Assessing Oral Performance

ALMEIDA, Virgílio Pereira de[1]

The idea of testing ability has always been closely connected with the teaching of the ability. However, only relatively recently has the principle of test reliability found its way into the field of testing. Spolsky (1978) defines the pre-scientific period[2] of testing as a period "characterized by a lack of concern for statistical matters or for such notions as objectivity or reliability" (p. v-vi). The linguist mentions that evaluation could rely totally on the judgment of experienced teachers, who could determine what grade to give after a short conversation with the student.

Madsen (1983) calls this period the *intuitive era* and admits that although instructors tried to evaluate students with a variety of instruments, including translations, essays and open-ended answers, subjectivity still played a key role in assessment.

The scientific period which followed attempted to eliminate all possibility of subjectivity playing a role in assessment. Objective written and oral tests were developed so as to allow consistent scoring even by untrained raters. In written tests, some words were removed from texts and test-takers were required to fill the gaps with appropriate terms. In oral evaluations, test-takers were required to distinguish between separate sounds. As a result, the grades became consistent and it was thought that the solution to eliminating subjectivity had finally been found. The terms *validity* and *reliability* were first associated with testing.

---

[1] Universidade Católica de Brasília

[2] Spolsky even goes as far as to call the pre-scientific period a *trend*, asserting it still holds sway in many parts of the world.

Specialists started to evaluate tests statistically, checking if they were objective enough, or if they needed to be perfected so as to avoid discrepancies resulted from assessment from different raters. However, testing during this period did not involve any real *use* of the language. Students were mechanically assessed in terms of knowledge about language rules, but were never asked to *perform* in the language. In a sense, there were no actual communication exchanges during an oral test. Nevertheless, the scholars and teachers regarded the fact as a minor inevitable drawback.

From the early 1970s, American sociolinguist Dell Hymes's theory of communicative competence began to exert an enormous influence on the field of language teaching. Hymes demonstrated that there was more than grammar rules and vocabulary involved in communication. Details of language in actual use integrated Hymes's area of analysis, which he named "the ethnography of speaking" (Hymes, 1962). The theorist started the analysis and description of language aspects which had not been formally addressed by other scholars, viz., the interrelations of speaker, addressee, audience, topic, channel, and settings, and the ways in which speakers use the resources of their language to perform specific functions. Hymes's theory expanded the range of what was expected from a language user, particularly introducing the idea of strategic competence. For the first time, instructors were aware that learners had to be taught that the language used to communicate with friends, for example, was not the same used in a job interview, or when talking to strangers.

As Hymes's new theory was gradually but steadily being absorbed by the EFL and ESL community, efforts started to be made so as to reflect such theory (and the

new teaching methodology which developed from it) in testing. In fact, Underhill (1987) points out that once teaching became more than ever directed to speaking and listening, the interest in oral evaluation increased dramatically. Speaking tests were developed to acquire evidence of learners' ability to communicate appropriately in different situations. Therefore, strategic competence was sided with grammatical and sociolinguistic competences in the development of testing instruments.

This new re-working of the concept of strategic competence originated a new problem for the rater since strategic competence is not a type of acquired knowledge, as are grammatical and sociolinguistic competences. Instead, strategic competence involves non-cognitive issues such as the ability to take risks, to negotiate meaning and understanding, among other skills. As McNamarra (2000) observed, "competent native speakers differ in their conversational facility and their preparedness to take risks in communication, and these differences of *temperament* rather than *competence* are likely to carry over into second language communication" (p. 19, my italics). Therefore, if raters are to assess performances taking strategic competence into account, they are to judge such performances with a specific standard in mind, considering the variability of such competence presented by competent native speakers. This may lead us to the conclusion that if an oral evaluation of English were applied to native speakers, they would be rated differently depending on their strategic rather than on their linguistic competence, considering they are all native speakers.

As a consequence, besides the specification of what kind of knowledge is expected in good communication, oral assessment tried to embrace slippery issues not

directly related to knowledge but which were felt should be included in the evaluation. McNamara (2000) even suggests that the “slowness with which the field [of oral language testing] has come to grips with the issues involved is perhaps motivated by a reluctance to face the difficulties of achieving a fair assessment in performance tests” (p. 19). One of the clearer difficulties is the desire to avoid subjective analyses of such personal skill as strategic competence.

However, there is a growing suspicion concerning the unfeasibility of eradicating subjectivity from any of the phases involved in language tests. The issue has been discussed by Bachman (1991), who, quoting Pilliner, reminds us that language tests are subjective in nearly all aspects, from the subjective decisions in producing test items, to the subjective judgments in scoring them.

Language testing has been reflecting the changes which occurred in language teaching due to the changes in course in many interrelated fields, besides the influence from Hymes’s studies. Bachman (1991) cites the influence of Chomsky’s theory of syntax which has become a “dominant paradigm for describing the formal characteristics of utterances” (p. 296). The author also mentions the development of sociolinguistics, pragmatics and the ethnography of communications. Broader assumptions about language and language teaching/acquisition have widened the range of tests as expressed by Canale (1984):

> Just as the shift in emphasis from language form to language use has placed new demands on language teaching, so too has it placed new demands on language testing. Evaluation within a communicative approach must address, for example, new content areas such as sociolinguistic appropriateness rules, new testing formats to permit and encourage creative, open-ended language use, new test administration procedures to emphasize interpersonal interaction in authentic situations, and new scoring procedures of a manual and judgmental nature (p. 79).

In the turn of the millennium, foreign language teaching finally started mirroring

new concepts and insights endorsed by sociolinguistics. Albeit somewhat indirectly, William Labov's findings did find their way into the field of foreign language teaching. When teachers and scholars in the area were shown (or maybe reminded) that a language has innumerable variables related to social context and that native speakers use those variables accordingly, changes started to occur in the way language was taught and therefore tested. In the late 1980s and early 1990s, the aim of foreign language teachers was to make their students speak with a near-native like accent. Nowadays, however, teachers are much more concerned with communicability and appropriateness of language. Osborn (1999) illustrates the fading practice, exemplifying that teachers taught "the 'standard' way of saying something, only to be confronted later by perplexed students wondering why native speakers use a different or even aberrant version" (p. 10).

The shift on the paradigm of what is expected from a foreign language learner was adopted even by course books used to teach English as a foreign or second language, which shifted from the dichotomy of American and British English, to a more neutral category of international English. Such a shift is in line with Trudgill's claim that "we can talk (…) about 'Canadian English' and 'American English' as if they were two clearly distinct entities, but it is in fact very difficult to find any single linguistic feature which is common to all varieties of Canadian English and not present in any variety of American English" (1979, p. 17).

Language testing research should face the challenges which the field presents after the blossoming of the various areas which can contribute to a more reliable and valid form of assessment. Bachman reminds us that the reasons oral interviews are not more widely used is that they are very time-consuming both to

administer and to score. However, considerations of efficiency cannot take precedence over reliability, validity and authenticity. Therefore, empirical evidence of any discrepancy in the oral assessment from native and non-native raters would confirm the need of including the issue in teacher training courses and undergraduate programs for language teachers.

The research described on this paper tried to contrast the way teachers whose mother tongue is English evaluated oral performance to the way Brazilian teachers evaluate the same oral performance.

The students interviewed for this research were recorded while being submitted to a CAE (Certificate of Advanced English) Speaking Paper, which has a rating grid containing the criteria or subskills which raters should follow to judge oral performance. Grades on a scale of zero to five, with intervals of 0.5 points, are awarded to examinees according the following criteria: (a) Grammar and Vocabulary (Accuracy and Appropriacy), (b) Discourse Management, (c) Pronunciation (Individual Sounds and Prosodic Features), and (d) Interactive Communication (Turn-taking, Initiating, and Responding)

The volunteers were paired and interviewed at different times on a radio studio. After the interviews had been recorded, some extracts were discarded so that each interview had a maximum of sixteen minutes. The interviews were recorded into a master CD which was copied and labeled to be distributed to the raters.

Copies of the CD were sent to the raters together with guidelines to rater, and a rating grid. The information to the raters contained explicit instructions that they were to grade the spoken extracts based on only one listening, i.e., they were to listen to the recordings only once to avoid grade discrepancies due to more careful

analyzes by some of the raters.

Five native teachers and five Brazilian-born teachers were given the material and asked to return after they had given the grades to each test-taker. Although the number of raters used in this study was restricted by the difficulty to locate raters from the first group who met the criteria, sociolinguistic research on subjective reactions to speech has long led to what became one of the basic principles of the field, which is the uniformity of attitudes towards language from a speech community (Labov, 1975). American linguist William Labov, in his studies which substantiated the influence of social aspects in linguistics, defended that “in fact, it seemed possible to define a speech community as a group of speakers who share a set of social attitudes towards a language” (Labov, 1966. p. 651). Therefore, in spite of the small number of raters from English speaking countries and from Brazil, it can be assumed that the findings presented by the analyses of data provided by these scorers are representative of the communities they represent, namely, the native speakers and Brazilian-born teachers of ESOL.

The data collected did not demonstrate homogeneity among the raters. The values of Standard Deviation show that native raters had similar judgments about the performance of only three test-takers, whereas non-native raters agreed on the grades of five test-takers. The standard deviation from these test-takers show a value below 0.5, which indicate a fairly similar judgment from the raters. On the other end of the scale, one of the test-taker had a highest standard deviation (0.85), indicating the largest incongruence among raters. On the part of the non-native raters, the highest standard deviation was 0.75, which also indicates heterogeneity among raters.

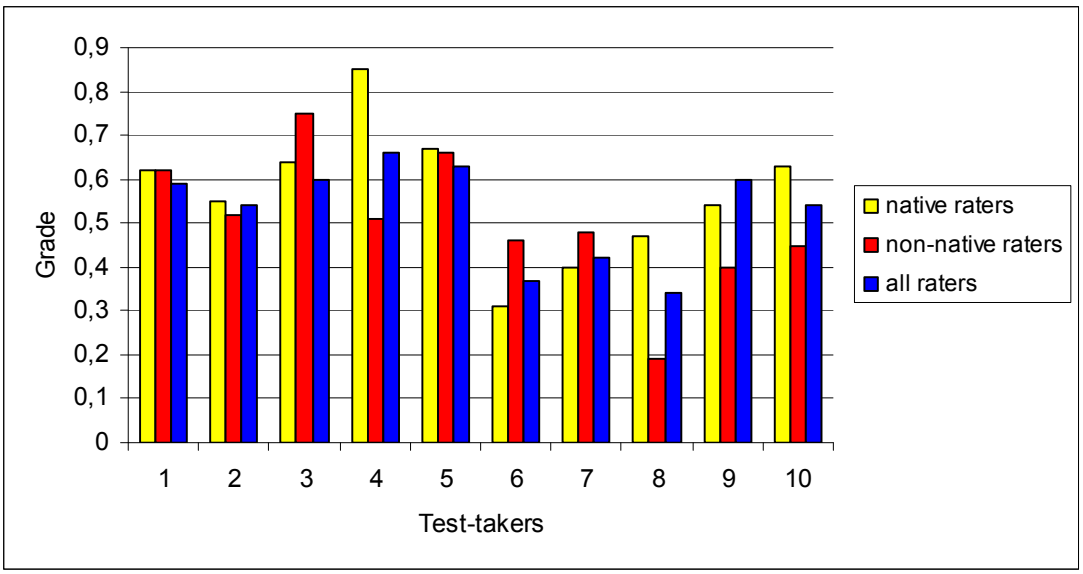

Figure 1 – Standard Deviation from Native, Non-native, and All Raters

The lack of homogeneity among each group of raters (native and non-native) can be further highlighted when we compare the standard deviation of each group, with the standard deviation from all raters combined (Fig. 1). According to the premise that a linguistic community has a similar set of attitudes towards language, it was expected that raters from the same community would rate performance with a higher homogeneity than raters from different communities. Therefore, it was expected that the standard deviation of each group on this research (native and non-native raters) would be lower than the standard deviation from the whole group of raters. Such situation was confirmed for only one of the test-takers (9), for whom the standard deviation of all raters (0.60) was higher than the standard deviation for each group (0.54 for native raters, and 0.40 for non-native raters). This means that each group, separately, was more homogeneous in their judgment toward the test-taker than all raters considered as a whole group. However, the fact that this situation occurred with only one of the test-takers does not corroborate with what

was expected from the study. Since grades varied from zero to five, at a scale of 0.5 points, a standard deviation of more than 0.5 would be significant enough to demonstrate heterogeneity, and therefore incongruence of attitude among raters.

The data collected also demonstrated that the grades awarded by native raters, in average, were not considerably higher than the grades awarded by non-native raters. Native raters awarded better grades to six out of the ten test-takers, but only two of the grades were significantly higher (test-takers 3 and 9 – Fig. 2).

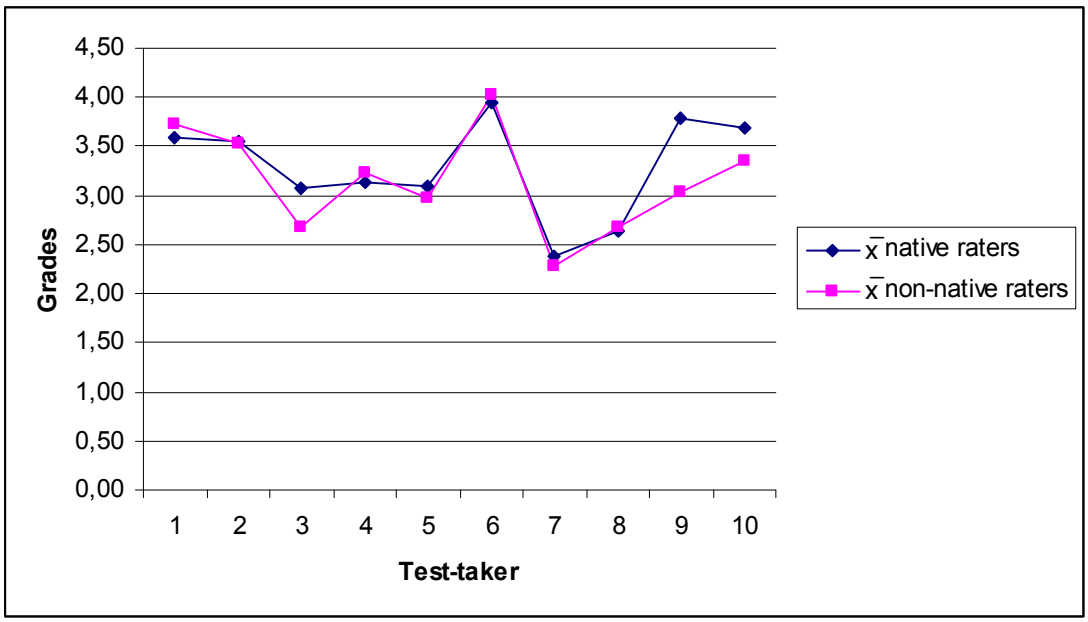

Figure 2 – Arithmetic Mean of Grades Awarded by Native and Non-Native Raters

Such evidence prevents this study from concluding that native raters are more lenient and thus grant better grades to oral performance than raters whose first language is the same as that of the test-takers.

When the average grades granted by native and non-native raters to each subskill are analyzed separately, it is observed that native raters granted higher grades than non-native raters for seven of the test-takers in two of the subskills, viz. the

Grammar and Vocabulary subskill and the Pronunciation subskill. However, the difference was substantial (more than 0.5) in only four instances. Since the data also demonstrate instances in which non-native raters granted substantial higher grades than native raters, it cannot be asserted that native speakers tend to grant better grades than non-native speakers.

The study observed a balance between the groups of raters related to the grades awarded for the Discourse Management and Interactive Communication subskills, that is, none of the groups awarded higher grades to the majority of the test takers for such criteria. However, a detail which challenges this supposed homogeneity is the fact that some of the test-takers received substantially different grades.

Nonetheless, the fact that native raters granted higher grades for the majority of test takers in the Grammar and Vocabulary subskill and in the Pronunciation subskill is, by itself, worthy of note.

Awarding higher grades in the Pronunciation subskill is a clear indication of an ampler view of what *pronunciation* means. Native speakers are, due to their own relation with the language, more aware of varieties of pronunciation which are used by native speakers depending on their geographic origin, social level, schooling, etc. Therefore, native speakers are more accepting of deviations from the pattern as long as communication is not compromised.

The same reasoning can account for the consistent higher grades awarded for the Grammar and Vocabulary subskill. Since raters were asked to evaluate “accurate and appropriate use of syntactic forms and vocabulary in order to meet the task requirements” (CAE – Guidelines to rater), the disparity of grades awarded by native raters also indicates that their concept of “accurate and appropriate”

syntactic forms and vocabulary is wider than that of non-native raters.

Nevertheless, the fact that only three test-takers received higher grades from native raters in all categories invalidates the assumption that native raters generally award better grades than non-native raters. Further research could attempt to deepen the analysis of the issue.

Although the research did not find strong evidence that native raters award better grades to oral performance than non-native raters, certain aspects must be noted.

All students of a foreign language have their oral communication skills evaluated at least once in a one-semester term. Many are tested twice a semester. Such tests are not as elaborate as the CAE Oral Examination, which was used in this research. In most of the language schools, raters are not given a rating grid with subskills to judge. They are simply required to interview a student and grade his oral performance. It is known that most raters, in such circumstances, evaluate pronunciation, grammar, and vocabulary. If this research proposed such an analysis of the test-takers, the findings might have been different and consonant with the literature. In other words, it may well be that the rating grid provided to raters during this research helped to contribute to a less heterogeneous judgment on their part, which is exactly the purpose of the grid on the actual testing situation.

The hypothesis proposed in the beginning of this study will require further empirical substantiation. A study of data acquired from a less controlled evaluation from raters, as is the case in most of the oral tests that language students take, may indicate a shift from the assumption to empirical evidence that native raters award better grades to oral performance.

Bibliography

Bachman, L. F (1991). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Canale, M. 1984. Testing in a communicative approach. In Gilbert A. Jarvis (ed.) *The Challenge for Excellence in Foreign Language Education*. Middlebury, Vt.: The Northeast Conference Organization: 79 – 92.

Hymes, D. (1962). The Ethnography of Speaking. In T. Gladwin and W. C. Sturtevant (Eds), *Anthropology and Human Behavior*. Washington, D.C. Anthropological Society of Washington.

Labov. W. (1966). *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

_____. (1975). *Sociolinguistic Patterns*. Pennsylvania: University of Pennsylvania Press, Inc.

Madsen. H. S. (1983) *Techniques in Testing*. Series Editors R. N. Campbell and W. E. Rutherford. New York: Oxford University Press.

McNamara T. (2000) *Language Testing*. Oxford Introductions to Language Study. Series Editor H. G. Widdowson. UK: Oxford University Press.

Osborne, D. (1999, April-June). Teacher-Talk. *The English Teaching Forum*, 37, 2. Retrieved on November, 14th, 2004, from http://exchanges.state.gov/forum/vols/vol3 7/no2/p10.htm.

Spolsky B. (1978) Introduction: Linguists and language testers. In B. Spolsky (Ed.), *Approaches to Language Testing* (pp. v – vi). Arlington, VA: Center for Applied Linguistics.

Trudgill, P. (1979). *Sociolinguistics: an introduction*. Great Britain: Penguin Books. (Original work published in 1974).

Underhill. N. (1987) Testing Spoken Language – A handbook of oral testing techniques. Cambridge Handbooks of Language Teachers. Great Britain: Cambridge University Press.