Problems of Assessing L2 Communicative Performance

Chul Joo Uhm
(Kwangju University)

Uhm, Chul Joo. 1996. Problems of Assessing L2 Communicative Performance. *Linguistics* **4, 95-108.** Evaluation is an especially problematic aspect of proficiency-oriented language pedagogy. Here, therefore, I briefly review a few important problems at thee levels that seem to cause difficulties in the design, development and use of communicative tests such as: i) lack of theory of communication and of a description of language in use, ii) the loose definition of "communicative proficiency" and the unclear idea of what to test, and iii) the search of a new way for achieving reliability. In this review we find out that the shift of focus in language teaching from language form to use is an exciting and promising development; however, we need to make endless efforts to develop the possibility of alternative assessment as valid and reliable procedures for this change. **(Kwangju University)**

1. Introduction

Foreign language teaching is a problematic area due to the variety of situations and individuals that it involves. When it comes to terms with testing, however, the situation becomes even more problematic because of the moral and human considerations involved (Skehan, 1991). This might be an explanation to the fact that testing is slow to catch up with methodological developments in the area of language teaching. Teachers are in general more aware of the risks and less ready to experiment. Experiments are not satisfied with the results in simulated testing situation, and, on the other hand, they do not feel like using a test as a definitive instrument if they are not sure of its validity and reliability.

As a consequence a lot of speculations are going on in the hope of clarifying ideas of language tests. Sometimes the discussion becomes an end in itself and causes paralysis. This is the case for

communicative testing, after so much talking we do not know yet what kind of jam it is (Harrison, 1983).

In the following we will not try to define the content of the label, but to discuss some of the reasons why the content is so difficult to define and some of the problems connected with communicative testing.

2. The Nature of Problems

On the features that might characterize a test as communicative, Many researchers have discussed in the related literature (e.g., Porter, 1983; Harrison, 1983; Anivan, 1991; Skehan, 1991; Weir, 1990).

Those features depend on the interpretation of the word "communicative" and on the ideas about testing in general. Here are some problems that seem to cause difficulties in the design, development and use of communicative tests:

Problems at construction level:

i) lack of theory of communication and of a description of language in use.

Problems at validation level:

ii) we do not know exactly what to test because of the loose definition of "communicative proficiency."
Consequently,
   a) shall we test communicative competence or communicative performance?
   b) shall we test the process of how communication is achieved or the final achievement of the communicative task?
   c) shall we test only language or also other aspects involved in communication?

Problems at assessment level:

iii) criterion-referenced vs. norm-referenced;

iv) definition of criteria;

v) the search of a new way for achieving reliability.

Now in the following we will look into the above problems.

2.1. Problems at Construction Level

Morrow writes (1979:147): "Starting from a certain set of assumptions about the nature of language and language learning will lead to language tests which are perfectly valid in terms of these assumptions, but whose value must inevitably be called into question if the basic assumptions themselves are challenged." Tests meant to assess the formal knowledge of the language had a consistent theory in the background. Now that theory has been challenged and new developments suggested.

We do not need to agree with Morrow that a theory can be completely invalidated and consequently the tests based on it. But if new dimensions are added to the definition of what language is, we need new measures of assessment. Language as a system is easy to analyze, describe and assess, language in use is so various and dependent on circumstances that any kind of analysis, description and systematization becomes a high demanding task and "a complete theory of communication will not be developed for a very, very long time." (Alderson, 1981b: 56)

Therefore, evaluation is especially problematic aspect of communicative approaches to language pedagogy in which the shift in emphasis from language form to language use has placed. Canale (1984) maintains that communicative testings must address new content areas such as sociolinguistic appropriateness rules, new testing formats to permit and encourage creative, open-ended language use, new test administration procedures of a manual and judgemental nature.

An adequate test of communication must satisfy two main conditions. First, it must be based on the sound description of communication. To the extent that this notion is inadequately defined and understood, tests will become less persuasive. Second, this description must be reflected not only in test content but also test method. This concern with the nature of communication and language in use has been addressed by many researchers (e.g., Canale, 1983; Omaggio, 1986; Bachman, 1990).

2.2. Problems at Validation Level

There is some level of agreement among the writers on what is meant by "communicative proficiency" (Morrow, 1979; Rea, 1979; Carroll, 1980; Oller, 1979; Omaggio, 1986; Weir, 1990). Communicative proficiency does not coincide with the knowledge of the system of the language (phonology, syntax and lexis) but it is the ability to use the system in contexts of meaningful communication to meet the linguistic demands of the various situations in real life.

The knowledge of a language does not coincide with the knowledge of its discrete parts, but it is the capability of using those parts creatively in order to produce effective and appropriate linguistic behavior.

There is less agreement among the writers when the focus moves from the features to the nature of linguistic proficiency. Is it composed of several independent abilities to be assessed separately or is it unitary in nature? A thorough discussion of this point can be found in Porter (1983). Porter recognizes the compound nature of language proficiency in contrast with Oller (1979)'s "unitary competence hypothesis." Oller sustains the existence of a single global linguistic ability undifferentiated in component skills; his statement is based on research. Porter questions Oller's experiments in terms of the statistical methods, the samples and the testing techniques employed. He sustains the necessity for complexity in communicative testing. There can be no single test of communicative proficiency for all candidates but several tests according to several needs. According to Porter, complexity is also a guarantee for authenticity.

2.2.1. Communicative Competence Vs. Performance

If we do not believe that communicative proficiency coincides or can be inferred from communicative competence, then we need to test also performance.

According to Canale and Swain (1980), communicative competence includes grammatical competence (knowledge of the rules of grammar), sociolinguistic competence (knowledge of the rules of use and rules of

discourse) and strategic competence (knowledge of verbal and non-verbal communicative strategies). Morrow (1979) thinks that testing communicative competence is not enough because we cannot infer to what extent the candidate is able to use this knowledge in meaningful communicative situations and process language under constraints of time. This suggests that tests will tend towards simulation with the attendant problems of extrapolation and reliability. Then we have to consider to what extent real-life conditions can be simulated in a testing situation (see Hughes, 1989; Weir, 1990). According to Weir (1993), this real-life approach in testing might be said to characterized by the following features: focus on meaning, contextualization, activity with an acceptable purpose, realistic discourse processing, use of genuine stimulus material, authentic operations on texts, unpredictable outcomes, interaction based, and performance under real psychological conditions.

Even though the ability to use 'real-life language' should be tested, we cannot easily build into our tests all the features of real life all the time. Alderson (1981a) says that "the pursuit of authenticity in our language tests is the pursuit of a chimera: it is simply unobtainable because they are language test." The problem of authenticity regards authenticity of the language used as input in the testing process and authenticity of the tasks.

Porter (1983) advocates the use of authentic texts for testing purposes, and we might add that simplified texts are not always easier to process.

Authenticity in devising the tasks is more difficult to achieve. It is said that authenticity of the tasks is not inherent to the nature of the tasks but is dependent on how far the testee feels the task to be authentic, purposeful and relevant.

A connected problem is that of the interactive nature of the communicative process. How far can interaction be simulated in a testing situation? In the current large scale or institutional proficiency testings such as UCLES/RSA[1] and IELTS[2] each individual's interactions

[1]University of Cambridge Local Examinations Syndicate/ Royal Society of Arts (UK)

[2]International English Language Testing System

cannot be measured properly (Weir, 1993).

There is also the problem of extrapolation. More samples of communicative behavior we have, the more we can generalize about the communicative ability of the candidate. This solution conflicts with the feasibility, practicality and economy of the test and makes it difficult to administer in a classroom situation where time, technical equipment and expertise are lacking.

2.2.2. Process Vs. Results

Then we have to decide if we want to test the process of communication or the results, that is successful communication. The problem is dealt with both by Morrow (1979) and by Alderson (1981a: 56-62). Alderson says: The question is whether tests are mirrors of reality, or "constructed instruments" from a theory of what language is, what language processing and producing are, and what language learning is.

This distinction is crucial because we have to decide if we want to define both types of tests as communicative or not.

In establishing goals and standards for courses and "communicative syllabi" (Brindley, 1989) it seems more legitimate to break down the macro-skills involved into micro-skills (Munby, 1978), because they must be developed and exercised. But is it necessary or even legitimate in a test of communicative proficiency? If the communicative task is successfully carried out, we must assume that the student has used the necessary skills to achieve it.

Morrow (1979: 151) recognizes that performance is by its very nature an integrated phenomenon and any attempt to isolate and test its discrete elements destroys the essential holism. This suggests a qualitative vs quantitative type of assessment and has implications for reliability. He expresses his opinion against the practice of isolating discrete points, even if they are functions, for the construction of proficiency and achievement tests. Nevertheless, in order to solve the problem of extrapolation, he ends by breaking up the global task and isolating underlying skills in the hope that these skills underlie other global tasks, and consequently give us a right to generalize about the

candidate's communicative ability. He says that these skills are operational and points out the difference between discrete-point tests of these enabling skills and discrete-point tests of structural aspects of the language system. We believe that there is a difference, but, as Morrow himself recognizes, it does not solve the problem of the relationship between the parts and the whole. Besides, we know what the structures of the language are, but we do not know very much about the enabling skills that underlie a communicative act and to what extent they contribute to its successful completion. (Weir, 1981: 34).

Multiple-choice (with a change of focus), cloze tests and dictation can be used to assess these skills. We go back to Alderson's distinction and to the question whether they are tests of communicative performance or not. It is not only a problem of "face validity" (Alderson, 1983: 90), it is a problem of "content validity." Talking of cloze tests and dictation Morrow writes (1979: 149):

> ... neither gives any convincing proof of the candidate's ability to actually use ... the language to read, write, speak or listen in ways and contexts which correspond to real life.

Carroll says (1980:35): "Clearly the multiple-choice test has few of the characteristic of normal communication and of cloze tests he says that they are "essentially usage-based," they do not represent genuine interactive communication and are only "indirect index" of potential efficiency in coping with day-to-day communicative tasks. This does not mean that they are not valid tests. There is no point in arguing for the superiority of communicative over non-communicative tests. Communicative ability is only one aspect of proficiency in a language and communicative tests would never meet all the needs that we have for testing. The problem is whether discrete-point items and blank-filling can be used in tests that claim to be a direct measure of communicative performance.

Porter (1983:193) says that both Morrow and Carroll make a distinction between test-task and test item and consequently items like blankfilling or multiple-choice might be used for assessing performance on communicative tasks, but he rightly points out that the notion of

"test item" is in conflict with the notion of authenticity. In fact, test-items are testing devices that have nothing to do with natural language use.

Canale (1984) says that it means simply reclassifying integrated tests as communicative tests, he tries to suggest what the communicativeness of a test is by providing three questions:

i) is the test task a reflection of a real-life situation? (with all that is thereby implied in terms of "real message" transmissions and of appropriacy to situation, discourse and intention);
ii) in scoring, is priority given to meaning rather than formal accuracy?
iii) is the test task one that the testee perceives as relevant?

Furthermore, in order to avoid the problems, Hamp-Lyon (1991) suggests the use of two types of tests - single global impression and multi-trait analytic approaches - to measure both "overall performance" and "strategies and skills used in achieving it."

2.2.3. Language Itself Vs. Other Aspects

Another problem is implicit in the nature of the communicative act. "Ability to communicate ... in specified sociolinguistic settings" (Rea, 1979: 47) involves innumerable components many of which are not linguistic. We know from psycholinguistics and discourse analysis that knowledge of social conventions and knowledge of the world play an important role in understanding spoken and written discourse. They are often cultural specific and cause misunderstanding. Where can we draw a line between what is not understood because of cultural differences or lack of background knowledge? The problem is if it is legitimate to test that as well. Some of the existent proficiency tests like the TEEP (Test in English for Educational Purpose), contain texts (e.g., Changes in the position of women) that need some cultural specific background knowledge.

Basically the emergency of testing performance is recognized by almost all the writers but they are not satisfied with it, they feel that it

is elusive, difficult to attain, and difficult to assess.

Almost all of them suggest two types (Morrow, 1979), two components (see TEEP test), two parts (Porter, 1983) or two tiers (Carroll, 1980) of the same test, one to assess linguistic competence and the other to assess performance related to specific communicative tasks. According to Carroll only the second tier must have characters of authenticity. This means, using Porter's words, that they are constructs in search of validation. Porter suggests that one of the two parts might be redundant.

2.3. Problems at Assessment Level

Testing communicative performance entails a lot of problems at the level of assessment. Moller (1981), in discussing the roles of the candidate and the assessor in the various kinds of tests writes: "In the discrete and integrative tests the candidate is an "outsider." The text of the test is imposed on him ... But in communicative performance tests the candidate is an "insider," acting in and shaping the communication, producing the text together with the person with whom he is interacting. It follows that the assessor is confronted with communication that is unpredictable and of varying quality. In other words, both performance and assessment are subjective. We are used to the scientific preciseness of the complicated statistical procedures of objective norm-referenced tests. Harrison (1983: 84) writes: ... the production and statistical justification of multiple-choice tests has made other more subjective assessments look weak by comparison.

Both Morrow (1979) and Carroll (1980) sustain that performance tests are criterion-referenced tests. Carroll says that the pre-specification of communicative tasks lends itself to criterion-referenced techniques but it is too early to abandon the elaborate and well-worked-out procedures of norm-based statistics. Morrow (1979: 150) says that a communicative test is criterion-referenced and not norm-referenced, because it is meant to show whether or not the candidate can perform a set of specified activities and not to discriminate. But, as Porter (1983) points out, a comparison between people is always implied and the qualitative modes of assessment might be converted into numerical scores to allow

statistics based on the normal distribution of scores to operate.

Weir (1993) points out that a criterion-referenced test may discriminate and even yield normal distribution but this would only be coincidental to the true purpose of the best, that is to indicate if the candidate has reached the criterion of satisfactory performance.

It is possible to conclude that communication-type tests are inherently difficult to grade objectively and reliably.

Rea (1979), in delineating the criteria for the construction and evaluation of any language test talks about two sets of assumptions, one derives from the field of educational measurements- the psychometric considerations- which are basic to test design (validity, reliability). It is thought that the criteria that we use to establish the reliability of a test depend also on the type of content and the purpose of the test, and the content of the tests depends on the first set of assumption. Consequently, when the first set of assumptions changes, the nature and the purpose of the test changes and we have to look to new criteria to establish its reliability. The objective procedures of norm-based statistics were the right measurements for tests that assessed the formal linguistic elements of the language but they cannot be used to assess unpredictable linguistic behavior. We need to find some other form of measurement that guarantees reliability.

The solution of the problem lies in the choice of dependable criteria for assessment and the expertise of the assessor. A careful specification of detailed and unambiguous criteria is crucial and it is not an easy task; therefore, we may run the risk that the result depends more on what the descriptions of the categories mean to the assessor than on the content of the student's utterance (Harrison, 1983: 81). We must admit that it is not easy to find out objective criteria that can be used to assess all the possible linguistic productions in a specific context with all the various degrees of acceptability. We need also some kind of model to refer to. Morrow's suggestion of comparing non-native speakers performance with that of native speakers has rightly been criticized by other writers (Alderson, 1981a: 49). Which native speakers and which performance?

The most comprehensive work in this field has been carried out by Carroll (1980). He has devised a set of ten performance criteria against

which the candidate's performance can be matched and assigned to a nine-point language band system that corresponds to a series of nine performance levels and ten language performance variables (the tenth is non-user consequently not assessable (p. 59). In the Appendix (pp. 134-136) each band is accompanied with a short description. Carroll says that the bands are convenient because the performance profile can be easily matched with the requirement profile of a particular job, the descriptions are less reliable but handy for the candidate that wants to know more about his performance. Only research can establish the reliability of these measurements.

This type of assessment requires expertise on the part of the assessor. Carroll writes (1980: 55): If assessors are give suitable training and guidance, judgements can achieve a fair level of reliability.

For the moment writers in this field do not feel like abandoning the safety of quantitative scores and suggest two types of assignment. Harrison (1983: 81) says that two kinds of "judgement" are required, a pre-test for objective marking and the practicality are not solved the assessment of linguistic performance will remain a difficult task to carry out in a classroom situation. The foregoing problem together with the problems involving reliability make these tests not feasible as creditation tests.

3. Conclusion

At the end it seems reasonable to ask the question: is it worth doing it? The answer depends on the answer to two more questions:

i) Do we need to test communicative performance?

There are situations in which we really need to know if people can operate through the language, for example when people need to use the language as a medium for their studies and their work. Rea (1978) in discussing ESP tests, points out the serious consequences, especially in the medical field, that can be caused by a misinterpretation of instructions given in writing, verbally, or by telephone. In this case an accurate description of the communicative needs of the testee is vital for the development of the test. Recently Garcia and Pearson (1994) in

pursuit of "alternative assessment" offers a good example of needs analysis with all the communicative skills involved. The TEEP test can be considered an example of a test developed on a detailed analysis of the needs of the candidate. For a coherent framework we need to look into the models of Canale and Swain (1980), Canale (1983) and Bachman (1990) and encourage a move from abstract theory to build in context.

The second question is:

ii) Is it possible to measure communicative performance indirectly?

In his experiment, Hughes (1978) has used conversational clozes to predict the ability of ELP students to take part in conversation or discussion in English. The results seem to correlate with the teachers' ratings of the students' ability. He concludes: "The conversational cloze test is an indirect measure of oral ability, and so runs counter to a trend towards making language tests simulate as closely as possible the conditions under which the language is to be used."

It is auspicious that more experiments like this will be carried out. We must not forget the old and discredit knowledge of grammar. We cannot accept Omaggio's remark (1986) that grammar is not vital for communication. Without the knowledge of the lexical, phonological and grammatical system of the language we lack the raw material to build up our utterances, we must accept that accuracy of form will contribute to effective communication. Davies (in Weir, 1981: 33) writes: " ... grammar is at the core of language learning ... Grammar is far more powerful in terms of generalizability than any other language feature."

In conclusion, since the traditional methods of assessment often mask what the students really know and what they can do, we need to make endless efforts to develop the possibility of "alternative assessment" as valid and reliable procedures. In so doing, we can gather "evidence about how students are approaching, processing, and completing 'real-life' tasks in a particular domain." (Garcia & Pearson, 1994: 375).

References

Alderson, C. 1981a. Reaction to the Morrow Paper in ELT Documents III. *Issues in Language Testing.* London: The British Council.

Alderson, C. 1981b. Report of the Discussion on Communicative Language Testing. ELT documents III. *Issues in Language Testing.* London: The British Council.

Alderson, C. 1983. "Who Needs Jam?" in K. Johnson & D. Porter, eds., *Perspective in Communicative Language.* NY: Academic Press.

Anivan, S. 1991. *Current Developments in Language Testing.* Singapore: SEAMEO Regional Language Center.

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

Brindley, G. 1989. *Assessing Achievement in the Learner Centered Curriculum.* Sydney: National Center for Teaching Language Teaching and Research.

Canale, M. 1983. "On Some Dimensions of Language Proficiency," in J. Oller, ed., *Issues in Language Testing Research.* Rowley, Mass: Newbury House.

Canale, M. 1984. "Testing in a Communicative Approach," *Northeast Conference on the Teaching of Foreign Languages,* 79-92. Lincolnwood, IL: National Textbook Company.

Canale, M. & M. Swain. 1980. "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing," *Applied Linguistics* 1(1), 1-47.

Carroll, B. 1980. *Testing Communicative Performance.* Pergamon Press.

Garcia, G. & P. Pearson. 1994. "Assessment and Diversity," in L. Darling-Hammond, ed., *Review of Research in Education.* Washington, DC: American Education Research Association.

Hamp-Lyons, L. 1991. *Assessing Second Language Writing in Academic Contexts.* New Jersey: Ablex.

Harrison, A. 1983. "Communicative Testing: Jam Tomorrow?" in K. Johnson & D. Porter, eds., *Perspective in Communicative Language.* NY: Academic Press.

Hughes, A. 1989. *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Moller, A. 1981. Reaction to the Morrow Paper. ELT documents III. *Issues in Language Testing.* London: The British Council.

Morrow, K. 1979. "Communicative Language Testing: Revolution or Evolution?" in C. Brumfit & K. Johnson, eds., *The Communicative Approach to Language Teaching.* Oxford University Press.

Oller, J. 1979. *Language Tests at School.* London: Longman.

Omaggio, A. 1986. *Teaching Language in Context.* Boston: Heinle & Heinle Publishers.

Porter, D. 1983. "Assessing Communicative Proficiency: the Search for Validity," in K. Johnson & D. Porter, eds., *Perspectives in Communicative Language.* NY: Academic Press.

Rea, M. 1978. "Assessing Language as Communication," *MALS Journal.* Birmingham: University of Birmingham.

Skehan, P. 1991. "Progress in language Testing: the 1990's," in J. C. Alderson & B. North, eds., *Language Testing in the 1990s.* London: Macmillan.

Weir, C. 1981. Reaction to the Morrow Paper. ELT Documents III. *Issues in Language Testing.* London: The British Council.

Weir, C. 1990. *Communicative Language Testing.* Hemel Hempstead: Prentice Hall.

Weir, C. 1993. *Understanding & Developing Language Tests.* New York: Prentice Hall.

Department of English
Kwangju University
592-1, Jinwol-Dong, Nam-Ku
Kwangju 502-703, Korea
FAX: +82-62-674-0078