東京家政学院筑波女子大学紀要第４集　75～82ページ　　2000年

Communicative language testing - an attainable goal ?

Nick Miyata-Boddy* and Clive S. Langham

* *The British Council, Tokyo*

Introduction.

In this paper we will first attempt to define the term 'communicative language testing'. We will then go on to examine ways in which communicative testing differs from other forms of language testing, both in the theoretical basis and what is tested. In the next part we will identify some of the problems communicative language testing faces, and look at how these problems have been addressed.

What is communicative language testing?

Communicative language testing is intended to provide the tester with information about the testee's ability to perform in the target language in certain context-specific tasks. It has to be recognised that given the constraints of time and practicality, only a small sample of the testee's language can be collected, and that however realistic the tasks may be intended to be, the testee's performance will inevitably reflect the fact that s/he was performing under test conditions.

Differences between communicative language testing and other forms of testing

We will address this by first briefly identifying other testing methods in the 'eras' preceding the emergence of communicative language testing, looking at what they were intended to measure and their theoretical basis. We will then turn to communicative testing and examine two of the communicative models on which it is based, and the characteristics which set it apart from other testing techniques.

Spolsky (1975) identified three periods of language testing: the pre-scientific, the psychometric -structuralist and the psycholinguistic-sociolinguistic. Although he has since (Spolsky 1995) offered an alternative view, we will use his original phases in this paper.

Spolsky first identifies the pre-scientific era. He recognises it as dating back to the Chinese civil service exams two thousand years ago, but believes it took its present form from the 18th

century Cambridge Tripos (Spolsky 1995). It was characterised by "the use of essays, open-ended examinations, or oral examining, with the results determined intuitively by an authorized and authoritarian examiner" (Spolsky 1995: 353). As the name suggests, testing in the pre-scientific era did not rely on linguistic theory, and reliability was considered less important than the production of a test that "felt fair" (Spolsky 1995:356).

After the pre-scientific era came the psychometric-structuralist era. The name was intended to reflect the joint contribution of the structural linguist, who identified elements of language s/he wanted testing, and the psychometrist, who produced objective and reliable methods of testing the candidates' control of those elements.

One of the first people to claim the need for input from these two sources was Lado, who was also responsible for the discrete point approach. The discrete point approach broke language down, using structural contrastive analysis, into small testable segments. Each test item was intended to give information about the candidate's ability to handle that particular point of language.

The main advantage of this was that it provided easily-quantifiable data. However, it also had numerous drawbacks, perhaps the greatest of which was pointed out by Morrow (1981:11), "An atomistic approach to test design depends utterly on the assumption that knowledge of the elements of a language is equivalent to knowledge of the language." As he says, knowledge of discrete elements is worthless unless the user can synthesise those elements according to the linguistic demands of the situation, or, in the words of Oller (1979:212, cited in Weir 1990), "the whole is greater than the sum of its parts..."

By the 1970s discrete point testing was no longer felt to provide a sufficient measure of language ability, and testing moved into the psycholinguistic-sociolinguistic era, with the advent of global integrative testing. Oller (1979, cited in Weir 1990) argued that global integrative testing, such as cloze tests, which required candidates to insert suitable words into gaps in a text, and dictation, provided a closer measure of the ability to combine language skills in the way they are used for actual language use than discrete point testing.

However, Oller's unitary trait hypothesis, which supposed that "language proficiency consists of a single unitary ability" (Bachman 1990:6), and upon which cloze and dictation were based, has since been disconfirmed (Bachman 1990) and the techniques have been heavily criticised. Alderson (1978, cited in Weir 1990) pointed out that results of cloze tests were affected according to the number of deleted items and where the deletions began.

Morrow (1979, cited in Weir 1990) states that neither technique allows for spontaneous production by the candidate, relying instead on the examiner for the language input. He also

criticised the techniques on the grounds that they tested competence rather than performance, in other words, they tested knowledge of how the language worked rather than an ability to use it.

The fact that discrete point and integrative testing only provided a measure of the candidate's competence rather than measuring the candidate's performance brought about the need for communicative language testing (Weir 1990). Before we look at the features which distinguish this form of testing, we will outline the models of communicative competence on which it is based.

According to Spolsky (1989:140), "Language tests involve measuring a subject's knowledge of, and proficiency in, the use of a language. A theory of communicative competence is a theory of the nature of such knowledge and proficiency. One cannot develop sound language tests without a method of defining what it means to know a language, for until you have decided what you are measuring, you cannot claim to have measured it".

Several attempts have been made to define what it means to know a language, but we only propose to discuss two of the more influential models. The work of Canale and Swain began in an attempt to "determine the feasibility and practicality of measuring what we shall call the 'communicative competence' of students enrolled in 'core' French as a second language programmes in elementary and secondary schools in Ontario.", (Canale and Swain 1980:1). Canale and Swain proposed a set of three competences which combine to produce communicative competence. The first, grammatical competence, included "knowledge of lexical items and rules of morphology, syntax, sentence grammar semantics and phonology" (Canale and Swain 1980:29). The second was sociolinguistic competence. This was made up of "sociocultural rules of use and rules of discourse" (Canale and Swain 1980:29). The third competence they proposed was strategic competence, which related to "verbal and non-verbal communicative strategies that may be called into action to compensate for breakdowns in communication due to performance variables or to insufficient competence" (Canale and Swain 1980:29). In 1983 Canale updated this model by subdividing sociolinguistic competence, which still relates to sociocultural rules, but he introduced a further competence, that of discourse. Discourse competence concerns mastery of cohesion and coherence in different genres.

The main implication this model had for communicative language testing was that since there was a theoretical distinction between competence and performance, the learner had to be tested not only on his/her knowledge of language, but also on his/her ability to put it to use in a communicative situation (Canale and Swain, 1980).

Bachman's framework (1990) was an extension of earlier models "in that it attempts to characterize the processes by which the various components interact with each other and with the context in which language use occurs" (Bachman 1990:81). The framework included three components: language competence, strategic competence and psychophysiological mechanisms

(Bachman 1990). Language competence comprises two further competences, organisational competence and pragmatic competence, each of which he further breaks down, with organisational competence covering grammatical and textual competence, and pragmatic competence covering illocutionary and sociolinguistic competence. Bachman defined language competence as "a set of components that are utilized in communication via language" (Bachman 1990:84).

Strategic competence consists of three components: assessment, planning and execution. It is the mental capacity to implement language competence appropriately in the situation which communication takes place, and involves sociocultural and real world knowledge. Psychophysiological mechanisms refer to the neurological and psychological processes involved in producing and comprehending language.

One notable advance on the Canale and Swain model is that Bachman acknowledges that test design and scoring might have a significant effect on the testee's performance as a result of strategic competence. Certain tasks lend themselves to use of strategic competence to compensate for a lack of competence in other areas, while tests which are assessed according to the "practical effect of the language performance" (Bachman 1990:105) may be affected by the strategic competence factor.

Canale and Swain and Bachman's are two of the more influential models of language competence, and, along with several others, they provide a useful framework for designing communicative language tests (Weir 1990). We will now go on to describe some of the features of communicative language tests which set them apart from other forms of testing.

Communicative language tests should have high content validity. If they are to be used to make judgements about how an individual can function in a normal situation outside the test, the test has to be as accurate a reflection of that situation as possible. This means that the sample of language collected and the tasks the candidate is called upon to perform should be as representative as possible of the language and skills needed to function in the real life context. Tests, therefore, need to be context-specific. If, for example, the objective is to test candidates to determine whether their second language ability is adequate to undertake a course at a higher education establishment, conducted in that second language, the tasks included in the test should be a fair reflection of the type of tasks the candidate will be required to perform as part of the course itself. As Weir (1990) points out, inauthentic tasks may interfere with the measurement of the construct which we seek. "Tests of communicative language ability should be as direct as possible (attempt to reflect the 'real life' situation) and the tasks candidates have to perform should involve realistic discourse processing" (Weir 1990:12). He advocates the use of genuine texts and that care be taken with regard to task length and processing in real time.

Face validity is also related to authenticity of tasks. Although not universally agreed upon,

many testers believe it is easier to gain acceptance for a test which appears to test real life skills than those which use formats such as cloze, which are not seen outside the test itself. Employing tasks which the testees might recognise also makes it easier to explain and justify the test to them. According to Morrow (1981:18), "Reliability, while clearly important, will be subordinate to face validity."

Tests of communicative spoken ability should have certain characteristics. They should reflect normal spoken discourse and give the candidate chances to initiate. There should also be an element of unpredictability. As Morrow (1981:16) points out, " The processing of unpredictable data in real time is a vital aspect of using language. "

The final aspect of communicative language testing we would like to address is that of assessment. Communicative tests should be assessed qualitatively rather than quantitatively (Morrow 1981). The behaviourist view was that learning took place through habit formation. Following from this, tests such as Lado's aimed to discover whether the correct habits had been formed. If they had, they were rewarded, but if they hadn't, they weren't. Passing the test meant obtaining a certain number of correct responses. However, Morrow (1981) argues that answers to tests are more than simply right or wrong, and that they should be assessed on the basis of how far toward an approximation of the native speaker's system they have moved. Tests should reveal the quality of the testee's language performance. Assessment which relates test performance to external criteria is called criterion referencing. It is an area of some contention, and it is the starting point for the next part of this paper.

Problems involved in communicative testing and ways in which these problems have been addressed.

We will identify the main problems associated with communicative language testing and in each case identify the ways in which testers have addressed them.

The first problem area we propose to address is that of assessment. In the psychometrist-structuralist era, reliability was considered of paramount importance. High reliability was claimed by the use of discrete point items which were either correct or incorrect. However, as was mentioned above, one of the characteristics of communicative language tests is that they are normally assessed qualitatively rather than quantitatively, which inevitably throws some doubt on their reliability because of the involvement of subjective judgements. As Weir points out (1990:13), "the holistic and qualitative assessment of productive skills, and the implications of this for test reliability, need to be taken on board." He also comments on the need to examine the criterion-referenced approach to communicative language testing.

In the CR interpretation of test scores, the candidate's ability is defined "in terms of his

successful completion of tasks from a set or domain of criterion tasks or his performance with reference to a criterion level that defines the ability in question" (Bachman 1990:210). Thus the score provides information about the candidate's ability to perform in a language rather than his/her ability relative to other candidates, which is essentially what norm-referencing does. However, this raises another problem; that of the rating scales used. As Brindley (1991:144) points out, although the scales are widely accepted, it is very difficult to find any empirical basis for them. He also cites a number of other drawbacks to the use of rating scales, including Bachman's point that unless there are upper and lower reference or end points, criterion-referencing is not possible (1989:17, cited in Brindley 1991). These end points exist only in theory because no one has either zero ability or the status of perfect speaker.

These two problems have been recognised and steps have been taken to address them. In the case of rater reliability, although at one time it was thought that subjective measures would never have a place in serious language testing, it is now possible, given "sufficient training and standardisation of examiners to the procedures and scales employed" (Weir, 1998:76) to obtain sufficiently high rater reliability for test results to be valuable. In controlled interviews, Clark and Swinton (1979) report average intra-rater reliabilities of 0.867 and inter-rater reliability of 0.75 for FSI type interviews (Clark & Swinton, 1980, cited in Weir, 1998:76). This does not, however, provide evidence of the construct validity of the scales, as Brindley (1991:157) points out.

One way of producing criteria to be used in proficiency testing is to consult expert judges, such as teachers (Brindley 1991), though the opinions of these experts has been called into question. Another possible source of opinion is so-called "naïve" native speakers (Brindley 1991), since these are the people the testees will encounter when using the language. The third suggested group of experts is the testees themselves, and self-assessment using learner-defined criteria is gaining ground in classroom-based assessment (Brindley 1991).

A second, and related, problem is sampling and extrapolation of results. As was mentioned above, tests take samples of language, and these samples are used for the purpose of inference about the candidate's ability outside the test situation (Weir 1990). Communicative testers endeavour to include contexts and tasks which reflect those which the candidate will encounter in real life. However, the specificity of the contexts reduces the generalisabilty of the information generated.

One way of obtaining a fuller sample of the candidate's language would be to include as many tasks in the test as possible. However, as Weir (1990) points out, this conflicts with the need for efficiency.

Bachman (1991:681) states that in order to make inferences or predictions "we need to demonstrate two kinds of correspondences: (a) that the language abilities measured by our

language tests correspond in specifiable ways to the language abilities involved in nontest language use, and (b) that the characteristics of the test tasks correspond to the features of a target language use context. The problem of sampling has been recognised and efforts are being made to address it.

The last problem we intend to discuss is task format. The method of testing can have a significant effect on test performance. According to Bachman (1991:674), "A number of empirical studies conducted in the 1980s clearly demonstrated that the kind of test tasks used can affect test performance as much as the abilities we want to measure". The topical content of test tasks is also reported to affect performance.

One of the features of Bachman's own model which extended upon the work of Canale and Swain was its recognition that test format can affect performance. As Skehan states (1991:10), "This aspect of the model implies a recognition of the fallibility of testing of the way in which part of the test result may be the result of the test format effects rather than underlying ability, and most ambitiously, that testers need to know about systematic effects of these sorts if they are to allow for them in actual test results, or, better still, to avoid them." One way this problem is being addressed is through introspection studies. Testers, according to Skehan (1991), assume that testees respond to their tests in the way they think they do. He goes on to point out that validation principally consists of testers analysing results with their original hypotheses in mind. However, studies which involve test takers introspecting, either during or after taking the test, show that they do not do what testers assumed they did, and often arrive at answers using a different linguistic process from that expected. This has resulted in a widening of testers concepts of validation (Skehan 1988a, cited in Skehan 1991) and an extension of how tests correspond to reality.

In conclusion, testing has progressed a long way since the pre-scientific era, with its disregard for reliability in favour of "fair" testing. It has passed through eras when reliability and objective testing were dominant to the period today when testers are more interested in how a candidate is able to use his/her knowledge of language in a communicative situation than a demonstration of the knowledge in isolation. It would appear, then, that the goal of communicative language testing is attainable. However, it is a form of testing which, like any other, has problems associated with it, and it is the responsibility of researchers and teachers to endeavour to find solutions to those problems.

References:

Alderson, J. C. (1981) *Report of the discussion on communicative language testing.* In J. C. Alderson and A. Hughes (eds.). Issues in Language Testing. ELT Documents 111. London: The British Council.

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

Bachman, L. F. (1991) *What Does Language Testing Have to Offer?* TESOL QUARTERLY, Vol. 25, No. 4.

Brindley, G. (1991) *Developments in Language Testing*. Singapore: Regional Language Centre, in Anivan. S (ed.)

Canale, M. and Swain, M. (1980) *Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. Applied Linguistics* Vol.1 No.1.

Canale, M (1983) *Language and Communication*. In J. C. Richards and R. W. Schmidt (eds.). London: Longman.

Morrow, K. (1981) *Communicative language testing: evolution or revolution?* In J. C. Alderson and A. Hughes (eds.). Issues in Language Testing. ELT Documents 111. London: The British Council.

Skehan, P. (1991) *Progress in Language Testing: the 1990s*. In J. C. Alderson and B. North (eds.). Language Testing in the 1990s. London: Macmillan.

Spolsky, B.(1989) *Communicative Competence, Language Proficiency and Beyond*. Applied Linguistics, Vol.10, No.2. Oxford: Oxford University Press.

Spolsky, B.(1990) *Measured Words*. Oxford: Oxford University Press.

Weir, C. J. (1990) *Communicative language testing*. London: Prentice Hall.