the standards they use or may not consistently apply those standards (McNamara1997:134). McNamara states:

> The rating finally given is thus a result of a host of factors interacting with each other; the process is like a piece of complicated machinery with a rating popping out at the end. This means that any idea that there is a transparent relation between the candidate's performance and the rating given is naïve (ibid.).

Recent advances in psychometrics have made it possible to take some control of the impact of variables such as rater characteristics and task characteristics in performance assessment. The multifaceted Rasch measuremen[53] and Generalizability theory (G-theory)[54] are particularly useful for this purpose. The nature of variability related to interlocutor or task has been investigated using qualitative approaches such as discourse analysis or conversational analysis. Considering all the complexity involved in the process, one must question how truly 'direct' is 'direct' measurement of oral proficiency in reality.

---

[52] Cultural difference is another variable. Any performance testing of second language oral proficiency is inevitably a cross-cultural encounter, which involves at least two cultures: that of the interviewer (or rater) and that of the candidate. The cultural differences may cause interviewer and raters to misunderstand the candidate's performance. For example, Ross's (1998) analyses of six English as a second language (ESL) and six Japanese as a second language (JSL) OPIs show that "the role of cultural background of the interviewer and the apparent differences in pragmatic strategies for dealing with interlocutor attempts to manage the interview may lead to dramatic differences in the interviewer's understanding of what sort of proficiency is being demonstrated (ibid., 47)"

[53] Rasch analysis is a branch of item response theory, the one-parameter model, developed by Georg Rasch (a Danish psychometrician). Rasch's Basic model was restricted to the analysis of dichotomously scored data. Further development made possible the analysis of data from rating scales. Linacre extended model, multi-faceted Rasch measurement, made it possible to analyse other facets of the assessment setting, such as the impact of the characteristics of raters (Davies, et al. 1999:160). See Linacre 1989; McNamara 1996.

[54] G-theory is an extension of classical test theory, differing in its view of measurement error. G-theory aims to consider various sources of error separately and estimate the contribution each makes to the overall error, using the statistical procedure called ANOVA. It estimates the effects caused by varying the number of items and tasks in a test or the number of raters involved in scoring performance. (Davies, et al. 1999:67). See Brennan, 1983; Shavelson and Webb, 1991.

3 Conclusion

In the 1950s, the US government Foreign Services Institutes (FSI) created an oral interview test in order to assess the speaking ability of their officers working overseas. The developers in FSI created their own rating scale and interview method without any opportunity for open scholarly interchange. Confidence in the FSI system led them to continue this 'experientially-based' approach, which in turn produced other tests like the ACTFL Oral Proficiency Interview. As praxis predated theory, theoretical studies related to oral proficiency and its assessment method have fallen behind.

I have investigated and analysed current theoretical discussions on two issues relating to oral proficiency: the description of oral proficiency and the assessment method for oral proficiency. I have also suggested possible future directions for research on these issues. The summary is presented below.

On describing oral proficiency levels (proficiency scales)

The FSI/ILR/ACTFL scales, which were the first verbal descriptors of oral proficiency level, are neither theoretically nor empirically based. Many researchers have criticised their intuitive approach, lacking both theoretical and empirical underpinning. ACTFL recently revised their Proficiency Guidelines (1999). The observable changes are as follows: the revised guidelines are characterised by an increased amount of information; the new level descriptors are given in descending order; advanced level is now divided into High, Mid, and Low, which makes a total of 10 levels. Judging from these 1999 revisions, ACTFL has actually not made any major change in their approach. If we expect to see drastic changes in their guidelines, they would have to make drastic changes to their philosophy.

Many researchers agree that future proficiency scales must be empirically and theoretically based. Yet the problem is, that we have not yet developed a satisfactory model of proficiency. In the US, the ACTFL Proficiency Guidelines are still dominant but I observe some new and promising attempts to formulate empirically based scales in Europe (e.g. North 1995; North and Schneider 1998). These studies were part of the Council of Europe (CE) project of developing a *Common European Framework of Reference*. CE has carried out a thorough investigation of the issues relating to the formulation of a proficiency scale. CE also demonstrates an important quality in test development; that is accountability. The ongoing development process is open to scholarly discussion. This CE project is valuable in this respect as well.

There are other considerations in developing future rating scales. The content of a rating scale should be determined according to the purpose of its use and the context in which it may be used. ACTFL Guidelines are now directly translated, albeit somewhat woodenly, and used in many other language contexts. However, because

sociolinguistic requirements differ from culture to culture, it is impossible to assume that one common rating scale can serve for all language contexts. Another problem of the ACTFL Guidelines is that there are several assessment criteria used in judging a candidate's performance but the final rating is given as a unidimensional score. Not all features of speech performance can be measured by a numerical score. An alternative format of feedback must be considered which supplements the score with diagnostic descriptions of the particular features of the candidate's performance.

On assessing methods of oral proficiency

The OPI, which became the prototype for many assessment methods of oral proficiency, has been criticised from different directions. The early criticisms of the OPI concerned mainly the validity and reliability of the test. The OPI has also been criticised for the inauthentic nature of the interview method. More recent researches from the discourse/conversation analytic perspective show that the interaction of the OPI is not like a natural conversation.

There have been two new developments in exploring other assessment methods of oral proficiency. The first is the development of OPI variations: the Simulated Oral Proficiency Interview (SOPI) and the Computerized Oral Proficiency Interview (COPI). The COPI and SOPI need careful evaluation. Practicality should not be placed before validity, reliability, and authenticity. The second is attempts at alternative assessment, e.g. portfolio assessment, self-assessment, self-monitoring, and project assessment. Many LT researchers admit that any language test is, by its very nature, inauthentic language behaviour. The solution is to develop a non-test, alternative assessment, which is an ongoing process involving the student and teacher in making judgments about a given student's progress in a given language. Possible alternative assessment methods of oral proficiency should be explored.

In the field of the Japanese language education, *the Japanese Language Proficiency Test*[55], which is basically a discrete-point test consisting of 1) writing-vocabulary, 2) listening and 3) reading-grammar, is most commonly used[56] and the Japanese OPI has also become increasingly popular for oral proficiency

[55] 'The Japanese Language Proficiency Test' is administered by the Association of International Education (AIEJ) in cooperation with the Japan Foundation.

[56] The popularity of this Japanese Language Proficiency Test simply may have more to do with the availability of the test rather than its content. The Japan Foundation, a Japanese government organization, which is probably the equivalent of ACTFL in US, actively promotes this test and administers it in countries such as Korea, China, Taiwan, Indonesia, Singapore, Thailand, Philippines, Malaysia, Vietnam, Myanmar, Bangladesh, Nepal, India, Sri, Lanka, Pakistan, Egypt, Australia, New, Zealand, U.S.A., Canada, Mexico, Argentina, Paraguay, Brazil, Peru, Bolivia, Italy, United, Kingdom, Greece, Germany, France, Hungary, Turkey, Spain, Bulgaria, and Russia (AIEJ 2002 at http://www.aiej.or.jp/examination/jlpt_e.html).

testing, especially in the US in the last decade.[57] These tests have been adopted by many institutions especially in the Third World, simply because of the availability of the tests made by government-related organisations such as the Japan Foundation. Considering the intricate sociocultural rules embedded into the Japanese language, there should be indigenous development of oral proficiency tests, or at the very least the proficiency scale should be revised to fit the needs of the specific Japanese sociocultural context.

The field of oral proficiency testing was pioneered by the FSI in the 1950s and developed with their 'experientially-based approach'. Now the field has been enriched by subsequent theoretical investigation and empirical studies. There has been increasing collaboration between researchers in other related disciplines, particularly, sociolinguistics, psychometrics and SLA. Advances in both quantitative and qualitative approaches have enabled test developers to take more control of the variables in assessment.

Other significant progress is found in awareness of accountability. The Council of Europe is making their ongoing project open to scholarly debate. This kind of transparency is required in all sectors in society today and should be encouraged more in the field of language testing.

Finally LT researchers have acknowledged the limitations of performance testing and testing in general. We expect further developments in oral proficiency testing in the coming decade, with confidence in accomplishing what we can do, and with humility accepting what we cannot.

[57] A combination of the Japanese Language Proficiency Test and the OPI began to be used for the assessment of language proficiency of non-Japanese instructors of the Japanese language. See Yokoyama et al. 1998 at http://www.jpf.go.jp/j/learn_j/jedu_j/kiyou8/ ronbun6.html#no6 .

Appendix 1:
The original FSI proficiency levels (Clark and Clifford 1987:130-131)

Level 1. Elementary Proficiency: Able to satisfy routine travel needs and minimum courtesy requirement. Can ask and answer questions on topics very familiar to him; within the scope of his very limited language experience can understand simple questions and statements, allowing for slowed speech, repetition or paraphrase; speaking vocabulary inadequate to express anything but the most elementary needs; errors in pronunciation and grammar are frequent, but can be understood by a native speaker used to dealing with foreigners attempting to speak his language; while topics which are "very familiar" and elementary needs vary considerably from individual to individual, any person at [Level 1] should be able to order a simple meal, ask for shelter or lodging, ask and give simple directions, make purchase, and tell time.

Level 2. Limited Working Proficiency: Able to satisfy routine social demands and limited work requirements. Can handle with confidence but not with facility most social situations, including introductions and casual conversations about current events, as well as work, family, and autobiographical information; can handle limited work requirements, needing help in handling any complications or difficulties; can get the gist of most conversations on nontechnical subjects (i.e., topics which requires no specialized knowledge) and has a speaking vocabulary sufficient to express himself simply with some circumlocutions; accent, though often quite faulty, is intelligible; can usually handle elementary constructions quite accurately but does not have thorough or confident control of the grammar.

Level 3. Minimum Professional Proficiency: Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics. Can discuss particular interests and special fields of competence with reasonable ease; comprehension is quite complete for a normal rate of speech; vocabulary is broad enough that he rarely has to grope for a word; accent may be obviously foreign; control of grammar good; error never interfere with understanding and rarely disturb the native speaker.

Level 4. Full Professional Proficiency: Able to use the language fluently and accurately on all levels normally pertinent to professional needs. Can understand and participate in any conversation within the range of his experience with a high degree of fluency and precision of vocabulary; would rarely be taken for a native speaker, but can response appropriately even in unfamiliar situations; errors of pronunciation and grammar quite rare; can handle informal interpreting from and into the language.

Level 5. Native or Bilingual Proficiency: Speaking proficiency equivalent to that of an educated native speaker. Has complete fluency in the language such that his speech on all levels is fully accepted by educated native speakers in all of its features, including breadth of vocabulary and idiom, colloquialism, and pertinent cultural features.

References

ACTFL-ALC Press. 1996. *Standard Speaking Test Manual.* Tokyo: ACTFL-ALC Press.

Alderson, J. C. 1991. 'Bands and Scores'. In J. C. Alderson and B. North, eds. *Language Testing in the 1990s: The Communicative Legacy.* London: Macmillan, 71-86.

American Council on the Teaching of Foreign Languages. 1988. *ACTFL Proficiency Guidelines*. Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

American Council on the Teaching of Foreign Languages. 1999. *ACTFL Gengo Unyoo Nooryoku —Wa-Ginoo1999 Revised (ACTFL Proficiency Guidelines* –Speaking revised 1999; Japanese version). Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

American Council on the Teaching of Foreign Languages. 2002. 'Proficiency Testing.' in *ACTFL.* Available from http://www.actfl.org/ (Accessed 7 November, 2002).

Association of International Education (AIEJ). *Guide to the 2002 Japanese Language Proficiency Test (administered in Japan).* Available from http://www.aiej.or.jp/examination/jlpt_e.html (Accessed 7 November, 2002).

Austin, J. L. 1962. *How To Do Things With Words.* Cambridge, Mass: Harvard University Press.

Bachman, L. F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. 1988. 'Problem in examining the validity of the ACTFL Oral Proficiency Interview.' *Studies in Second Language Acquisition* 10(2): 149-163.

Bachman, L. F. and Cohen, A. D. 1998. *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press.

Bachman, L. F. and Eignor, D. R. 1997. 'Recent advances in quantitative test analysis.' In C. Clampham and D. Carson, eds. Encyclopedia of Language and Education, Vol.7: Language and Assessment, 227-242. Dordrecht: Kluwer Academic.

Bachman, L. F. and Palmer, A. S. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

Bachman, L. F. and Palmer, A. S. 1981. 'The construct validation of the FSI oral interview.' *Language Learning* 31: 67-86.

Bachman, L. F. and Savignon, S. 1986. 'The evaluation of communicative language proficiency: A Critique of the ACTFL Oral Interview.' *The Modern Language Journal* 70(4): 380-390.

Banerjee, J. and Luona, S. 1997. 'Qualitataive Approaches to test validation.' In C. Clampham and D. Carson, eds. Encyclopedia of Language and Education, Vol.7: Language and Assessment, 275-287. Dordrecht: Kluwer Academic.

Barnwell, D. P. 1996. *A History of Foreign Language Testing in the United States: from its beginnings to the present.* Tempe, Arizona: Bilingual Press.

Barnwell, D. P. 1989. "Naïve" native speaker and judgements of oral proficiency in Spanish. *Language Testing* 6(2): 152-163.

Barnwell, D. P. 1987. 'Oral proficiency testing in the United States.' *British Journal of Language Teaching* 25(1): 35-42.

Brennan, R. L. 1983. *Elements of Generalizability Theory.* Iowa City, IA: The American College Testing Program.

Breiner-Sanders, K., Lowe. P. Jr., Miles, J. and Swender, E. 2000. 'ACTFL Proficiency Guidelines –Speaking Revised 1999.' *Foreign Language Annals* 33(1): 13-17.

Brindley, G. 1989. *Assessing Achievement in the Learner-centred Curriculum*, NCELTR Research Series. Sydney: Macquarie University, 112-140.

Brindley, G. 1998. 'Describing language development? Rating scales and SLA.' In L. F. Bachman and A. D. Cohen, *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press.

Buck, K. ed., Byrnes, H. and Thompson, I. contributing eds. 1989 *ACTFL Oral Proficiency Interview: Tester Training Manual.* Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

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

Carroll, J. B. 1968. 'Language testing.' In A. Davies, ed. *Language Testing Symposium: A Psycholinguistic Approach*. London: Oxford University Press, 46-49.

Center for Applied Linguistics (CAL1). *Computerized Oral Proficiency Interview.* Available from http://www.cal.org/projects/copi.html (Accessed 7 November, 2002).

Center for Applied Linguistics (CAL2). 'Simulated Oral Proficiency Interviews.' In *Foreign Language Test Development (Speaking Tests)*. Available from http://www.cal.org/tests/fltests.html (Accessed 7 November, 2002).

Chalhoub-Deville, M 1995. 'Deriving oral assessment scales across different tests and rater groups.' *Language Testing* 12(1): 16-33.

Clark, J. L. D, and Clifford, R. T. 1988. 'The FSI/ILR/ACTFL proficiency scales and testing techniques: development, current status, and needed research.' *Studies in Second Language Acquisition* 10(2): 129-147.

Dandonoli, P and Henning, G. 1990. 'An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure.' *Foreign Language Annals* 23(1): 11-22.

Davies, A. and Brown, A., Elder, C., Hill, K., Lumley, T., McNamara, T. 1999. *Dictionary of Language Testing* Cambridge: Cambridge University Press.

Fulcher, G. 1996. 'Invalidating validity claims for the ACTFL Oral Rating Scale.' *System* 24(2): 163-172.

Fulcher, G. 1997. 'The testing of speaking in a second language.' In C. Clampham and D. Carson eds. *Encyclopedia of Language and Education*, Vol.7: Language and Assessment, 75-85. Dordrecht: Kluwer Academic.

Gumperz, J. J. and Hymes, D. 1964. *Ethnography of Communication*. Washington: American Anthropological Association.

Hadden, B. L. 1991. 'Teacher and nonteacher perceptions of second-language communication.' *Language Learning* 41(1): 1-24.

Hancock C. R. 1994. 'Alternative Assessment and Second Language Study: What and Why?' *ERIC Clearinghouse on Languages and Linguistics Digest* July 1994. Available from http://www.cal.org/ericcll/digest/hancoc01.html (Accessed 7 November, 2002).

Henning, G 1992. 'The ACTFL oral proficiency interview: validity evidence.' *System* 20: 365-72.

Henning, G. 1990. 'An investigation of the construct validity of the ACTFL Proficiency Guidelines and oral interview procedure.' *Foreign Language Annals* 23(1): 11-22.

Hymes,D. H. 1972. 'On communicative competence.' In J. B. Pride and J. Homes, eds.

Sociolinguistics: Selected Readings, 269-293. Harmondsworth, Middlesex; Penguin. Mar Ban PRI

Johnson, M. and Tyler, A. 1998. 'Re-analyzing the OPI: How much does it look like natural conversation?' In R. Young, and A. W. He, eds. *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency.* Amsterdam/Philadelphia: 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(2): 163-188.

Kramsch, C. J. 1986. 'From Language Proficiency to Interactional Competence.' *The Modern Language Journal* 70(4): 366-372.

Kunnan, A. J. 1999. 'Language testing: fundamentals.' In B. Spolsky, ed. *Concise Encyclopedia of Educational Linguistics.* Oxford: Elsevier Science.

Kuo, J. and Xixiang, J. 1997. 'Assessing the assessments: the OPI and the SOPI.' *Foreign Language Annals* 30(4): 503-512.

Lantolf, J. P. and Frawley, W. 1985. 'Oral proficiency testing: a critical analysis.' *Modern Language Journal* 69(4): 337-345.

Lantolf, J. P. and Fawley, W. 1988. 'Proficiency: Understanding the Construct.' *Studies in Second Language Acquisition* 10(2): 181-195.

Lazaraton, A 1992. 'The structural organization of a language interview: a conversation analytic perspective.' *System* 20: 373-386.

Lewkowicz. J. A. 2000. 'Authenticity in language testing: some outstanding questions.' *Language Testing* 17(1): 43-64.

Linacre, J. M. 1989. *Many-faceted Rasch Measurement.* MESA Press, Chicago, IL.

Liskin-Gasparro, J. E. 1984. 'The ACTFL Guidelines: A historical perspective.' In T.V. Higgs, ed. *Teaching for Proficiency: The organizing principle,* 11-42. Lincolnwood, IL: National Textbook Co.

Malabonga, V. 2000. 'Trends in Foreign Language Assessment: The Computerized Oral Proficiency Instrument (COPI)' The NCLRC Language Resource 4(1). Available from http://www.cal.org/nclrc/caidlr41.htm#BM8 (Accessed 7 November, 2002).

McNamara, T. 1996. *Measuring Second Language Performance.* London: Longman.

McNamara, T. 1997. 'Performance Testing.' In C. Clampham and D. Carson eds. *Encyclopedia of Language and Education, Vol.7: Language and Assessment,* 131-139. Dordrecht: Kluwer Academic.

Nevo, D. and Shohamy, E. 1986. Evaluation standards for the assessment of alternative testing method: An application. *Studies in Educational Evaluation* 5:149-158.

North, B. 1995. 'The development of a common framework scale of descriptors of language proficiency based on a theory of measurement.' *System* 23(4): 445-465.

North, B. and Schneider G. 1998. 'Scaling descriptors for language proficiency.' *Language Testing* 15(2): 217-262.

Omaggio, A. C.1983. 'Methodology in transition: the new focus on proficiency.' *The Modern Language Journal* 67(4): 330-341.

Pienemann, M. and Johnson, M., and Brindley, G. 1988. 'Constructing an acquisition-based procedure for second language assessment'. *Studies in Second Language Acquisition* 10: 217-234.

Raffaldini, T. 1988. 'The Use of Situation Tests as Measures of Communicative Ability.' *Studies in Second Language Acquisition* 10(2): 197-216.

Reve, T. 1991. 'From testing research to educational policy: a comprehensive test of

oral proficiency.' In J. C. Alderson and B. North, eds. *Language Testing in the 1990s: The Communicative Legacy.* London: Macmillan.

Rivera C ed. 1983. *An Ethnographic/Sociolinguistic approach to language proficiency assessment.* Cleveland, OH: Multilingual Matters.

Ross, S. 1992. 'Accommodative questions in oral proficiency.' *Language Testing* 9(2): 173-186.

Ross, S. 1998. 'Divergent frame interpretations in language proficiency interview interaction.' In R. Young, and A. W. He, eds. *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

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

Salaberry, R. 2000. 'Revising the revised format of the ACTFL Oral Proficiency Interview.' *Language Testing* 17(3):289-310.

Saville-Troike, M. 1989. *The Ethnography of Communication: An Introduction.* Oxford: Basil Blackwell.

Searle, J. R. 1965. 'What is a speech act.' In M. Black ed. Philosophy in America. London: Allen and Unwin.

Searle, J. R. 1981. *Speech Acts (2nd ed.).* London: Cambridge University Press.

Seliger, H. W. 1985. 'Testing authentic language: the problem of meaning.' *Language Testing* 2 (1): 1-15.

Shavelson, R. J. and Webb, N. M. 1991. *Generalizability Theory: A Primer.* Sage: Newbury Park, CA.

Shohamy, E. 1998. 'How can language testing and SLA benefit from each other? The case of discourse.' In L. F. Bachman and A. D. Cohen. 1998. *Interfaces between Second Language Acquisition and Language Testing Research.* Cambridge: Cambridge University Press.

Shohamy, E. 1988. 'A proposed framework for testing the oral language of second/foreign language learners.' *Studies in Second Language Acquisition* 10(2): 165-179.

Shohamy, E. 1990. 'Language Testing Priorities: A Different Perspective'. *Foreign Language Annals.* 23, 385-394.

Shohamy, E. 1994. 'Validity of Direct Versus Semi-Direct Oral Tests.'*Language Testing* 11: 99-123.

Shohamy, E. and Reves, T. 1985. 'Authentic language test: where from and where to?' *Language Testing* 2: 48-59.

Spolsky, B. 1989. *Conditions for Second Language Learning.* Oxford: Oxford University Press.

Spolsky, B. 1977. 'Language Testing: art or science in Proceedings of the Fourth International Congress of Applied Linguistics.' Stuttgart: Hoschschulverlag: 7-28.

Spolsky, B. 1995. Measured Words: the development of objective language testing Oxford:Oxford University Press.

Spolsky, B. 1981 'Some ethical questions about testing.' In C. Klein-Braley, and D. Stevenson, eds. *Practice and Problems in Language Testing.* Frankfurt-am-Main: Peter D. Lang.

Spolsky, B. 1985. 'The limit of authenticity in language testing.' *Language Testing* 2(1): 31-40.

Stern, H.H. 1983. *Fundamental Concepts of Language Teaching.* Oxford: Oxford University Press

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

Stevenson, D. K. 1985. 'Authenticity, validity and a tea party.' *Language Testing* 2 (1): 41-48.

Swender, E., Breiner-Sanders, K., Mujica-Laughlin, L., Lowe. P., and Miles, J. 1999. *ACTFL Oral Proficiency Interview Tester Training Manual.* Hasting-on-Hudson, NY: The American Council on the Teaching of Foreign Languages.

Swender, E., Breiner-Sanders, K., Mujica-Laughlin, L., Lowe. P., and Miles, J. *1999. ACTFL Oral Proficiency Interview Tester Training Manual Japanese Edition.* Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

The Europe of Cultural Co-operation (ECC1). *European Language Portfolio.* http://culture2.coe.int/portfolio/inc.asp?L=E&M=$t/208-1-0-1/main_pages/contents_portfolio.html (Accessed 7 November, 2002).

The Europe of Cultural Co-operation (ECC2a). *Modern Languages: Learning, Teaching Assessment. A Common European Framework of Reference* Strasbourg. http://culture2.coe.int/portfolio/documents_intro/common_framework.html (Accessed 7 November, 2002).

The Europe of Cultural Co-operation (ECC2b). 1996. "Scaling and Levels." in *Modern Languages: Learning, Teaching Assessment. A Common European Framework of Reference* Strasbourg. http://www.rimini.com/provveditorato/didattica/saperi/linguestraniere/links/quadro_europeo/inglese/paragrafi/eedu2_4i.htm (Accessed 7 November, 2002).

Turner, C. and Upsher, J. 1995. 'Constructing rating scales for second language tests.' *ELT Journal* 49(1): 3-12.

Valdman, A. 1988. 'Introduction'. *Studies in Second Language Acquisition* 10(2): 121-128.

van Lier, L. 1989. 'Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency as conversation.' *TESOL Quarterly* 23 (3): 489-508.

Weir, C. J. 1988. *Communicative Language Testing: With Special Reference to English as a Foreign Language.* Exeter: University of Exeter. 1990. London: Prentice Hall.

Yokoyama, N., Kitani, N. and Yanashima, F. 1998. 'Proficiency analysis of non-native Japanese language teachers.' Japan Foundation Japanese-Language Institute, Urawa Bulletin 8 (March, 1998). http://www.jpf.go.jp/j/learn_j/jedu_j/kiyou8/ronbun6.html#no6 (Accessed 7 November, 2002).

Young, R. 1995. 'Conversational Styles in Language Proficiency Interviews.' *Language Learning* 45: 3-42.

Young, R. and He, A. W. eds. 1998. *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Young, R. and Milanovic, M. 1992. 'Discourse Variation in Oral Proficiency Interviews.' *Studies in Second Language Acquisition* 14: 403-424.

Addendum