DOCUMENT RESUME

ED 394 321 FL 023 725

AUTHOR Cui, Songren
TITLE Assessment: Individual Achievement and Program Effectiveness.
PUB DATE Nov 95
NOTE 12p.; Paper presented at the Annual Meeting of the American Council on the Teaching of Foreign Languages (29th, Anaheim, CA, November 18-21, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Chinese; *Criterion Referenced Tests; Individual Differences; Language Proficiency; *Language Tests; *Norm Referenced Tests; Program Effectiveness; Program Evaluation; Second Language Programs; Second Languages; *Student Evaluation; Testing; Test Interpretation; Test Use

ABSTRACT

A discussion of student and program assessment in the context of Chinese language instruction looks at theory and methods of assessment, and proposes that norm-referenced language proficiency testing is inadequate to evaluate student achievement and program effectiveness. It is further proposed that criterion-referenced tests are more appropriate in assessing both student learning and program effectiveness in Chinese language programs. Deficits identified in norm-referenced tests currently in common use include discrepancies between instructional objectives and test content, descriptive ambiguity, inadequate test interpretation, threat to content validity due to built-in item selection that systematically eliminates "easy" items, and inappropriateness for summative selection. It is argued that criterion-referenced tests, by comparison, are based on instructional objectives, have descriptive clarity, can be interpreted according to degrees of mastery, and can include some easier items because there is no intent to maximize differences among individuals. (Contains 19 references.) (MSE)

Assessment: Individual Achievement and Program Effectiveness

ASSESSMENT: INDIVIDUAL ACHIEVEMENT AND PROGRAM EFFECTIVENESS

Songren Cui
University of Massachusetts at Amherst

1. Introduction

The past decade has witnessed rapid developments in Chinese language testing, evidenced by the emergence of various kinds of Chinese tests. The Oral Proficiency Interview (OPI) by ACTFL, the Simulated Oral Proficiency (SOPI), Chinese Speaking Test (CST), Chinese Proficiency Test (CPT), and Preliminary Chinese Pr . iency Test (Pre-CPT) by the Center of Applied Linguistics (CAL), SAT II Chinese Language Test by ETS, and Hanyu Shuiping Kaoshi (HSK -- *Chinese Proficiency Test*) by Beijing Language Institute, just to name a few of them. These tests have played, and will continue to play an important role in the measurement of individuals' language proficiency, particularly in helping decision makers to find out an individual's relative rank with reference to the performance of other individuals, and to determine whether an individual is qualified for a specific status either by professional or academic standards. In this aspect, the significance of these proficiency tests are undeniable, not to mention their profound implications for the instruction of Chinese.

It should be noted, however, that these tests are, by design and by nature, proficiency tests. They are not constructed to test the achievement of specific instructional objectives, nor purported to evaluate the effectiveness of a Chinese language program. When we stretch out to embrace these new tests, we must remember this "blockbuster truth: *Different educational purposes require differing educational tests and differing uses of those tests*" (Popham, 1981:10). This is similar to the common sense in everyday life. There are various instruments for measurement: for example, a ruler is for length and a scale for weight. But we are not supposed to measure weight with a rule. Unfortunately, as Popham points out,

> "There are too many educators who unwarrantedly assume that 'a test is a test is a test'. Even though this phrase possesses enticing poetic and metaphysical quality, it is also in error. To employ a test for instructional improvement that was created to sort out youngsters for other assessment purposes can be a serious mistake. Similarly, to use a test designed for ins'ructional improvement to try to evaluate a program's instructional effectiveness may turn out to be a major blunder" (Ibid).

In this paper, I intend to address two issues: 1) assessing individual students' achievement, and 2) evaluating the effectiveness of a Chinese language program. I will first brief some notions concerning testing and evaluation. I will then assert that using the above mentioned proficiency tests is inadequate for assessing achievement and evaluating program effectiveness. Next, I will discuss the needs for criterion-referenced tests in assessing achievement of Chinese language students and evaluating effectiveness of Chinese language programs, and the advantages of criterion-referenced tests. Last, I will suggest that it is necessary to set up a test group to further study these issues in order to improve Chinese testing practice in our field.

2. Language Testing and Program Evaluation

To facilitate the discussion, it is necessary to review a few terms related to language testing and program evaluation.

2.1 Norm-Referenced and Criterion-Referenced Tests

There are, basically, two major types of tests: norm-referenced tests and criterion-referenced tests. By definition, "A norm-referenced test is used to ascertain an individual's status with respect to the performance of other individuals on that test" (Popham, 1981:26). That is, by conducting a norm-referenced test, the tester seeks information about an individual's ranking percentile, or standing point with reference to other individuals who have taken the same test. The emphasis in developing a norm-referenced test, therefore, is to maximize the possible differences among testees, so as to distinguish one from another. The quintessential norm-referenced test is the "standardized test". This type of test has two distinctive features: 1) it is administered in a standard way under uniform condition and 2) it has been tried out with large groups of individuals, whose scores provide standard 'norms' or reference points for interpreting scores (Bachman, 1989:248). The familiar examples of norm-referenced tests are TOEFL, GRE, and SAT. In the case of Chinese language tests, the Hanyu Shuiping Kaoshi (HSK) by Beijing Language Institute, Chinese Proficiency Test (CPT), and Preliminary Chinese Proficiency Test (Pre-CPT) by the Center of Applied Linguistics (CAL), and SAT II Chinese Language Test by ETS all fall into this category. The test scores of a norm-referenced test is interpreted and reported *relatively*. For example, a college student Susan who is enrolled in an Intermediate level Chinese language class took a Chinese Proficiency Test (CPT) and scored 110 (not the raw score) on the Listening Comprehension section. According to the CPT Interpretation Table, her percentile is 81%, which means that Susan equaled or exceeded the performance of 81% of the Intermediate Level examinees in the entire group who took the CPT in the Listening Comprehension section. It is because we make our interpretation of examinee's scores by relating or referencing them to that of examinees in the norm group that we refer such tests as *norm-referenced tests*.

Criterion-referenced tests, on the contrary, is used to ascertain an individual's status with respect to a defined behavioral domain. Its focus is on the nature of an examinee's test performance, and the interpretation of the test scores is *absolutely*. That is, the tester tries to find out whether an individual can demonstrate one's mastery of a defined set of criterion behaviors, or a behavioral domain. Thus, a basic requirement of developing criterion-referenced tests is to specify a context domain or criterion of performance. Since such domain is often specified in terms of behavioral objectives through instruction, criterion-referenced tests are sometimes also referred to as "*objective-based*" tests. A typical example of is teacher-made classroom tests whose items have been constructed to measure an instructional objective. Objective-based tests have a long history, and are still widely used in most classroom testing contexts. However, because of the brevity of behavioral objectives which usually describe intended learner outcomes in short-hand, "most of objective-based tests fall far short of carving out the well-defined behavioral domain that constitutes the essence of any truly virtuous criterion-referenced test" (Popham, 1981:29). Unfortunately, this is also true to the situation of Chinese language testing. In a strict sense, criterion-referenced tests should be domain-referenced ones which is broader and more precise than objective-based tests, although the two notions of criterion-referenced tests and domain-referenced tests are essentially interchangeable.

2.2 Formative and Summative Evaluation

Evaluation is a set of complex interactive activities which defined by Brown in these words:

> *Evaluation* is the systematic collection and analysis of all relevant information necessary to promote the improvement of a curriculum, and assess its effectiveness and efficiency, as well as the participants' attitudes within the context of the particular institutions involved. (Brown, 1989:223)

There are, at least three dimensions, or perspectives from which to evaluate a given program. These dimensions are: formative versus summative, process versus product, and quantitative versus qualitative. For the specific purpose of our discussion, this paper will concentrate on formative and summative evaluation only.

The terms of formative and summative evaluation are coined by Scriven (1967) in the context of curriculum evaluation. Formative evaluation is conducted during the development of a program and in the process of the implementation of its curriculum. The purpose is, then, to improve instruction and the program. Its major concerns are: 1) to determine the results of the program, and 2) to diagnose strength and weakness of the program. Formative evaluation is generally conducted for in-house staff and normally for remains in house. It may, however, be done by an internal or external evaluator, or, ideally, a combination of both. The decisions made from such evaluation are relatively small scale and numerous, and will result in modification and fine tuning of the current program design, such as revising instructional objectives, adjusting teaching methods, upgrading education facilities, changing curriculum content, and staff selection or development. Summative evaluation, on the other hand, takes place when a program is completed. Its purpose is to determine whether a program is successful and effective enough to adopted. Evaluation of this type provides information for educational policy makers and administrators. Their decisions from such analysis will be considerably large scale, and may affect the education system at different levels (for example, the continuation of funding a program or its cancellation). Thus summative evaluation contributes to public relations and aids planning.

According to Cronbach et al. (1980), formative evaluation is more influential and significant than summative evaluation because while the program is still fluid, such evaluation will contribute more to the improvement of education than evaluation used to appraise a product that is has already been on the market. As people are more reluctant to tear apart a supposedly finished product, evidence of weakness and suggestions of improvement provided midway will have a greater impact through the process of program development. Moreover, providing feedback to both teachers and students about success or failure in mastering the specified skills or content in the curriculum is an essential part of the teaching-learning process. In Brown's (1989) view, virtually all education evaluation should be formative, for all information is to be used to modify and improve teaching and learning. "Typically, language programs are ongoing concerns that do not conveniently end; consequently, a summative evaluation is difficult to perform" (Brown, 1989:230). Hence, he suggests that formative evaluation should be conducted constantly with the purpose of providing information and analysis that will be useful for program improvement.

3. Inadequacies of the Existing Chinese Tests for Achievement Assessment and Program Evaluation

Although some of the above-mentioned norm-referenced Chinese tests claim that they have the function of "evaluation of Chinese instructional programs"[1], I would like to echo the argument of a number of evaluation researchers (for example, Weiss, 1972; Millman, 1974; Baker, 1974; Popham, 1978 and 1981; Bachman, 1989) that standardized tests are inadequate for the purposes of assessing student achievement and evaluating a program, either formatively or summatively.

The first defect in the use of norm-referenced tests for assessing students achievement and evaluation program effectiveness is the discrepancies between instructional objectives and testing content. Since all standardized-test publishers, especially those commercial firms, have tried hard to sell their tests as widely as possible, they don't want to tie up with any curriculum objectives, any particular programs or any institutions. In the case of Chinese testing, though testing agents such as CAL, BLI, or ETS may not necessarily have the same motivation in the development of those Chinese proficiency tests, they would certainly like to see their tests have a greater scope of adoption. However, given the fact that there are hundreds of Chinese programs in the United States, and their instructional objectives and text materials vary so much, it is almost impossible to use a few standardized tests to measure the achievement of the instructional objectives, particularly for beginning and intermediate levels. This, in turn, prohibits the Chinese proficiency tests from supplying sufficient information for instructional improvement, which is the major purpose of formative evaluation.

A second deficit of norm-referenced tests for achievement assessment and program evaluation is their descriptive ambiguity. As norm-referenced tests attempt to have a wide distribution, and to measure a more general category of examinees' competencies (for example, reading comprehension), the description of what to measure in a specific test is very general. This is typical for those commercial standardized-tests, the abilities to be measured can be interpreted only indirectly with reference to specific instructional objectives. Such a vagueness in description of the measured competencies can hardly provide clear information about achievement and effectiveness, thus making the program evaluator unable to offer on-target assistance to teachers and administrators formatively.

A third weakness in using norm-referenced tests for achievement assessment and program evaluation results from the interpretation of test scores. When an individual's scores of proficiency standardized tests are interpreted and reported by percentile with reference to others' performance, the teacher will have no idea to what degree this student has mastered the skills or knowledge. For achievement tests, teachers are more concerned about the degree of students' mastery of certain content or skills as defined in the curriculum objectives. Knowing a given individual's status in relation to others, even when it is a high percentile such as the 90th, may not be sufficient to tell what one can do, and how well he or she can do. If the entire group of examinees happen to perform below a standard that we are willing to accept as an indicator of mastery, choosing a 90th percentile of that group is just like what the Chinese saying goes: *to pick a general from among the dwarfs*. Since we don't know the degree of a student's mastery, we can't pinpoint his or her deficit. Accordingly, we are unable to provide the necessary remediation to get a solid fix on the particular skill a student has not mastered -- it defeats the purpose of formative evaluation.

Another problem of norm-referenced tests for evaluative purpose is the threat to the content validity due to their built-in tendency of item selection which systematically eliminates the "easy" items. In order to maximize differences among individuals, norm-referenced tests usually use certain statistical criteria to select items that are of medium difficulty and that can discriminate well between high and low groups of examinees. If a test item is answered correctly by most examinees, it will be considered to have very low or even no discriminating efficiency. Generally, such items are to be discarded because their inadequate contributions to response variance, that is, they will lower the descriminatabilty of the entire test. By doing so, a lot of items that emphasize the significant aspects of instructional objectives will be sifted out. However, if we want to assess students' achievement, shouldn't we test their mastery of the knowledge and skills defined in the curriculum? If we want to evaluate a program to improve instruction, aren't we going to measure the outcome and see how effective our teaching is and how well students have learned? When we use norm-referenced tests to assess students' achievement or to evaluate a Chinese program, we may, unawarely, put ourselves in two conflicting positions.

Finally, norm-referenced tests is inadequate for summative evaluation purpose, either. The focus of summative evaluation is to determine whether a given program is sufficiently effective to adopt or to continue implementing. To make the decision, information that is relevant to both the defined curriculum objectives and unexpected outcomes is absolutely indispensable. As mentioned previously, norm-referenced test scores are inappropriate to be used as indicators of students' achievement of the instructional objectives. Although sometimes a few standardized tests such as TOEFL may be used as indicators of broader outcomes, there is certain limitation. That is, in Bachman's view, "the program developer or evaluator is willing to accept the definition of language abilities that informs the test" (Bachman, 1989:250). As far as for Chinese instructors concern, the degree of willingness seems much lower to accept the definition of Chinese language abilities that the CPT, Pre-CPT, SAT II Chinese Language Test are based -- namely, the ACTFL Chinese Proficiency Guidelines, primarily for its strong bias toward the western languages and the misrepresentation of the nature of Chinese language.

In addition, when we use a norm-referenced test for summative evaluation, especially when we try to make comparisons between two programs that may already differ from each other in their curriculum objectives, we are imposing a third set of objectives (represented by the test) onto these two programs which may or may not relevant to these two programs. Not to mention the norms to which the test is referenced to may not necessarily be appropriate to those students in the program.

Theoretically speaking, all the Chinese speaking tests mentioned above, that is, the Oral Proficiency Interview (OPI) by ACTFL, the Simulated Oral Proficiency (SOPI), and Chinese Speaking Test (CST) by Center for Applied Linguistics (CAL) are criterion-referenced tests because an individual's test scores are not interpreted with reference to others' performance, either in the same group nor in a "norm" group. They directly reference an individual's performance to a defined behavioral domain, or a level of proficiency as specified in the ACTFL Proficiency Guidelines. However, being criterion-referenced tests themselves may not guarantee that these tests can fulfill the task of assessing individual student's achievement nor evaluating effectiveness of a Chinese program. The reasons have been summarized by Bachman, which I can't agree more:

> The major problems in measuring learner achievement in language program evaluation in the past have, I believe, been twofold: i) the inadequacies of norm-referenced measurement theory and of tests developed within this theory for addressing the needs of program evaluation, and ii) the incompleteness of our definition of language proficiency.
>
> (Bachman, 1989:243)

Because these three tests, the OPI, SOPI, and CST were all constructed according to the same description of behavioral domain, or to be more precise, levels of proficiency, and correlated with each other, none of them can be immune from any flaws of that given specification of levels of proficiency. (For discussions of the limitations and flaws of the ACTFL Guidelines, please refer to Cui, 1993 and 1994). In addition, it is also a problem to use the term *criterion* (as a desired behavior conception) to refer to a level of proficiency. According to Popham,

> It is now apparent that to interpret *criterion* as a level of examinee proficiency yield almost no dividends over traditional testing practices. In fact, by using that conception of criterion, one could magically transform any norm-referenced test into a criterion-referenced test merely by setting a specific proficiency level for the test. If criterion-referenced tests are going to constitute a unique contribution to our measurement arsenal, it will be because they yield a more accurate depiction of an examinee's performance, not in relative terms, but in absolute terms. In other words, if criterion-referenced testing is going to provide any substantial payoff, it will be because we can secure a more precise notion of an examinee's status with respect to a clearly delimited domain of behaviors. The contribution to educational measurement that criterion-referenced tests are supposed to make is predicated on their *increased descriptiveness*. (Popham, 1981:28)

A level of proficiency, for example, the Intermediate-high in speaking as defined in the ACTFL Guidelines, covers an array of subskills, functions, and grammatical structures. It is quite possible that an individual can make "simple comparisons", but fails to sustain the correct use of particles of *le, zhe, guo* as stipulated by the Guidelines to show one "has flexibility in expressing time relationships"[2]. However, because the ACTFL Guidelines contain "a non-comp the de ensatory core", the rating of Intermediate-high cannot be assigned to that individual even if other aspects of performance signal the given level. Apparently, such a conception of criterion is too broad, and is inappropriate to be used for achievement assessment.

The last argument that these tests are not appropriate for assessing individual achievement and evaluating program effectiveness is quite self-evident. If we want to measure the learning outcome after a period of time, the test should target on the scope of instruction, both knowledge and skills. Likewise, when we evaluate a program, any tests used as a component of the evaluation to collect information must be relevant to the program to be evaluated. If the tests are irrelevant, or in Scriven's term, *program-free* (Scriven, 1974), they can't provide the needed information. As the ACTFL Proficiency Guidelines are designed to test proficiency, they are not sensitive to any particular curriculum, textbook, teaching methodology, or language program. Accordingly, tests such as OPI, SOPI, CST can hardly meet the needs of achievement assessment or program evaluation, especially for students who are at the lower end of the Chinese proficiency.

4. The Needs for Criterion-Referenced Tests of Chinese Language

For purposes of achievement assessment and program evaluation, criterion-referenced tests have several advantages over norm-referenced tests. First of all, unlike norm-referenced tests which are "goal-free" or "program-free", criterion-referenced tests are based on the behavioral domain of instructional objectives. As these tests measures the intended outcomes only, they can provide direct and detailed information about students' achievement and avoid any possible mismatch between what has been taught and the what is tested. Secondly, the descriptive clarity of well constructed criterion-referenced tests makes these tests ideal for planning on-target instructional activities. In addition to designing instructional sequence to promote a desired outcome, criterion-referenced tests can help teachers to locate a student's specific problem or to discover the particular skills a student lacks. Criterion-referenced tests can supply information for diagnosis whereas norm-referenced tests can't. Third, an individual's test scores of criterion-referenced tests are interpreted and reported in relation to the degree of mastery of the instructional objectives rather than to other examinees' ranking. This information is much more meaningful for both teachers and students to find out the strengths and deficiency in their teaching and learning. Finally, those "easy" test items that usually bear emphasis of instructional objectives need not be eliminated, since the focus of criterion-referenced tests is not to maximize the differences among individuals. The problem of content validity caused by item selection in standardized tests can be avoided. All these taken into consideration, naturally, criterion-referenced tests are a better choice than norm-referenced tests, either for achievement assessment or for program evaluation. When the above-mentioned Chinese proficiency tests cannot serve the purpose of assessing individual's achievement and evaluating program effectiveness, accordingly, there is a need to look for other alternatives.

5. Conclusion

To date, according to my own observations, most of us rely heavily on teacher-made tests to assess student achievement, and program evaluation is dominantly based on qualitative data collected from class observations and teacher/student interviews. Criterion-referenced tests are, rarely used for both purposes of assessing student achievement and/or evaluating program effectiveness. While there is nothing wrong with using teacher-made tests and collecting qualitative data per se, well-constructed criterion-referenced tests can certainly help us do a better job in achievement assessment and program evaluation. We cannot always follow the beaten path, if we want to upgrade our instruction and main-stream Chinese teaching in the United States. This becomes even more urgent as we are approaching the twenty-first century, the so-called Pacific-Rim-Century. I suggest that, therefore, we should set up a special task force, say a testing group, to further study these issues, and to coordinate research on testing and evaluation. With collective effort and proper management, the dream is not absolutely unrealizable to develop practical and valid criterion-referenced tests that can provide us with meaningful information about students' achievement, and about the effectiveness of a Chinese language program.

[1]. Chinese Langauge Testing Program, Center for Applied Linguistics. 1992. *Chinese Proficiency Test and Preliminary Chinese Proficiency Test Combined Test Interpretation Manual*. Washington, D.C. p. 4.

[2]. 'ACTFL Chinese Proficiency Guidelines'. *Foreign Language Annals*. 20.5:474.1987.

REFERENCES

American Council on the Teaching of Foreign Languages (ACTFL). 'ACTFL Chinese Proficiency Guidelines.' *Foreign Language Annals.* 1987, 20.5: pp.47-481.

Bachman, Lyle. 1989. The development and use of criterion-referenced tests of language ability in language program evaluation. In Robert K. Johnson (Ed.) 1989. *The Second Language Curriculum.* pp. 242-258. Cambridge: Cambridge University Press.

Backer, Eva L. 1974. Formative evaluation of instruction. In James W. Popham (Ed.) 1974. *Evaluation in Education: Current Applications.* pp. 533-85. Berkeley, California: McCutchan.

Brown, James D. 1989. Language program evaluation: a synthesis of existing possibilities. In Robert K. Johnson (Ed.) 1989. *The Second Language Curriculum.* pp.222-41. Cambridge: Cambridge University Press.

Chinese Language Testing Program, Center for Applied Linguistics. 1992. *Chinese Proficiency Test and Preliminary Chinese Proficiency Test Combined Test Interpretation Manual.* Washington, D.C. p.4

Cui, Songren. 1994. Taking ACTFL guidelines as curriculum objectives: some considerations. *Journal of Chinese Language Teachers Association.* 29.2: pp. 47-69.

Cui, Songren. 1993. Conceptualizing language proficiency. *Journal of Chinese Language Teachers Association.* 28.2: pp.1-23.

Cronbach, Lee J., S. R. Ambron, S.M. Dornbusch, R.D. Hess, R.C. Hornick, D.C. Phillips, D.F. Walker and S.W. Weiner. 1980. *Toward a Reform of Program Evaluation.* San Francisco, California: Jossey-Bass.

Liu, Yinglin. (Ed.) 1989. *Hanyu Shuiping Kaoshi Yanjiu.* (Research on Chinese Proficiency Testing). Beijing: Modern Press.

Lynch, Brian. 1992. Evaluating a program inside and out. In Charlse J. Alderson and Alan Beretta (Eds.) 1989. *Evaluation Second Language Eduation.* pp. 61-99. Cambridge: Cambridge University Press.

Millman, Jason. 1974. Criterion-referenced measurement. In James W. Popham (Ed.) 1974. *Evaluation in Education: Current Applications.* pp. 309-97. Berkeley, California: McCutchan.

Office of National Chinese Proficiency Testing Committee. 1993. *Hanyu Shuiping Kaoshi Dagang*. (Testing Syllabus of Chinese Proficiency Test). Beijing: Modern Press.

Popham, James W. 1988. *Educationo! Evaluation*. Englewood Cliffs, New Jersey: Prentice-Hall.

Popham, James W. 1981. *Modern Educational Measurement*. Englewood Cliffs, New Jersey: Prentice-Hall.

Popham, James W. 1978. *Criterion-Referenced Measurement*. Englewood Cliffs, New Jersey: Prentice-Hall.

Popham, James W. 1974. (Ed.) *Evaluation in Education: Current Applications*. Berkeley, California: McCutchan.

Scriven, Michael S. 1967. The methodology of evaluation. In Ralph Tyler, Robert Gagne and Michacl S. Scriven. *Perspectives of Curriculum Evaluation*. Series on Curriculum Evaluation. No. 1. pp.39-83. Chicago, Illinois:Rand McNally.

Scriven, Michael S. 1974. Evaluation perspectives and procedules. In James W. Popham (Ed.) 1974. *Evaluation in Education: Current Applications*. pp.3-93. Berkeley, California: McCutchan.

Weiss, Carol H. 1972. *Evaluation Research: Methods of Assessing Progr:n Effectiveness*. Englewood Cliffs, New Jersey: Prentice-Hall.