DOCUMENT RESUME

ED 366 214 FL 021 831

AUTHOR	Weiping, Wu
TITLE	TOCFL: Problems and Guidelines.
PUB DATE	91
NOTE	22p.; Paper presented at the Annual Meeting of the Association of Asian Studies New York (Ithaca, NY, 1991).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Guides - Classroom Use - Teaching Guides (For Teacher) (052) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Chinese; Cultural Context; Language Proficiency; Language Role; *Language Tests; Listening Skills; Reading Skills; *Second Languages; *Standardized Tests; *Test Construction; Test Format; *Testing Problems; Test Reliability; Test Validity
IDENTIFIERS	Authentic Materials; *Oral Proficiency Testing

ABSTRACT

Problems in the testing of Chinese as a foreign language (CFL) are examined, focusing on proficiency testing needs and test standardization. Particular attention is paid to listening and reading assessment. The first part of the discussion looks at specific problems with five existing proficiency tests, including such aspects as inadequacy of the test's scope, the threat to validity posed by a bilingual test format, inherent difficulty in reading characters when the test is in Chinese only, and reliability problems arising from the use of authentic materials or nonstandardized texts. The second part of the discussion offers guidelines for designing a standardized CFL test. Guidelines are that: the test should be in Chinese only so that it can be used by all learners of Chinese regardless of background or native language; it should be based on the core of the language, that shared by all Chinese speakers; its content and design should be politically neutral; and close attention must be paid to the cultural loading of test content. Specific procedures for selecting phonological content and vocabulary are outlined. It is suggested that such a test will be increasingly useful as the global economy develops. (MSE)

****

TOCFL: Problems and Guidelines

Weiping Wu

Georgetown University

(AAS/NY Annual Conference, 1991. Cornell University, Ithaca, NY)

****

TOCFL: PROBLEMS AND GUIDELINES

The present paper consists of two parts. Discussed in the first part are the problems with existing proficiency tests for the Chinese language and in the second part, the guidelines for designing a standardized test for Chinese, named TOCFL with apparent reference to TOEFL. There are three guiding principles for designing such a test, scientific, neutral and flexible.

Due to the nature of the four basic language skills (listening, speaking, reading and writing), it is much easier to have a standardized test, in the sense that it can be machine-scored and relatively easy to administer, for the decoding process (listening and reading) rather than the encoding one (speaking and writing). Since the TOEFL also focuses on the comprehension skills and the TOCFL I am talking about is along the same line, the discussion in this paper will concentrate on the proficiency test for listening and reading.

Given the fact that Chinese is the language used by the most number of speakers, and that its users are found in more than one culture, it is expected that there are quite a few varieties in almost every aspect of the language, including the three basic components of phonology, grammar structure and vocabulary. On the other hand, no matter what variety or where it is used, they all share the same core. That is why it is feasible to have a standardized proficiency test that measures the language skills of all users of the language. The presupposition of such a feasibility underlies all the arguments in this paper.

Over the years, quite a few proficiency tests for Chinese have been developed by various institutions. Among the more prominent ones in regular use now are DLPT, the FSI TEST, CPT, Pre-CPT, and *HSK*.

DLPT: Chinese Proficiency Test developed by the Defense Language Institute and is used mainly to test the proficiency of military personnel in listening comprehension and reading comprehension. Currently, the Mandarin version is called DLPT IV, though DLPT III is still in use.

FSI TEST: Chinese Proficiency Test developed by the Foreign Service Institute of the State Department and is mainly used to test the oral and reading proficiency of diplomats.

CPT: Chinese Proficiency Test developed by the Foreign Language Education and Testing Division of the Center for Applied Linguistics (CAL) in 1983. It is a proficiency test in listening and reading comprehension for English-speaking learners of the Chinese language.

Pre-CPT: The new Chinese Proficiency Test, developed by CAL under a grant from the US Department of Education. It is more than just an updated version of the first CPT, since one of its purposes is to stretch the lower end of the first CPT so that learners with a lower proficiency can be differentiated.

HSK: *Hanyu Shuiping Kaoshi*, the initials in Chinese *Pinyin* for Chinese Proficiency Test, developed by Beijing Languages Institute and the National Office in charge of teaching Chinese to foreigners in China, with its stated goal of "testing the Chinese proficiency of foreigners, overseas Chinese and the non-Han people" (peoples who are not the Han people ethnically). It has been through the trial period and is now administered regularly on 15th of January, June, and October in Mainland China; on June 15th in Singapore and on October 15th in Australia.

While it is undeniable that each and every test listed above measures a certain aspect of the learner's proficiency in Chinese, none of them can claim to be "The Test" for the Chinese language. In reference here is the status the TOEFL enjoys in English. Though there have been complaints about some side effects of the TOEFL, one should always remember that it has served, and still serving the need to have a proficiency test for the English language ever since its inception in 1963. In spite of the existence of a variety of other proficiency tests designed and used by individual institutions, the TOEFL is not replaceable as an effective instrument in measuring English proficiency on a global level. It is in this sense that it can be considered as "The Test" for the English language. A closer look at the existing Chinese proficiency tests will give us some idea what is lacking in this relatively new field of Chinese proficiency testing.

Following Harris (Harris, 1963), a good test has three characteristics: validity, reliability and practicality. Since this is not an overall evaluation of the tests mentioned above, I will just concentrate on some of the problems in light of these three criteria for a good test. The TOCFL, which is still a blue print for the time being, should be free of these problems if it is to play the same role as the TOEFL does in English.

In terms of language, DLPT and the two CPTs use both Chinese and English in the test, while *HSK* uses only Chinese. In both cases, however, one serious defect remains: the content validity of the test.

When English is used, as in CPT and DLPT, it is in the question and the choices. The stimulus is in Chinese (In both the listening and the reading sections). Although the use of English in a Chinese proficiency test designed for English speakers can be justified on the ground that it is impossible to have beginners read the choices in Chinese, its side effect is obvious. In the reading test, especially in DLPT which measures proficiency up to level 3 (ILR), the examinee can usually pick up clues from both the question and the choices in English. In such a situation, it is difficult to claim that the test measures the examinee's reading ability accurately. So the validity of the test is in question.

Theoretically, one can argue that, if the point being tested is not be revealed in either the question or the choices, the use of English would have no side effect. The nature of reading above the elementary level, however, makes it almost impossible to do so. One reason for this is that context plays a key role in reading comprehension beyond the elementary level. It is not surprising if the effort to separate a certain word or phrase from its context fails. One of my students, who received training before being sent to work in Beijing and scored a 2+ in the DLPT reading test upon the completion of his training, told me that "the English helped a lot".

Another side effect with this bilingual format is the unavoidable process of translation. A monolingual decoding process is certainly different from a bilingual one. When hearing the Chinese stimulus but facing the English choices, the examinee has to either translate the stimulus into English or the choices into Chinese before a decision can be made. Because of this process, one more variable is at work during the test. Thus one cannot be sure if the failure of the examinee is due to misunderstanding or to other problems in the translation process. Moreover, the same word may have different connotations in the two languages. The examinee may pick the right choice for the wrong reason because of the connotation carried by the English word.

When it is all Chinese, as in *HSK*, we face the other side of the problem. Examinees with elementary proficiency can in no way read the Chinese characters in the given time slot and make a decision. the complexity here is more than what it appears to be. The obvious objection to the use of the Chinese characters comes from the fact that most of the examinees would not know the word in the first place, such as *guantou* (can food), *zuoqujia* (composer), *tongqing* (sympathetic) in the first section of listening comprehension in *HSK*. The length of the choices is another problem. One of the items actually has 4 very long choices. There are 55 characters altogether and the longest choice has 16 characters. Even intermediate students find it hard to cover all the choices within the given time, let alone making a decision.

Another objection, which is often neglected, against the use of characters is from the perspective of information processing time. Both listening and reading belong to the decoding process. The difference between the two tests (listening and reading) is that in testing reading comprehension, only one process (to understand the reading material) is involved. In listening comprehension with Chinese characters in the choices, two decoding processes, both listening and reading, are involved. Even if we assume zero difficulty in the reading part, the time needed for decoding the reading message cannot be denied. It is therefore unfair and unscientific to require the examinee to finish the listening comprehension within approximately the same time slot as that for reading comprehension. The results of an experiment described below lend supporting evidence to such a rationale.

The same *HSK* was given to 4 students from the Department of Defense. all of them had a proficiency between 2+ to 3+ for both listening and reading according to the DLPT scale. The average rate of correct answers for the listening comprehension section is only 59.5%, the lowest among the four sections of the test. Reading comprehension, on the other hand, has the highest rate, 86%. For grammar structure, it is 78.5%, and for comprehensive (including filling in blanks), 84%. Individually, all the four students scored lowest in their listening section. Teacher evaluation, which serves as an outside criterion, disagrees with the test result. The disparity in their proficiency in listening and reading is not as great as indicated by the percentage. One of the possible reasons here is the lack of time for each item in the listening section. One may argue that even the TOEFL adopts such a format and allows no significant difference in time between items in the listening and the reading parts. For those who are familiar with the Chinese language, it is clear that the gap between written and spoken Chinese, however, is not the same as the one between written and spoken English.

The FSI Test consists of two parts: speaking and reading. I will focus on the reading test since speaking is production rather than comprehension. The test differs from all the others in that it requires an evaluation team for each examinee and is administered on an individual base. Since it is not actually a standardized test, its reliability remains questionable if not administered properly. Although qualified examiners are certified and are supposed to have a consistent way of scoring, they are human beings and therefore are not free from the influence of mood, environment, and many other factors. Thus scorer reliability can be a question.

Most of the materials for the test are chosen from actual language data, like newspapers. The advantage of such an approach is the authenticity of the test. If a student can read the article during the test, most likely he or she can read those of similar difficulty level in the newspaper. the disadvantage also derives from the same fact. Since authentic materials are not written for the testing purpose, there are problems in subject matter, style, vocabulary that may lead to differences in difficulty level in the articles chosen for the test. If the five articles given to student A for briefing are not of exactly the same level of difficulty as the 5 given to student B, for example, test reliability cannot be guaranteed. Even if the same 5 articles are given to both students (which is not often the case), there is no way to guarantee that the two students will pick the same article. The flexibility in the choice, while helping the students, actually hurts the test reliability.

In spite of these problems, the FSI test is still one of the most adequate and accurate means of measuring the proficiency of the examinee, especially the speaking test. This is because real language is for communication and is interactive. The format of the test makes it possible to have a real situation in which the examinee can work with real language data (reading) or use the language (speaking). The price for this is the lack of practicality. Imaging having more than 5,000 examinees from 17 countries (that is just one percent of the total number of people who take the TOEFL annually) taking the test every year, how many qualified testers would be needed! It may serve the FSI purpose well, but certainly cannot be used on a large scale, like the way most proficiency tests are used.

Also along the line of practicality is the format of both CPTs and the DLPT. Due to its format, the test is not available to non-English speakers. If the TOEFL is to be taken as an example of proficiency test, the bilingual format in CPT and DLPT can be considered as a defect in a Chinese proficiency test. A monolingual test, in the sense that everything is in the language to be tested, is not the automatic solution either. Again take *HSK*, the instructions alone present an unsurmountable barrier to examinees with a lower proficiency. Since the instruction is an external part, it can be easily corrected. Instructions in other languages can be provided as needed. This will not affect the validity and reliability of the test, but will certainly increase its practicality. On the other hand, the problem of characters in listening comprehension discussed before is not as easily solved. The solution to such a problem is not the focus of this paper, but two suggestions may be mentioned here. One is the use of pictures for listening items for the Novice Levels, similar to some of the items in Basic English Skills Test (CAL, 1982). The other is to include not just the stimulus, but the question and the choices in the tape as well.

The problems with the existing proficiency tests in Chinese can be summed up as follows. In terms of scope, they cannot cover examinees from the elementary level to the advanced level, or from ACTFL Novice Low to level 5 in the ILR scale, to be specific. This does not mean that all these tests fail to meet their intended purposes, since each was designed to be used only within certain levels of proficiency. It does mean, however, that we are still looking for a test that can cover all proficiency levels. The two CPTs can be regarded as good tests for the beginning level. DLPT goes as far as intermediate but then the English part becomes damaging to validity. *HSK* claims to be for beginning and intermediate levels. If what is meant by "beginning" is the same as the CPT designers understand it, then it certainly fails to hold up to such a claim. In terms of practicality, those with English in the test already enforce a limit to its application, while the one with Chinese alone excludes a large number of people with low proficiency.

It is true that the TOEFL has a lot of unwanted side effects and is not liked by all, one can hardly deny that it is The Test for the English language. I am not saying that the TOEFL as a standardized test will accurately measure a testee's English proficiency in every case. Due to its importance in college admission for foreigners, it has been known to have misled the direction of English teaching in many places. It is unfair, however, to blame the TOEFL for these unwanted evils. The recognized fact is, it is an effective instrument, if used properly, to measure the English proficiency of the learner. One does not have to know another language in order to take the test. Nor would one be likely to fail the listening test because of insufficient reading comprehension. Moreover, it does cover learners from beginning to advanced.

From the discussion above on the problems with existing proficiency tests in Chinese, it is clear that there is still a long way to go before we have a standardized test that can be accepted by all. To have a proficiency test for the Chinese language that resembles the TOEFL in English, one thing is clear: The test should be in Chinese so that it can be used by all Chinese learners regardless of their background or their native languages.

Once the problem of language is solved, the next thing to worry about is how the test should be designed. Among the many factors that may come into play in designing a proficiency test, the three guidelines mentioned at the beginning of this paper are certainly worth our attention. Discussed below are these three guiding principles for designing TOCFL: scientific, neutral and flexible.

The first principle, to be scientific, covers the linguistic part of the test. The test should be based on the core of the language. Such a core is shared by all Chinese speakers no matter where they are. Of the three major components of the language, phonology, grammar and vocabulary, the first two can be considered as close systems and the selection of the core features are relatively easy. This does not mean that there is no problem. In phonology, for example, the distinction between the pronunciation for the word "and", which is pronounced as second tone *he* in Mainland but as forth tone *han* in Taiwan, makes it necessary to avoid the use of this most common word in the stimulus of a listening item.

To insure content validity, a list of key points to be tested in each aspect of the language must be worked out before actually writing the items. The task to construct such a list is relatively simple in phonology. The obvious thing to pay attention to here is perhaps the few differences in pronunciation of certain common words, like the example cited above. With grammar structure, it is a much tougher job but still manageable. Textbooks used at various levels are a good source to start. Efforts have been made in this field by most CPT designers and others interested in Chinese teaching and testing. The structure list collected by Kubler (Kubler, 1988) is one example in this respect.

To insure acceptability by all, at least the textbooks for elementary Chinese in Mainland, Taiwan, Hong Kong, Singapore, The United States, Australia and other places where a sizeable Chinese community exists, should be consulted. One problem with a hastedly constructed list is that item writers and reviewers for the test often spend valuable time arguing whether or not a certain feature should be included in the first place. Agreement on key features to be tested before item writing would help a great deal in preventing such waste of time.

The most demanding part of the task is the vocabulary list. With approximately 45,000 individual characters and virtually countless words, any claim that a small list of several thousand words represents the vocabulary is not easy to justify. The only objective way, relatively speaking of course, is frequency count. In this respect, the several steps in selecting the final word list adopted by *HSK* designers are worth mentioning.

There are four levels of vocabulary in *HSK*. Frequency counts of various kinds played a key role in the process of selection. A description of the process will illustrate this point. First of all, the selection was based on previous frequency studies. There are three sources from which the Vocabulary Guideline was constructed: (1) frequency dictionaries, (2) vocabulary lists, and (3) textbooks. Examples include Frequency Dictionary of Modern Chinese (FDMC, 1895) in the first group; A List of 3,000 Frequently Used Words in Mandarin (1958), A Practical List of Frequent Words in Chinese for Foreigners (1981) and Frequent Words in Teaching Chinese to Foreigners (1986) in the second group; and 16 textbooks compiled by both Mainland and Taiwan in the third group. It goes without saying that this latest Vocabulary Guideline used for *HSK* is a further development of the existing ones. Even the third group, textbooks, can be considered as some type of vocabulary list because compilers heavily rely on some sort of vocabulary studies in the first place.

Three steps are used in selecting the vocabulary for each level. Take Level A for example. There are 1,011 words in this level. These are considered to be the most common words in the Chinese language. The first step is to single out all the words that have a Frequency of Use (FU), as opposed to Frequency of Occurrence (FO) of 120 or above from the FDMC. The distinction between FU and FO is a new index used by FDMC. At the risk of oversimplifying the whole issue, FU can be defined as a complex index which takes into consideration the number of each occurrence, the type of data in which a word occurs and the number of articles in which it occurs. It is a more accurate measure of the frequency than FO, a single index indicating the total number of occurrences.

The second step is to compare this list of words with the second and third group of the source materials (previous vocabulary lists and textbooks) and, in accordance with years of experience in teaching Chinese, delete certain words that have high FU but are not considered as essential to learners of Chinese. Among the deleted words are, for example, *gongchandang* (The Communist Party) and *guomindang* (The Nationalist Party). Due to the large proportion of newspapers and periodicals in the data base, there are quite a few words of a political nature with high frequency. These are not considered as Level A vocabulary for the learners. One the other hand, words like *shengri* (birthday) and *shangwu* (morning) are added to the list of Level A words, though these words are not among the top (120 FU or above) frequency words in the FDMC.

The third step is to consolidate the remaining words from the first step and the added words from the second step. In every step, as is clear from the description here, different kinds of frequency indexes remain the major criteria for the selection. The vocabulary for all the other three levels, B, C and D, is selected in the same manner as the A Level. The total number of words from all the four levels in the Vocabulary Guideline is 8,000, which is used as the vocabulary base for *HSK*.

The main problem in fulfilling the first criterion lies in constructing the core of each major component of the language. Whether or not the core features quantitatively represent the language proportionately in terms of importance directly affects the content validity of the test. Once it is agreed upon by the working committee that the core lists represent the most common elements of the language, then the actual item writing will begin.

One important point in this respect is naturalness of the items. The worst way of writing an item is perhaps to have the structure in mind and then write an item for it. The core list should be used as the reference rather than the actual materials. Item writers should be given only very general guidance and told to write, or collect from authentic materials if possible, what is the most natural utterance. These items are then checked against the core list to see if they include the key features or if they are beyond the level intended.

To be neutral is a criterion that deals with the politics of the TOCFL. As a linguistic project, the TOCFL will never meet the need of all Chinese speakers if linguists overlook the politics behind. Language is always associated with the society in which it is used. Given the sensitive issues between Mainland and Taiwan, for example, certain words, no matter how common they are, should be avoided for the test to be accepted by all. Most people are now aware that words like *Gongfei* (Communist bandits) and *Jiangfei* (The KMT bandits under Jiang) have become history. It is a fact that no people would call each other bandits these days. But a lot of the more subtle ones, like *jiefangqian* (before the liberation, i.e. before 1949) used in Mainland, should also be avoided.

Another aspect along the same line is external to the test but as important as any internal issues. That is, a lot of politically subtle but intriguing issues will be involved. Any of these issues, if not handled properly, can ruin all the efforts and lead to the failure of the whole project. Some of these seemingly insignificant issues are: name of the test (CPT or *HSK* or TOCFL, or TOCSL, for example), sponsorship (what will be the hosting organization?), the right to administer the test, the procedures involved, and so on.

Given the significance of such a project as TOCFL, it is certainly worthwhile for all parties concerned to make concerted efforts for its development. All parties concerned will include at least the following: individuals who were involved in existing tests as well as linguists and teachers who are experienced in the field, organizations like CAL, ACTFL and CLTA in the United States, The National Office in charge of Teaching Chinese to Foreigners in China, organizations in Taiwan, Hong Kong and other places with a sizeable Chinese speaking community, as well as any institutions with the resources for the project.

To be flexible, as the third criterion, addresses the cultural aspect of the TOCFL. It has been mentioned at the very beginning that Chinese speakers are all over the world and their cultural contacts can be quite different from group to group. Although Mainland has the most number of speakers, it does not follow that the particular subculture it represents is the only culture for the Chinese speakers. How much culturally bounded information should be admitted in the test is a question that cannot be ignored.

Recognizing the close relation between language and culture, we cannot hope to have a test that is purely linguistic. Advanced level students are differentiated from those of a lower level proficiency mainly by their ability to decode culturally loaded messages, and not just linguistically complex structures. It is nothing new to language teachers to hear "I know every word in the sentence but still don't know what it means!" The few examples cited by Zhang Kai (See Liu, 1989) demonstrate the importance of cultural knowledge in the test.

As is the case in the second criterion, the major problem here is also vocabulary. To be flexible means to allow some of the idiomatic expressions tied to a certain subculture to come into the test. Without such flexibility, even the simplest word can cause problem. Take words commonly used in greeting strangers or customers, *xiaojie* (Miss), *xiansheng* (Mr) *taitai* (Mrs) versus *tongzhi* (comrade). One may say that the first three are mainly used in Taiwan, Hong Kong and Singapore and insist on not having it in the test. The counter argument is that these are common words of the language and no matter where they are used, a language learner should know these words and therefore they can be included in the test. Moreover, it is more and more common that people in Mainland, especially those in the coastal areas, are now using these terms in their daily life, though the majority (in terms of absolute number) of the people still tend to use *tongzhi*. Taking a different stand, one may also say that *tongzhi* is so common (still so with most of the Mandarin speakers) that there is no reason to avoid using the word just because it is only used in the mainland. The same counter argument would also work here. The point here is that the test should be flexible enough to admit common words like these. Detailed guidelines for constructing a list of words and expressions that are bounded to each subculture will of course need more research and coordination.

It is easy for linguists and teachers to agree that the TOCFL should be scientific linguistically, neutral politically and flexible culturally. It is much harder, if ever possible, for all of them to agree to the means that would lead to such an aim. Generally speaking, items for the intermediate proficiency level are the easiest to write and agreed upon. Most of the weaknesses of the existing tests concentrate on either end of the test. Items intended to test the lower end of the proficiency are usually too difficult. The higher end of the proficiency is still a blue print for the time being, perhaps due to the complex linguistic and cultural issues involved.

This situation is easily explained from a sociolinguistic point of view. As the examinee knows more about the language, he is getting more and more into the cultural side of the language, such as allusion, connotation, historical background, political coloring, cultural value and association and so on. One can hardly exhaust such a list. The test of such cultural knowledge as associated with the Chinese language is a complex issue because of the diversity in the subcultures. This suggests that a lot of problems will be involved in designing the items for the advanced level of the test.

With the trend towards a global economy, communication among peoples with various language backgrounds becomes more and more important. The study of foreign languages, rather than something people play with in high school, has become directly connected with economic success. It is against such a background that proficiency tests of foreign languages have become an increasingly important issue.

If the economic power of the Chinese speaking communities in Asia and other parts of the world is something that economists cannot ignore, the various issues related to the teaching and testing of the Chinese language is also something that linguists cannot afford to ignore. With the prospect that Mainland China will eventually be part of the international family economically and politically, any investment in the research of the Chinese language will finally pay off. Moreover, the sheer number of speakers for the Chinese language all over the world makes it worthwhile to invest in a project like the TOCFL.

References:

Harris, David. 1969. Testing English as a Second Language. McCraw-Hill Book Company.

Kubler, Cornelius C. 1988. Chinese Grammar and Expression Check List. JCLTA Vol. XXIII: No. 1, pp57-85.

Liu, Yinglin (ed). 1989. *Hanyu shuiping kaoshi yanjiu* (Research on Chinese proficiency test). Xiandai Publishing House, Beijing.

Wang, Lih Shing and Stansfield, Charles W. 1988. Chinese Prodiciency Test: Test Interpretation Manual. CAL, Washington, D.C.