Revising the revised format of the ACTFL Oral Proficiency Interview

Rafael Salaberry *Rice University*

Since the early 1980s proponents of proficiency examinations such as ACTFL (American Council on the Teaching of Foreign Languages) have been criticized for the low validity and reliability of tests such as the OPI (Oral Proficiency Interview). Despite these strong concerns, the most recent edition of the ACTFL Tester Training Manual (Swender *et al.*, 1999) does not reveal substantial changes from the previous manual published in (Buck *et al.*, 1986). While a complete elimination of proficiency tests such as the ACTFL-OPI may be neither feasible nor necessary, some practical changes may be appropriate. In this article I describe minor structural changes of the ACTFL-OPI framework that would not eliminate the practical benefits of a proficiency test, but which address substantial concerns related to the validity and reliability of the instrument.

I Introduction

Several universities in the United States have established second language (L2) proficiency requirements in order to overcome the problems attributed to L2 curricula that are based solely on the so-called 'seat-time requirement'. This term refers to the fact that students can satisfy language proficiency requirements simply by being 'seated' in the required course sequence, which typically consists of three or four semesters of instruction. Proficiency tests – generally based on the format established by ACTFL (American Council on the Teaching of Foreign Languages) – are intended to overcome the deficiencies associated with the attainment and evaluation of proficiency through course work only (e.g., Barnes *et al.*, 1990; Villar and Meuser-Blincow, 1993). For instance, Barnes *et al.* state that a proficiency requirement at the University of Minnesota was implemented in response to a number of factors that negatively affected the teaching of language courses. The authors cite factors, such as student apathy, lack of funding for courses with increasingly more students per section, TA job alienation and the increasing work load of language course supervisors. Barnes *et al.* claim that the ACTFL proficiency

Address for correspondence: Rafael Salaberry, Department of Hispanic Studies, Rice University, 6100 Main Street, Houston, TX 77251–1892, USA; email: salaberry@rice.edu

Language Testing 2000 17 (3) 289–310 0265-5322(00)LT183OA

guidelines were selected because 'despite their shortcomings, they provide a comprehensive set of criteria that nonspecialists can easily understand' (1990: 36).

It is my contention, however, that performance tests as currently represented in the ACTFL-OPI (American Council on the Teaching of Foreign Languages – Oral Proficiency Interview) may not adequately address the basic concern brought about by the perceived shortcomings of academic L2 programs.[1] In this article I support this argument with a critical analysis of the ACTFL Proficiency Guidelines for Speaking (ACTFL, 1986; 1999) and the ACTFL-OPI Tester Training Manual (Buck *et al.*, 1986; Swender *et al.*, 1999; henceforth, these are referred to as ACTFL-TTM, 1986 and 1999, respectively). The analysis makes reference to both the old editions of the Guidelines and the Manual published in 1986 and the newly revised ones published in 1999.[2] Despite the argument that the revision of the ACTFL Guidelines and the Tester Training Manual follow a 'process of revision [of the definition of language abilities] ... usually occasioned by new or expanded insights' (Lowe, 1998: 358) few substantial changes are noticeable in the revised edition of 1999. In fact, as will be illustrated in the following sections, both the Guidelines and the Tester Training Manual (TTM) have preserved, in almost verbatim format, the basic tenets of the OPI test proposed in 1986. On the other hand, I intend to show that the OPI may be substantially improved by way of incorporating a limited number of changes that do not necessarily affect the basic tenets that led ACTFL originally to propose and develop the OPI.

The sections that follow focus on the three major factors that are necessary for a thorough evaluation of performance tests of L2 proficiency:

1) the assessment of professional standards and accountability, social and educational values and legal consequences;
2) the analysis of the theoretical construct of proficiency as represented in the ACTFL-OPI; and
3) the identification and evaluation of aspects of the ACTFL-OPI; that may be improved following a critical assessment of (1) and (2) above.

---

[1] In fact, some proponents of ACTFL tests argue that the fulfillment of proficiency requirements may be addressed fruitfully through course work only if the course objectives are adequately implemented (Alice Omaggio, personal communication, November 1997). Furthermore, Brindley (1998: 48) comments that some descriptors of outcomes-based assessment projects 'are similar to those found in language-proficiency rating scales'.

[2] I have received official training through ACTFL to elicit speech samples and rate them according to the standards established in the Guidelines and the Tester Training Manual.

The article is structured as follows. In Section II I assess the importance of consequential validity within the larger framework of standards and accountability, ethical values and legal consequences. In Section III I analyze how the constructs of proficiency and communicative language ability are embodied in performance tests such as the ACTFL-OPI. In Section IV I identify several components of the ACTFL-OPI that are amenable to substantial changes, but that do not necessarily require major restructuring of the basic ACTFL framework. Finally, in Section V, I highlight the importance of the above-mentioned proposal within the larger scope of the quality of college-level L2 requirements.

II Accountability, validity and ethical values

1 Accountability and practicality

The use of the ACTFL proficiency test has been justified along the lines of accountability and practicality. Firstly, it is important to point out that the effort of ACTFL to improve foreign language learning in the USA has increased the level of awareness of language educators about performance tests. For instance, Shohamy (1990: 385) states that ACTFL has been 'successful in drawing attention to goals, standards, and accountability'. Similarly, De Jong (1995) claims that ACTFL has shifted the attention of teachers and learners from the written mode to the spoken mode, and, in so doing, has focused their attention on the communicative demands of proficiency testing. Most important, Bachman and Savignon (1986: 380) note that 'guidelines for measuring language proficiency can enhance accountability and strengthen the profession'. Finally, Bachman (1988: 160) emphasizes the potential for improvement of the basic underlying proposal: '[I]s the ACTFL oral interview a valid measure of communicative language ability? As it is currently designed and used, I believe it is not. This is not to say, however, that it cannot be'.

Second, supporters of the ACTFL-OPI test consistently make a valid point: there is no practical alternative to proficiency testing. Pienemann *et al.* (1993: 500) state that linguistic profiling and proficiency approaches to language testing have different objectives:

> [T]he opposition between proficiency-oriented approaches to language testing and linguistic profiling exists at the level of construct validity. At the practical level, however, these approaches are designed for very different purposes . . . most proficiency-oriented approaches are designed to capture the global picture of a person's ability in a language. Approaches to profiling are currently unlikely to be able to achieve that objective.

In fact, some researchers accept the practical benefits of a performance test such as the OPI, as long as its drawbacks are acknowledged. For instance, van Lier (1989: 501; italics added) argues that it is 'possible to sidestep the issue of construct validity altogether and be satisfied with measuring *whatever oral language use happens to be elicited by the OPI*, since it is in any case the best instrument available'. Hence, because the OPI test may be the only practical instrument available, it is important to understand the limitations of a performance test in order to assess the intended and unintended consequences of the outcome of the OPI.

2 Validity and reliability: legal and ethical consequences

L2 performance tests have been portrayed as real-life direct tests of L2 ability. However, Messick (1994: 14) states that 'the portrayal of performance assessments as authentic and direct has all the earmarks of a validity claim but with little or no evidential grounding'. Accordingly, many researchers have argued against the validity of the ACTFL guidelines due to the lack of theoretical and empirical support (e.g., Lantolf and Frawley, 1985; 1988; 1992; Savignon, 1985; Bachman and Savignon, 1986; Kramsch, 1986; Raffaldini, 1988; Valdman, 1988; Bachman, 1990; Shohamy, 1990). For instance, Shohamy (1990: 386) points out that tests such as the ACTFL-OPI are limited in scope in terms of assessment of developmental stages as well as communicative interaction:

> [T]he [ACTFL] guidelines reduce language to a set of simplistic descriptions which are believed to reflect and represent language proficiency and its stages of development. The guidelines cannot fit all situations, all purposes, all levels, all languages and there is no empirical research to support their descriptions.

Even proponents of the ACTFL-OPI testing procedure have recognized its lack of theoretical and empirical support and share some of the concerns voiced by ACTFL's harshest critics. In effect, Dandonoli and Henning (1990: 11) acknowledge that 'the most significant' criticism against the use of the ACTFL-OPI is that there is no study that supports the test's validity.

Indeed, issues of validity and reliability are non trivial. Messick (1994: 13) claims that 'test validity and social values are intertwined and that evaluation of intended and unintended consequences of any testing is integral to the validation of test interpretation and use'. Similarly, Moss (1994) points out that the concern over reliability relates directly to epistemological (e.g., degree of generalization) and ethical (e.g., fairness) issues. For these reasons, several researchers have questioned the *institutionalization* of the ACTFL proficiency

guidelines as an assessment instrument due to the lack of an appropriate theoretical or empirical foundation. Most important, the lack of theoretical support for the type of language testing conducted by any institution – especially regarding validity and reliability – may have legal ramifications. According to Shohamy (1988: 178), the evaluation of the theoretical assumptions of oral language tests, as well as their development and their implementation have 'special significance nowadays, when test scores are subject to court challenges'. As a consequence, Shohamy claims that 'scientific/empirical information is needed to justify the scores assigned to test takers'. More dramatically, Bachman (1988: 161) asserts that the consequences of ignoring the finding that a rating is valid both as an indicator of proficiency as well as for the purpose of making specific decisions 'can be found in the annals of our judiciary system'. Shohamy (1990: 391), for instance, reports on the case of Debra Turlington who obtained the ruling from a United States Court of Appeals that 'a graduation test must possess content validity'.

3 Face validity

Face validity has been cited as a primary justification for the OPI (e.g., Shohamy, 1988; van Lier, 1989; Dandonoli and Henning, 1990). In other words, the OPI appears, to both testers and test-takers, to assess communicative language ability in a realistic situation. For instance, Dandonoli and Henning (1990) justify their claim that the OPI has a high level of face validity by signalling that, in their study, the ratings of naive native speakers correlated highly with the ratings of ACTFL-trained raters (0.934 for English samples and 0.929 for French samples). Ironically, however, this high level of face validity raises serious questions about the practical use of ACTFL training workshops for OPI testers and raters. In other words, if naive native speakers are able to judge speech samples as well as any ACTFL-trained teacher, why waste resources in training and standardization workshops when an untrained native speaker can do the same job? In this regard, several questions are relevant:

- What are the criteria that native speakers use to arrive to those judgements?
- Do their criteria differ from the ones used by the ACTFL-trained teachers? and
- If they do differ, how come the ratings of both groups coincide?

These questions highlight the fact that we may not know exactly what is behind the OPI ratings. In other words, we may lack evidence to justify the particular rankings assigned to test-takers, and we may,

therefore, be unable to defend these rankings from either ethical or legal perspectives. Crucially, however, 'face validity by itself cannot provide proof of validity' (Shohamy, 1990: 386).

III The theoretical construct of proficiency

In the ACTFL-OPI test 'language proficiency itself is not defined, but rather, a domain of actual, or "real-life" language use is identified that is considered to be characteristic of the performance of competent language users' (Bachman, 1990: 41). Messick (1994), however, points out that proponents of performance assessment sometimes confuse performance per se with performance as a vehicle of assessment. Examples of the former are 'an arts contest or an Olympic figure-skating competition or a science fair' (p. 14). In such cases 'inferences are not to be made about the competencies or other attributes of the performers, that is, inferences from observed behavior to constructs such as knowledge and skill underlying that behavior' (p. 14). On the other hand, proponents of the oral proficiency interview insist that 'proficiency is . . . like judging a gymnast's routine on the parallel bars, where a score is based on the impressions of an observer and is in keeping with a set of conventions and standards' (Hagen, 1990: 50). In fact, according to the ACTFL-TTM (1986) 'some shortcomings may reflect performance errors rather than flaws in the underlying competence of the speaker' (pp. 2–9). But, how should we distinguish competence from performance? In this respect, the 'real-life' approach advocated by defenders of the ACTFL-OPI test misses the point that mental abilities are not directly observable, 'but must be inferred on the basis of observed performance' (Bachman, 1990: 256). Not surprisingly, then, Shohamy (1996) questions the neglect of the important distinction between competence and performance among defenders of the OPI.

Pursuing such important criticisms of the OPI, Bachman (1990) argues that the development of language tests should follow a sequence of three steps:

1) identification and definition of the theoretical construct;
2) operational definition of the construct; and
3) the identification of rules and procedures to quantify observations.

Skipping the first step of construct validation (i.e., empirically testing hypotheses of the relationship between test scores and proposed abilities) is appealing for practical reasons, because one need not contend with a yet underdeveloped theory of L2 acquisition and competence. This step, however, is essential because it distinguishes 'the

construct we wish to measure from other similar constructs by defining it clearly, precisely and unambiguously' (p. 41). In the Sections 1 and 2 below, I address the need:

1) to identify the theoretical construct of L2 proficiency; and
2) to establish a clear set of guidelines to measure global oral performance.

1 Communicative language ability

One problem associated with the lack of explicit identification and description of the theoretical construct of L2 proficiency is that factors beyond language proficiency per se may become part of the operational construct under analysis (i.e., construct-irrelevant variance). For instance, it is plausible that high levels of strategic competence in the management of conversational interaction may help some learners obtain higher ratings than other learners with similar language proficiency but less skill in managing conversational exchanges. Bachman (1990: 105), for example, claims that in tests such as a picture description task 'it may well be that performance is affected more by strategic competence than by the specific language ability the test was originally intended to measure'. Furthermore, even if the learner were skilful in the management of conversation, he or she may consider such non-linguistic factors irrelevant for the test outcome (e.g., focus on the command of inflectional morphology versus strategies for turn-taking). Bachman (1990: 106) explicitly states that:

> rather than considering strategic competence solely an aspect of language competence, I consider it more as a general ability, which enables an individual to make the most effective use of available abilities in carrying out a given task.

In essence, Bachman argues that the management of conversational exchanges should lie outside the realm of 'L2 proficiency'. In contrast, Savignon (1985: 131) specifies that 'communicative competence certainly requires more than knowledge of surface features of sentence-level grammar. And educated native-speaker grammatical competence is not necessary for communication' (see also Skehan, 1998: 159–67).

Clearly, the identification and operationalization of the theoretical construct that we want to measure requires theoretical consistency and carries important ethical and political consequences. For instance, if we demand that non-native speakers be skilful in the managing of conversational interaction in the target language, should we not expect the same from native speakers? This is problematic because we know that native speakers are not necessarily good at the strategic manage-

ment of conversation. In fact, the opposite argument could be made, since nonnatives may become accustomed to managing few linguistic resources in demanding communicative situations. It is popular wisdom that some non-native speakers are far better than native speakers in precisely the strategic management of conversational interaction, while deficient in strictly linguistic modules (e.g., Henry Kissinger as a skilful political negotiator and debater despite a noticeable German accent). Alternatively, if we decide to maintain the above mentioned non-linguistic components of conversational exchanges in our notion of proficiency, what, then, is the theoretical concept (construct) of proficiency that serves as the yardstick (given that the native speaker norm will no longer be the target)? The explicit identification of our theoretical construct may have far-reaching consequences, particularly if we use such constructs for the development of tests that decide practical aspects of our lives such as certification in language proficiency for job search or job promotion.

2 Reliability

The ACTFL-TTM (1986; 1999) does not provide a comprehensive description/analysis of the various factors that can affect the assessment of interview data by human raters: variation in topics discussed (e.g., Young, 1995), raters who act as test interlocutors versus raters with access to tape-recordings only (e.g., McNamara and Lumley, 1997), face-to-face versus telephone interviews (e.g., Thompson, 1995), interlocutor variables such as linguistic competence or rater severity (e.g., McNamara, 1996; Lynch and McNamara, 1998) and individual strategies (e.g., Purpura, 1998). Hence, inconsistent ratings within and across raters (i.e., intra-rater and inter-rater reliability) can be traced back to the lack of a detailed set of criteria that specifies the expected factors that may have an impact on the assessment of proficiency. For instance, Bachman (1990) states that:

> [P]ractically anyone can rate another person's speaking ability . . . but while one rater may focus on pronunciation accuracy, another may find vocabulary to be the most salient feature. Or one rater may assign a rating as a percentage, while another might rate on a scale from zero to five . . . the different raters in this case did not follow the same criteria or procedures for arriving at their ratings.

The above-mentioned problem appears to be more acute when speech samples lie on the borderline between levels and even more so when administrative decisions are based on pre-specified cut-off points. In such cases it appears that decisions may rely on some specific components of the scale, particularly accuracy. For instance, van Lier (1989: 494; italics added) argues that:

raters of OPI in practice (that is, aside from whatever recommendations are made in training programs and rating scales) come *to rely heavily on certain criterial linguistic features, often of a very discrete nature*, especially when decisions to cross bands or levels are at stake.

Contrary to this claim, Halleck (1992: 228) claimed that ACTFL raters tend to be 'primarily concerned with communicative strategies rather than with the grammatical accuracy of the interviewees'. Halleck's claim is based on the analysis of a questionnaire about the justification that raters provided for their ratings. We should bear in mind, however, that some of the communicative factors mentioned by Halleck may also be interpreted, in practice, as discrete grammatical features (e.g., hypothesizing, narrating, connected discourse).

In summary, the reliability of the test of oral proficiency may be compromised if we are not able to distinguish the abilities we attempt to measure from other factors. If human raters are required to make global decisions on performance, it is essential that the target ability be properly identified (i.e., as an explicit theoretical construct). In turn, the explicit identification of the target ability must be accompanied by a set of explicit procedures that guide human raters in the assessment of speech samples.

IV Areas of improvement

The previous discussion highlights an important dilemma for L2 testers. On the one hand, research in L2 acquisition has not yet provided language testers with a practical alternative to the ACTFL-OPI (e.g., Pienemann *et al.*, 1993; van Lier, 1989). On the other hand, the underlying framework of performance tests such as the ACTFL-OPI is weakened by the lack of specification of construct validity and the absence of a detailed set of criteria to measure global oral performance. Given that a comprehensive solution to this problem will take time to become available, it is important to identify and describe areas of improvement within the current ACTFL-OPI framework. In the following sections I identify some of the most crucial problems in the ACTFL-OPI test and I describe some modifications that could overcome these deficiencies.

Before describing the proposed changes, however, it is important to point out that the 1999 revision of the ACTFL-TTM has not introduced substantial changes to the 1986 edition. Firstly, a cursory look at the two volumes reveals that the seven chapters of the original edition have been maintained in the new one with two minor exceptions:

1) The chapter on the elicitation of role plays of the old edition has

been subsumed under the one that explains the structure of the OPI in the new edition; and

2) The section that describes the ACTFL Proficiency Guidelines for Speaking, originally placed in an appendix in the old edition, has become a new chapter in the revision.

Second, and more importantly, the content of the chapters remains largely unchanged except for very minor modifications related to the internal organization of paragraphs and some other editorial changes. For example, a quick review of Chapter 1, titled 'What is the oral proficiency interview' in both editions, reveals that the information on pages 1–1 and 1–3 of the old edition are maintained in almost verbatim format on pages 2 and 3 of the revised edition. It is important to point out, however, that Chapter 1 in the revised edition does include some additional information such as:

1) a new section on how the ACTFL-OPI relates to the recently proposed National Standards of Foreign Language Education (p. 6);
2) a statement on the purported validity and reliability of the ACTFL-OPI (p. 4); and
3) a statement on the purported interactive, adaptive and learner-centred nature of the ACTFL-OPI as a testing instrument (p. 4).

In the following sections I refer the reader to page numbers in both ACTFL-TTM (1986) and in ACTFL-TTM (1999) that substantiate the above-mentioned claim and, consequently, justify the analysis proposed in this article.

1 Selection of tasks

The ACTFL-TTM (1986, 1999) provides OPI testers with a standardized procedure for the OPI interview. Standardization is deemed necessary 'since, to assure reliability in assessing different speech samples a prescribed procedure must be observed' (1986: 1–1).[3] Furthermore, the ACTFL-TTM is specific about the conversation format/content used to obtain speech samples: 'the OPI should resemble, to the greatest extent possible, a natural conversation' (1986: 1–2; 1999: 4). I believe, however, that there are three problems with this procedure. Firstly, the type of linguistic dialogue portrayed in a typical OPI interview (predicated on the exchange of questions

[3] Young (1995) claims that, in relative terms, the ACTFL-OPI provides language testers with fewer constraints to elicit spontaneous data than similar performance tests such as the Cambridge First Certificate Examination.

and answers, like interviews) is not typical of a 'natural conversation'. Van Lier (1996: 175) argues that 'interaction is conversational to the extent that it is oriented towards symmetrical contributions'. By contrast, OPI interviews are power-laden and, by definition, not symmetrical in terms of conversational interaction. In this respect, Lazaraton (1992; 1996) and van Lier (1989; 1996) offer an extended analysis of the theoretical and practical importance of this issue.[4]

Second, if the OPI is going to be used to assess 'language performance in terms of ability to use the language effectively and appropriately in real-life situations' (1986: 1–1; 1999: 1), then a wide range of interaction formats are essential to represent such real-life situations. However, one of the clearest weaknesses of the OPI test is the fact that a limited range of tasks is assessed. For instance, Shohamy *et al.* (1986) argue that discussions, reports, interviews and conversations are all different. In view of this, a particular test will only tap a sample of the features of overall communicative language ability. Similarly, Savignon (1985: 132) claims that 'among the many contexts *not* sampled [by the ACTFL-OPI test] are small-group discussion, playing a game, or conducting a survey'. Savignon emphasizes the importance of these contexts because they require 'very different discourse strategies, strategies that teachers often encourage, or would like to encourage, in their classrooms'.

Third, the ACTFL guidelines address performance in 'general rather than job-specific language' (Lowe, 1986: 393). This runs counter to the importance of contextual features of interaction (i.e., job-specific language) to language learners, who typically have limited access to environments other than the classroom setting. In fact, Kramsch (1987: 358) questions the validity of proficiency ratings 'that are supposed to predict performance across conversational contexts, interlocutors, topics and purposes'. Similarly, Bachman and Savignon (1986: 386) argue that the ACTFL Guidelines offer the 'apparent claim' that such Guidelines 'are based on the actual needs of language learners'. But, Bachman and Savignon wonder: 'What language learners? . . . Do language majors, for example, have the same goals as students who are majoring in other areas?' In order to address this issue, Henning (1990: 382) pleads that 'we can and should better articulate a taxonomy of contextualization features to guide matters such as choice of reading and listening passages, setting of test administration, topic of writing prompts, and appropriate use

[4]Incidentally, the ACTFL-TTM explicitly acknowledges that 'the OPI format as well as the intent to assess language use inherently introduces a power relationship in favor of the interviewer into the interview' (pp. 5–3).

of role play activities in language assessment'. Bachman and Savignon (1986: 388) suggest, as a practical alternative, that raters be provided with 'a list of content areas and contexts that might be useful in eliciting samples of speech for rating'.

In sum, the three problems of the OPI interview mentioned above (i.e., lack of features of conversational interaction, limited range of interactional contexts and lack of specification of content areas to be addressed) can be avoided if:

1) we broaden the scope of interactional formats represented in oral performance tests; and
2) we incorporate ways to delimit the content areas that classroom learners are expected to know.

A possible solution for this dilemma is the extended use of simulations such as role plays (Di Pietro, 1989), because they 'can approximate the appearance, form and effect of an authentic situation' (Shohamy, 1988: 172). Along the same lines, Kormos (1999: 165) claims that a more extended use of guided role plays would address the problems of the non-symmetrical nature of interaction in non-guided interviews (see also Raffaldini, 1988).[5]

2 *Criterion reference test*

The ACTFL-TTM portrays the OPI test as a criterion-referenced test (1986: 1–3; 1999: 3). However, Lantolf and Frawley (1985: 343) reject this claim because the educated native speaker is 'taken as the norm against which all L2 speaker performance is judged'. In fact, ACTFL workshop trainers explicitly identify a loosely defined notion of an educated native speaker as the target norm of the OPI. So prevalent is this feature of the 'educated native speaker norm' that non-native speakers who otherwise show high degrees of fluency, accuracy and complexity in their speech may fail to rate as Superior (highest level) if they also fail to qualify as 'educated'. This becomes an even more serious issue when native speakers themselves fail to qualify as Superior because they are not considered 'educated'. By this measure, illiterate native speakers, by definition, will fail to qualify as Superior.

As a solution to this dilemma, ACTFL trainers exemplify the target

[5]In fact, both editions of the Tester Training Manual encourage testers to create their own role plays. In principle, this could be regarded as a built-in feature of the OPI procedure that would allow testers to implement some of the above-mentioned modifications. On the other hand, these proposed modifications may eventually affect the standardized nature of the procedure (unless the changes themselves are also standardized).

norm by identifying and describing profiles of successful non-native speakers, such as diplomats or spies, who are able to pass as natives. The relevance of these profiles, however, is problematic. For instance, imagine a spy who needs to take the role of an 'uneducated' farmer. Such a spy would be easily identified – and plausibly killed – if he or she attempted to speak as an 'educated' speaker of the target language. Hence, the target norm of 'educated native speaker' may not necessarily be the choice in all circumstances. In other words, language use is not independent of context. Indeed, awareness of prejudice in favour of particular dialects, especially standard educated ones, should lead towards a better design of the testing instrument through identification of a range of valid target norms (for the use of naive-native-speaker informants, see Barnwell, 1989).

3 Developmental criteria and exponential development

The ACTFL-TTM makes an explicit claim about the developmental nature of its rating criteria: 'each major level subsumes the criteria for the levels below it' and facility 'with language increases exponentially' (1986: 2–5; 1999: 11). Problems with this claim include the following:

1) Individual differences may have an important effect on overall proficiency outcomes (e.g., Kramsch, 1987; Henning, 1990; Purpura, 1998).
2) No substantive reason is provided for the assignment of particular abilities to specific levels.
3) The exponential development of language abilities may not be theoretically justified; and
4) The underlying skills-acquisition framework espoused by the guidelines is only one of the various theoretical frameworks that describe L2 development.

Firstly, the monolithic developmental path represented in the ACTFL-TTM may not necessarily correspond to any one learner. Indeed, Henning (1990: 381) points out that 'the relation between time of learning and position on a scale of proficiency needs clearer articulation for individuals of varying aptitude and language background exposed to various educational treatments'.

Second, the ACTFL-TTM specifies that 'each of the four major levels encompasses a range of performances'. More importantly, it is noted that 'some of these constellations might lend themselves to judgments of a "stronger" or "weaker" performance; others may not affect rating at all' (1986: 2–5; 1999: 11). It is not clear, however, how these constellations of factors are assessed as relevant for each

level. For instance, regarding the gradation of functional abilities, one wonders how ACTFL has determined that *explanation* can only occur at the Superior and Advanced level, that *narration* appears at the Advanced level, that *comparison* emerges at the Intermediate level, and that *description* occurs at the Novice/Intermediate level. This is not a trivial issue, because the above mentioned gradation has important consequences for the rating of speech samples. For instance, as currently structured, level checks may prevent subjects from getting role-play cards that enable them to show proficiency in Advanced or Superior level categories. That is, these interviewees are being denied the opportunity to show their proficiency in those areas.

A similar example from the category context (or more adequately sociolinguistic competence) reveals the same unsubstantiated claim: the management of *informal settings* has been placed at the Intermediate level whereas the management of *formal settings* has been placed at both Advanced and Superior levels (1999: 23–24). To my knowledge, no empirical evidence permits us to presume that formal contexts are more difficult (or that they will be more difficult to learn/control) than informal contexts. By the same token, it is difficult to justify the placement of *abstract and unfamiliar* topics (in the same category of context) at the Superior level, while *concrete and factual* topics belong to the Advanced level. Could we not argue the opposite for the case of an academic learner who may be a specialist in academic discussion and could discuss topics in his or her field of specialization? For instance, in the case of the L1 Spanish–L2 English contrast, the use of the Latinate vocabulary in English (as opposed to the Germanic lexicon) may help the Spanish speaker to excel in a conversation on an academic topic rather than in a more mundane and everyday one. In short, the above-mentioned developmental criteria must be properly substantiated because they have so much bearing on the rating of the OPI interview.[6] As far as grammatical competence is concerned, Tschirner (1996) claimed that L2 acquisition research (e.g., development of word order in German) provides empirical justification for the developmental sequences embedded in the ACTFL-OPI descriptions. Tschirner, however, acknowledges that his interpretation is speculative because he extrapolated 'the kinds of grammar structures that may be expected at specific OPI levels' from the OPI definitions that describe language functions and text types.

Third, the ACTFL-TTM scale 'presumes that facility with a

[6]As a consequence of the above mentioned deficiency in the gradation of functional and structural abilities the assignment of specific role-play cards to each level is untenable. It appears necessary to modify such stringent requirements and to give access to all interviewees to all role-play cards.

language increases exponentially within the various global tasks and throughout the hierarchy of tasks, rather than growing linearly in a merely additive fashion' (1986: 2–3; 1999: 11). Bachman (1990: 45), however, contends that proficiency levels may not represent equal intervals with reference to 'the amount of training required to move from one level to the next'. This is an important consideration that raises questions about the potentially misleading effect of having similar distribution of sublevels within any major level: low, intermediate and high for both novice and intermediate learners. Finally, the ACTFL guidelines make implicit assumptions about the cognitive processing of a second language. For instance, it is stated that:

> [a] stage of *conceptual awareness* will be followed by *partial control* of the feature and ultimately its *full control*. Depending on the complexity of the feature in the target language ... this development from initial awareness of a concept to its *full control in performance* may be a very extended process (1986: 2–9; 1999: 13; italics added).

In other words, ACTFL relies on a skill acquisition theory of language development (see also Omaggio, 1986), a theoretical position plagued by two problems. Firstly, ACTFL has claimed that the oral proficiency interview is atheoretical, a claim contradicted by the above-mentioned quotation from the ACTFL guidelines. In fact, it is hard to imagine any atheoretical language performance test. Second, irrespective of the explanatory value of the process of skill-acquisition in adult L2 development, such a categorical position is incompatible with the potential validity of other psycholinguistic models of L2 acquisition that rely on implicit learning (see Schwartz, 1993; Paradis, 1994; Ellis, 1996). Given that this issue is deeply embedded in a constantly changing research landscape, it would be premature to reject the value of any theoretical approach.

4 Reliability: separation of assessment criteria

As a measuring instrument, the ACTFL-TTM (1986; 1999) establishes that 'the ability to assess oral language use depends on the existence of criteria by which use can be judged' (p. 2–1). The problem is that the criteria provided by the ACTFL-TTM may not be clear enough to allow raters to reliably assess language proficiency. For instance, the ACTFL-TTM establishes that the oral proficiency test 'does not measure discrete aspects of language or knowledge about the language. There are four categories of assessment criteria' (1986: 1–1; 1999: 2). However, should the four categories – and their respective descriptions – be considered discrete? In fact, even supporters of the ACTFL-OPI scale have pointed out potential incompatibilities in the ACTFL description of sub-components of global oral

proficiency. Clark and Clifford (1988: 143), for example, acknowledge that mixing descriptions of linguistic and functional components 'tends to reduce . . . the potential accuracy and effectiveness of both kinds of information'. A possible solution for this dilemma is to do an assessment of sub-components on a principled basis. For instance, it is possible to have a series of tasks (e.g., a series of role plays) that assess language performance in different combinations of sociolinguistic conditions (see Raffaldini, 1988; Shohamy, 1988; 1990; Kormos, 1999).

In general, the inadequate or inconsistent rating of speech samples is predicated on two main factors:

1) the lack of identification of the criteria used to rate performance within each module of communicative language ability; and
2) the relative weight attributed to the different components of communicative ability (see Section 5 below).

With reference to the former, notice that there is no clear separation of the components that make up the rating criteria. The original elaboration of the concept of proficiency into three separate components – function, content and accuracy – (Clark and Clifford, 1988) has been expanded into four components in the ACTFL-TTM (both 1986 and 1999): *global tasks/functions, context/content, accuracy and test type*.[7] However, the classification of the above-mentioned components is inconsistent as exemplified by the following contradictory descriptions:

1) The description of *global function* features such as 'Can describe and narrate in major time/aspect frames' (for Advanced Level) cannot be determined by assessing the learner's accurate use of inflectional morphology. This is because the 'accurate' use of verbal morphology should be more properly considered an *accuracy* feature. In effect, previous research has consistently shown that speakers can narrate in major time–aspect frames using non-inflectional morphology such as adverbials, calendric reference, interlocutor scaffolding, etc. (for example, see Schumann, 1987; Trévise, 1987; Sato, 1990; Perdue and Klein, 1992; Dietrich *et al.*, 1995).[8]

---

[7]In the 1986 edition Table 3-e separates the category context (more properly, sociolinguistic competence) from the category content, whereas Table 3-b keeps them together as is also the case of the description in pages 3–2 to 3–4.

[8]Even though it would be possible to circumvent this problem by accepting the use of non-inflectional morphology as a valid way of describing and narrating in major time–aspect frames, in practice, this is not the case. To solve this ambiguity, an explicit description of the specific mechanisms that may be accepted may be necessary (for the analysis of a similar situation regarding different interpretations of what constitutes circumlocution based on the ambiguity of the definition provided by the Guidelines, see also Liskin-Gasparro, 1996).

2) Conversely, the category *accuracy* is described in terms of *global functions* (i.e., communicative success) (1999: 28):

- 'May be difficult to understand' (Novice level);
- 'Can be understood . . . by speakers accustomed to non-native speakers' (Intermediate level);
- 'Can be understood . . .. by speakers unaccustomed to dealing with non-native speakers' (Advanced level);
- 'Errors virtually never interfere with communication or distract the native speaker from the message' (Superior level).

It is worth noting that the notion of communicative success should not necessarily be related to accuracy of use of the target language. Conversely, it is good interactional competence (e.g., circumlocution, rephrasing, and even translation) that leads to establishing successful communication in the midst of grammatical inaccuracies (see Subsection 1 in Section III above). In sum, the obvious solution to this problem is a clear reassessment of how the categories are identified (see Liskin-Gasparro, 1996).

5 Relative weight of each category of assessment criteria

The ACTFL-TTM claims that 'linguistic components are viewed from the wider perspective of their contribution to overall speaking performance' (1986: 3–1; 1999: 21). However, the discrimination of categories that make up communicative language ability may lead to the assignment of differential weights to each one of those components. In fact, several researchers have questioned the 'unconscious' differential weights assigned to different components of the established criteria (e.g., van Lier, 1989; Douglas, 1994). For example, Douglas (1994: 126; italics added) underscores the potential effect of differential weighting on the reliability of the testing procedure:

> [S]ince test designers cannot completely control raters' interpretations of the scales, *particularly in terms of the weights individual raters may unconsciously assign to various components* . . . the likelihood is strong that no two raters will arrive at the same rating for the same reasons.

Not surprisingly, in most cases, the grammatical component is the one that tends to be favoured in the global assessment of communicative language ability. In fact, from a historical point of view, Savignon (1985: 131) states that 'assignment of ratings on the original FSI scale was made on the basis of separate, weighted scores assigned to discrete linguistic features. Of all the features evaluated, the grammar scale received the heaviest weight'. Potential support for the emphasis on grammatical development comes from studies that show the

deleterious effects of fossilization. For instance, Higgs (1984: 7) proposes that 'postponing linguistic accuracy is an approach that promises a terminal profile'. Notice, however, that fossilization need not be limited to morphosyntactic development, but could, in principle, affect any other component of language development.

The predominance of some components over others in the global measurement of speaking performance – as pointed out by Savignon among others – may address quite different constructs of proficiency. For example, Bachman (1990: 243) asserts that:

> [T]he label 'oral proficiency' for example, has value implications quite different from those of 'communicative language abilities,' as these are related to very different views of language ability (Lowe, 1988), with perhaps different educational and social ideologies as well.

In brief, the differential weight assigned to the components of communicative language ability is conceivably the most significant liability of the ACTFL-OPI model. Consequently, language testers cannot afford to ignore the consequences of not explicitly addressing such concerns. It seems evident that a more specific set of criteria beyond what is provided in the 1986 ACTFL-TTM as well as the minimally revised 1999 version will be necessary to remedy this situation.

V Conclusions

More than a decade ago, Kramsch (among others) criticized the proponents of an across-the-board implementation of proficiency examinations because the 'proficiency movement has broadened its claims without refining its basic premise' (1987: 361). A few years later, Shohamy (1990: 391) asserted that 'constructing tests that are based on intuition alone is irresponsible'. Despite such strong concerns in the profession, the latest revision of the ACTFL-TTM (1999) does not reveal substantial changes from the previous model (ACTFL-TTM 1986). While a complete elimination of proficiency tests such as the ACTFL-OPI may not be feasible, some practical changes may be appropriate and even necessary. These changes can only be implemented, however, if the profession identifies the inherent weaknesses of the test. For these reasons, it is necessary to weigh carefully the status of a testing system – largely unchanged – that has failed to reflect the concerns of the profession in the 1990s. In other words, it is imperative for language testers to review recent developments in the field of L2 acquisition research, teaching and testing in order to make necessary and appropriate modifications to the current model.

To this end, the principled selection of tasks (i.e., conversational interaction samples) to be included in an oral proficiency interview (Subsection 1 of Section IV), the selection and identification of what

criterion or norm will be pursued (Subsection 2 of Section IV), some specification of the developmental process of L2 learning (Subsection 3 of Section IV), the explicit identification and description of the components of communicative language ability (Subsection 4 of Section IV), and the explicit assignment of weights to each category of overall competence (Subsection 5 of Section V) should provide points of departure for the modifications of any future revision of the ACTFL-TTM. The targeted modifications of the ACTFL-OPI test suggested in the previous sections ultimately constitute minor structural changes of the ACTFL-OPI framework. Indeed, although these structural modifications would retain the practical benefits of a proficiency test, they would also incorporate substantial changes that can help language testers address major concerns related to the validity and reliability of the testing instrument.

VI References

American Council on the Teaching of Foreign Languages (ACTFL). 1986: *ACTFL proficiency guidelines for speaking*. Hasting-on-Hudson, NY: ACTFL.

—— 1999: *ACTFL proficiency guidelines for speaking*. Hasting-on-Hudson, NY: ACTFL.

ACTFL-TTM 1986: see Buck *et al.*, 1986.

—— 1999: see Swender *et al.*, 1999.

Bachman, L. 1988: Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition* 10, 149–64.

—— 1990: *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. and **Savignon, S.** 1986: The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. *Modern Language Journal* 70, 380–90.

Barnes, B., Klee, C. and **Wakefield R.,** 1990: A funny thing happened on the way to the language requirement. *ADFL Bulletin*, 35–39.

Barnwell, D. 1989: 'Naive' native speakers and judgments of oral proficiency in Spanish. *Language Testing* 6 (2), 152–63.

Brindley, G. 1998: Outcomes-based assessment and reporting in language learning programmes: a review of the issues. *Language Testing* 15 (2), 45–85.

Buck, K., Byrnes, H. and **Thompson, I.** 1986: *ACTFL Oral Proficiency Interview Tester Training Manual*. Hasting-on-Hudson, NY: ACTFL.

Clark, J. and **Clifford, R.** 1988: The FSI/ILR/ACTFL proficiency scales and testing techniques: development, current status, and needed research. *Studies in Second Language Acquisition* 10, 129–47.

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, 11–22.

De Jong, J. 1995: The needs for standards in language education. *System* 23 (4), 441–45.

Di Pietro, R. 1989: *Strategic interaction: learning languages through scenarios*. Cambridge: Cambridge University Press.

Dietrich, R., Klein, W. and **Noyau, C.** 1995: *The acquisition of temporality in a second language*. Philadelphia, PA: John Benjamins.

Douglas, D. 1994: Quantity and quality in speaking test performance. *Language Testing* 11, 125–44.

Ellis, N. 1996: Sequencing in SLA: phonological memory, chunking and points of order. *Studies in Second Language Acquisition* 18, 91–126.

Hagen, L.K. 1990: Logic, linguistics, and proficiency testing. *ADFL Bulletin* 21, 46–51.

Halleck, G. 1992: The oral proficiency interview: discrete point test or a measure of communicative language ability? *Foreign Language Annals* 25 (3), 227–31.

Henning, G. 1990: Priority issues in the assessment of communicative language abilities. *Foreign Language Annals* 23, 379–84.

Higgs, T. 1984: *Teaching for proficiency, the organizing principle: the ACTFL foreign language education series*. Lincolnwood, IL: National Textbook.

Kormos, J. 1999: Simulating conversations in oral proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing* 16, 163–88.

Kramsch, C. 1986: From language proficiency to interactional competence. *Modern Language Journal* 70, 366–72.

—— 1987: The proficiency movement: SLA perspectives. [Review of Teaching for proficiency: the organizing principle; Teaching language in context: proficiency-oriented instruction; and Defining and developing proficiency: guidelines, implementations, concepts]. *Studies in Second Language Acquisition* 9, 355–62.

Lantolf, J. and **Frawley, W.** 1985: Oral proficiency testing: a critical analysis. *Modern Language Journal* 69, 337–45.

—— 1988: Proficiency: understanding the construct. *Studies in Second Language Acquisition* 10, 181–95.

—— 1992: Rejecting the OPI – again: a response to Hagen. *ADFL Bulletin* 23, 34–37.

Lazaraton, A. 1992: The structural organization of a language interview: a conversation analytic perspective. *System* 20, 373–86.

—— 1996: Interlocutor support in oral proficiency interviews: the case for CASE. *Language Testing*, 151–72.

Liskin-Gasparro, J. 1996: Circumlocution, communication strategies, and the ACTFL Proficiency Guidelines: an analysis of student discourse. *Foreign Language Annals* 29 (3), 317–30.

Lowe, P. 1986: Proficiency: panacea, framework, or process? A reply to Kramsch, Schulz and, particularly, to Bachman and Savignon. *Modern Language Journal* 70, 391–97.

—— 1998: Keeping the optic constant: a framework of principles for writing

and specifying the AEI definitions of language abilities. *Foreign Language Annals* 31 (3), 358–80.

Lynch, B. and **McNamara, T.** 1998: Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing* 15 (2), 158–80.

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

McNamara, T. and **Lumley, T.** 1997: The effect of interlocutor and assessment mode variables of speaking skills in occupational settings. *Language Testing* 14 (2), 140–56.

Messick, S. 1994: The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23, 13–23.

Moss, P. 1994: Can there be validity without reliability? *Educational Researcher* 23, 5–12.

Omaggio, A. 1986: *Teaching language in context*. Boston, MA: Heinle and Heinle.

Paradis, M. 1994: Neurolinguistic aspects of implicit and explicit memory: implications for bilingualism and SLA. In Ellis, N., editor, *Implicit and explicit learning of languages*. London: Academic Press.

Perdue, C. and **Klein, W.** 1992: Why does the production of some learners not grammaticalize? *Studies in Second Language Acquisition* 14, 259–72.

Pienemann, M, Johnston, M. and **Meisel, J.** 1993: The Multidimensional Model, linguistic profiling, and related issues. *Studies in Second Language Acquisition* 15, 495–503.

Purpura, J. 1998: Investigating the effects of strategy use and second language test performance with high- and low-ability test takers: a structural equation modelling approach. *Language Testing*, 333–77.

Raffaldini, T. 1988: The use of situation tests as measures of communicative ability. *Studies in Second Language Acquisition* 10, 197–216.

Sato, C. 1990: *The syntax of conversation in interlanguage development*. Tübingen: Gunter Narr.

Savignon, S. 1985: Evaluation of communicative competence: the ACTFL provisional proficiency guidelines. *Modern Language Journal* 69, 129–34.

Schumann, J. 1987: The expression of temporality in basilang speech. *Studies in Second Language Acquisition* 9, 21–41.

Schwartz, B. 1993: On explicit and negative data effecting and affecting competence and linguistic behavior. *Studies in Second Language Acquisition* 15, 147–63.

Shohamy, E. 1988: A proposed framework for testing the oral language of second/foreign language learners. *Studies in Second Language Acquisition* 10, 165–79.

Shohamy, E. 1990: Language testing priorities: a different perspective. *Foreign Language Annals* 23, 385–94.

—— 1996: Competence and performance in language testing. In Brown, G., Malmkjær, K. and Williams, J., editors, *Performance and competence*

in second language acquisition. Cambridge: Cambridge University Press.

Shohamy, E., Reves, T. and **Bejerano, Y.** 1986: Introducing a new comprehensive test of oral proficiency. *ELT Journal* 40, 212–20.

Skehan, P. 1998: *A cognitive approach to language learning*. Oxford: Oxford University Press.

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: ACTFL.

Thompson, I. 1995: A study of interrater reliability of the ACTFL oral proficiency interview in five European languages: data from ESL, French, German, Russian, and Spanish. *Foreign Language Annals* 28 (3), 407–22.

Trévise, A. 1987: Toward an analysis of the (inter)language activity of referring to time in narratives. In Pfaff, C., editor, *First and second language acquisition processes*. Cambridge, MA: Newbury House.

Tschirner, E. 1996: Scope and sequence: rethinking beginning foreign language instruction. *Modern Language Journal* 80 (i), 1–14.

Valdman, A. 1988: Introduction. *Studies in Second Language Acquisition* 10, 121–28.

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

—— 1996: *Interaction in the language curriculum: awareness, autonomy and authenticity*. New York: Longman.

Villar, S. and **Meuser-Blincow, F.** 1993: Proficiency requirement-based and nonproficiency requirement-based second language programs: how do students rate? *Foreign Language Annals* 26 (1), 49–62.

Young, R. 1995: Conversational styles in language proficiency interviews. *Language Learning* 45, 3–42.