Constructing Tests that Reflect Instructional Goals

Jianhua Bai, 1998, JCLTA

1. Introduction

One recommendation from Kubler and others (1997:137) is that assessment reflect program goals and be coupled with feedback to improve learning and teaching. When observing our current practice we find that many tests do not provide teachers with specific information about students' performance, and there is a need of constructing tests that better reflect instructional goals. There is a need of constructing tests that better identify students specific needs and strengths so that teachers and students can be appropriately informed about students' learning progress and learning difficulties. For instance, we have been emphasizing communicative competence as one of our instructional objectives, but have we paid enough attention in testing? I think that we should make particular efforts to develop task-based tests that assess specific dimensions of students' communicative competence. We have realized more and more that learning strategies should be an important part of teaching and learning, but have we developed tests that assess our students' learning strategies? Probably not. We should develop process-oriented tests that access our students procedural knowledge. It seems to me that testing often falls behind our teaching practice.

The current testing instruments available in our field are either home-made tests or standardized proficiency tests which are designed to identify how well an individual is capable of communicating in Chinese. They are useful tools for certifying competence or evaluating programs, but less useful in diagnosing students' needs and strengths at different stages of learning. We need to make more efforts and develop better instruments that can keep students and teachers better informed about the needs and strengths of the learners in order to enhance the learning and teaching process. I am sure there are good home-made tests in our field and hope we can share them with each other. My objective in this paper is to 1) discuss why it is important to develop tests that provide more detailed information for teachers and students and 2) present some guidelines for developing more informative tests.

2. The Need for Constructing Tests that Reflect Instructional Goals

Testing is part of evaluation which is the process of collecting information for making decisions. Foreign language tests are generally classified into 1) aptitude tests, 2) proficiency tests, 3) placement tests, and 4) achievement tests. Aptitude tests such as MLAT serve to indicate an individual's facility for acquiring specific skills and learning. They are designed to predict, before beginning language studies, a student's probable capacity of acquiring the language. This kind of testing is mostly used in research studies in our field. Proficiency tests are designed to identify how well an individual is capable of communicating in the language at the time of testing. They are often used for certifying competence or evaluating programs. There are several standardized instruments available for testing Chinese as a foreign language: 1) CPT (Chinese Proficiency Test by Center of Applied Linguistics), 2) OPI (Oral Proficiency Interview by ACTFL), 3) CAT (computer adaptive testing) Programs by Ted Yao of University of Hawaii and 4) HSK (Hanyu Shuiping Kaoshi from the People's Republic of China). Placement tests are utilized for instructional tracking and grouping.

Many of the Chinese summer programs such as Middlebury have developed their own tests for placing students at the appropriate instructional level. Ted Yao's computer adaptive test is also used as a placement instrument by the Middlebury Chinese Summer School. Achievement tests are used to indicate the extent to which an individual has mastered the specific language skills in a formal learning situation. Most of the achievement tests in our field are home-made by teachers themselves.

As mentioned in the above section, the current testing instruments available in our field are standardized proficiency tests which do not provide teachers with specific information about how well the students are doing in acquiring Chinese as a foreign language. It is very important for us to design tests that identify our students' specific needs and strengths. Moreover, testing often has a strong "wash-back" effect and has an important impact on how students learn. Students learning and study habits are often influenced by the way they are tested. Nitko (1989) pointed out some of the undesirable consequences of using tests that are inadequately linked and integrated with instruction: 1) Teachers and students may be inappropriately informed about students learning progress and learning difficulties. 2) Students' motivation for learning could be reduced. 3) Critical decisions about students might be made unfairly. 4) the effectiveness of instruction may be evaluated incorrectly.

A timely diagnosis of students strengths and needs is of great necessity. Different language skills do not advance evenly and a score on the student's general proficiency does not provide enough information for instruction. From my observation of the results of the last five years placement testing at the Middlebury Chinese Summer school I have found that an individual who has a degree of competence in vocabulary does not necessarily have a corresponding competence in syntax or reading or speaking. In order to be useful and enhance teaching and learning tests have to provide teachers and learners sub-scores that indicate various specific dimensions of students' performance.

The effect of weaknesses in some essential skills of the students can be cumulative. Without appropriate guidance at the right time, the learner will fail to acquire the skills essential for normal development of language competence. When minor difficulties occur and are not recognized and promptly corrected, their effects tend to be cumulative and result in severe learning difficulty in acquiring Chinese.

Without knowing students specific needs and strengths teachers often teach toward an ill-defined "average" group and fail to take into account the abilities and needs of those students in the class who are at either end of the scale. When this happens students at either ends will become frustrated, less motivated or dropped the class.

The above discussion shows the importance of integrating testing with instruction. It will be nice if we can develop a formal instrument for wider use, but, before that becomes realized, we can follow the guild lines suggested in the next section to develop our own tests for our own individual program.

3. Some Guidelines for Constructing Tests that Reflect Instructional Goals

3.1. The first guideline in constructing a good test is to understand the nature of language acquisition. In particular we need to address the construct validity: "the test score reflects the area of language ability we want to measure, and very little else (Bachman and Palmer, 1997:39). For instance, if we want to design a test to measure the end-of-semester achievement of our advanced students, we need to know 1) What are the communicative competencies that successful advanced learners are able to demonstrate? 2) What should the objectives of the curriculum be for nurturing competent advanced students of Chinese? If we have an adequate understanding of the above questions, we can use it to guide our test design. We should use our best judgement based on the available research findings and our classroom experience as well. When we take a closer look at some of the currently-used tests for the advanced curriculum we find that many of the test items do not reflect advanced competencies that we want to measure. There are many test items that access vocabulary and sentence level comprehension which are more appropriate for assessing beginning and intermediate-level students; there are not enough test items that really measure "advanced" competencies. For measuring the achievement of advanced learners of Chinese we need to, among other things, construct test items that assess students' ability of comprehending and producing texts of extended discourse. We need to construct test items that assess students' ability of abstract expression both in writing and speaking. We need to construct test items that assess students' advanced communicative competencies such as those described by the ACTFL proficiency guidelines. The point that I have tried to make in this paragraph is that tests should truly reflect what we want to accomplish in teaching for the particular group of learners that the tests are designed for.

3.2. The second guideline is to be aware of the different natures of discrete-point testing and integrative testing, which has been an important issue in the literature of language testing. There has been a long debate on discrete-point testing and integrative testing. Under the influence of the audiolingual approach, Lado (1961) and his followers such as Harris (1969) and Valette (1967) developed the theory of discrete-point testing. The basic tenet is that skills such as grammar, vocabulary, pronunciation should be tested separately. Language is treated as a system of discrete categories such as words, phrases and sentences. They feel that testing a representative sample of the total language items would provide an accurate estimate of the learner's language proficiency.

As the cognitive approach to teaching became popular, the theory of discrete-point testing was criticized by many such as Oller (1976). They pointed out that answering individual items is not of much value. They argued that language is not a set of unrelated bits, but that the bits must be integrated and tested in combination with one another. They stressed contextulizing all teaching points. Consequently, they developed the theory of integrative testing. They believe that language tests should be designed to measure the global proficiency of total communication instead of the discrete linguistic components. Some types of the integrative tests include cloze, dictation and oral interview. Little attention is paid to particular skills.

While the discrete-point theory and the integrative theory compete in foreign language testing, Rivers (1981) proposed that "Just as the teacher needs to identify the specific skill he wishes to test, so he must distinguish carefully the various aspects of that skill and test these one by one, as well

as finally testing them as part of an all-round performance." I would suggest that we integrate both theories in our testing procedures. For instance, in our evaluation of students' speaking performances, we emphasis that communication is the key and provide students with specified communicative holistic tasks. But, at the same time, we specify clearly what discrete dimensions they are supposed to accomplish such as vocabulary, grammar patterns etc. Appendix A is an example. For that test we have clearly described the setting for the communicative task, the roles of the two students and the communicative competence they need to demonstrate. We also specify and later report on the specific dimensions such as their pronunciation and intonation, their command and accuracy of vocabulary and grammatical patterns, their fluency and the social appropriateness of their utterances. For the test to be useful to both teachers and students it has to yield more specific sub-scores that provide instructive and informative feedback that enhances teaching and learning.

3.3. The third guideline is to make sure that testing procedures measure the students behaviors and cognitive processes that have been proved to be desired outcomes of instruction. Types of performance and dimensions of performance should be specified and the outcome criteria be listed. For instance, when we ask students to break a sentence apart and then name the parts of speech of each part of the sentence (Please see Appendix B), we are treating this kind of task as desired outcome of our instruction. Is that a true indicator of successful language learning? In particular is it a true indicator of grammar awareness? Probably not! What we are testing by asking students to name parts of speech is linguistics training rather than language use. On the other hand, as proved empirically by research studies, sentence anagram (Please see Appendix C) is a true indicator of grammar awareness. Sentence anagram refers to the task which requires students to put randomly ordered parts of a sentence into meaningful and correct sentences. And research has indicated that sentence anagram task is an effective way of developing students language ability. Therefore, in order to test students' grammar awareness, we should not ask them to name parts of speech; we should ask them to do sentence anagram, which has proved to be a desired outcome of instruction. Another type of test items is to ask students to correct ungrammatical sentences, but there is no empirically-based research evidence to show that this kind of task is a desired outcome of instruction that foster successful language learning. It seems to me that there is a need of examining our current practice and find out if any of the testing procedures we utilize is flawed in the sense that it is not a true indicator of successful language learning.

3.4. The fourth guideline is to take into consideration the procedural knowledge, which is often ignored in testing. In reading, for instance, research has demonstrated that schemata cannot enhance reading comprehension without being activated. Students need training in developing their metacognitive skills and testing should reflect it. Reading research (Pearson et al 1992) has shown that good readers: 1) search for connections between what they know and the new information they encounter in the text they read, 2) monitor the adequacy of their reading comprehension, 3) take steps to repair faulty comprehension once they realize they have failed to understand something, 4) learn early on to distinguish important from less important ideas in texts they read, 5) are adept at synthesizing information within and across texts and reading experience, and 6) draw inferences during and after reading to achieve a full and integrated understanding of what they read. These are proved to be essential procedural knowledge required of successful readers, but they are not incorporated appropriately in testing of reading skills. We need to make lots of efforts to improve our testing practice in this area, i.e. to take into consideration of strategic competencies or their

development of metacognitive skills which make learning more efficient and effective.

3.5. The fifth guideline is to consider the concept of criterion-referenced testing, which is important for assessing students learning outcomes. Unlike norm-referenced testing which deal with the comparison of performances among students, criterion-referenced testing is concerned with how well and how much an individual student has achieved in her or his acquisition of a particular competency. Learning occurs on a continuum of acquiring a particular competency. We need practicing over time when we learn a language. According to the information processing theory (McLaughlin, Rossman and McLeod 1982; Anderson 1985) two stages are involved in the process of acquiring language skills: controlled and automatic processing. Controlled processing requires attention and takes time, but through practice, sub-skills become automatic and controlled processes are free to be allocated to higher level of processing. It is controlled processes that regulate the flow of information from working memory to long-term memory. It suggests that repeated practice gradually leads students to fluency in a foreign language.

The direct implication to testing is that, in addition to find out how well our students are going in comparison with others or native speakers, we should also pay attention to where our students are on a continuum of acquiring what we try to teach. For instance, we all know that a strong vocabulary is essential for enabling our students to become communicatively competent. Are we paying enough attention to the testing of both the width and depth of vocabulary knowledge? Probably not. In testing, we often pay attention to how many words our students know, not how well they know the words. Bai (1995) pointed out the problems of using L1 definitions of L2 words and suggested eight dimensions to look at when considering depth of vocabulary knowledge in acquiring Chinese as a foreign language. For measuring the command of syntax Bachman and Palmer (1997: 214) suggest five levels that lead to mastery: the first level is zero; the second level is characterized as small range, poor, moderate or good accuracy; the third level is distinguished as medium range and moderate and good accuracy within range; the fourth level is characterized as large range with few limitations and few errors and the fifth level is described as no evidence of restrictions in range and evidence of complete control except for slips of the tongue. For measuring knowledge of appropriate register, Bachman and Palmer (1997:215) also mentioned five levels. For instance, the second level shows evidence of only one register in formulaic expression; the third level is characterized as two registers in formulaic expressions with few errors; and the full knowledge is described as completely appropriate use of two registers with no errors. We do not necessarily have to use their levels, but I think the point here is that, in constructing tests, we need to keep in mind the necessity of identifying where our students are on a continuum of acquiring what we try to teach. Is the learning outcome full and complete or is it partial. If it is partial, what is missing specifically and what else should be done to help our students towards complete mastery?

3.6. The sixth guideline is to notice the difference between formative and summative testing. Summative testing is concerned with the final outcome, such as in the form a letter grade, of student learning and formative testing deals with the on-going performance for the purpose of improving the teaching and learning process. Bachman and Palmer (1997:98) pointed out that "Information from language tests can be useful for the purpose of formative evaluation, to help students guide their own subsequent learning, or to help teachers modify their teaching methods and materials so as to make them more appropriate for their students' needs, interests, and capabilities." Much of our

everyday assessment should be formative in nature. Test results should not be used only for grading. More importantly they should be used to inform the teacher and students for further improving students' performances. It is important that the test provide, in addition to the total score, a performance profile that allows the teacher to identify the students' specific strengths and needs.

In conclusion, testing should reflect our instructional goals and should help us make valid inferences about students' performances. Effective testing and assessment should be an on-going process and an integral component of the curriculum (See Appendix D) and should be justified on the basis that they provide informative and useful feedback to students and teachers, which in turn, enhance teaching and learning.

References

Anderson, John R. 1985. *Cognitive Psychology and Its Implications*. New York: W. H. Freeman and Co.

Bai, Jianhua. 1995. "Teaching vocabulary: 8 faces of a word." *ERIC on Languages and Linguistics*, ED342245.

Bachman, Lyle F. and Adrian, S. Palmer. 1997. *Language Testing in Practice*. Oxford University Press.

Harris, David P. 1969. *Testing English as a Second Language*. New York: McGraw-Hill.

Hughes, Arther. *Testing for Language Teachers*. New York: Cambridge University Press. 1989.

Kubler, Cornelius et al. 1997. *Guide for Basic Chinese Language Programs. Pathways to Advanced Skills*, Vol. III. NFLC at Ohio State University.

Lado, Robert. 1961. *Language Testing*. New York: McGraw-Hill.

McLaughlin, Barry, Tammi Rossman and Beverly McLeod. 1982. "Second Language Learning: an Information-processing Perspective." *Language Learning*, 33: 135-158.

Nitko, Anthony. 1989. "Designing Tests That are Integrated with Instruction." *Educational Measurement*. Ed. R. Lin. London: Longman.

Oller, John. W. 1976. "Language Testing." *A Survey of Applied Linguistics*. Ed. H. D. Brown. Ann Arbor: University of Michigan Press.

Pearson, P. David. et al. 1992. "Developing expertise in reading comprehension." *What research has to say about reading instruction*. Ed. S. J. Samuels, and A. E. Farstrup. Newark: International Reading Association.

Rivers, Wilgar. 1981. *Teaching Foreign Language Skills*. Chicago: University of Chicago Press.

Skehan, Peter. 1988 "State of Art: Language Testing." *Language Teaching and Linguistics Abstracts*.

Valette, Roberta. 1967. *Modern Language Testing*. New York: Harcourt, Brace and World.

Appendix A:
Communicative-task-based Testing that also Examines the Various Dimensions of Performance.

Test Description: a role play

An American introduces herself/himself (Student A) to a Chinese (Student B) who recently arrived in the States and then initiate a conversation. The following are some of the possible topics. You don't have to follow the exact order.

The American can explore about the newcomer: 1. Recent arrival? Traveled alone? 2. doing what here and where? Both the American and the Chinese each want to know 1. about each other's families. 2. if selected members of these families can speak or read or write other languages. 3. the manner in which they speak or read or write other languages. 4. the geographical area of the family. The American is interested in life in China and want to know about traveling in China, food, social customs etc.

Evaluation Criteria:

The criterion for evaluating your performance (1-5 scale, 5 being very strong) is to see how well you can accomplish the specified communicative task which require 1) good pronunciation and intonation, 2) accuracy of use of vocabulary and grammatical patterns 3) fluency, which does not equal speed; a fluent conversation is natural, interactive and meaningful. 4) quantity, which means you should try and use as many as you can of the words and grammatical patterns we have learned. 5) social appropriateness, which means that your utterances are socially appropriate in accordance to the specified social contexts.

Appendix B:
Undesired Test Items for Testing Grammar Awareness
"Pointing out the subject, predicate, attributive and adverbial adjuncts:"
(Insert examples here)

Appendix C: Desired Test Items for Testing Grammar Awareness
Sentence Anagram: (Insert examples here)

Appendix D: Testing as an Integral Component of the Curriculum