15 *The development and use of criterion-referenced tests of language ability in language program evaluation*[1]

Lyle F. Bachman

Introduction

The role of measurement in program evaluation has become increasingly unclear in the past few years, and this is for several reasons. First, evaluators have come to recognize that the processes that take place in a language program are at least as important as the products of the program. That is, what happens in the classroom is of interest not only because of its relationship to program outcomes, but in its own right. Most of these processes, indeed those that are of greatest interest, that take place in the minds of learners, are extremely difficult to measure. A second reason for questioning the role of measurement in program evaluation is the emphasis that many current evaluation theories place on qualitative sources of information, in some cases to the virtual exclusion of quantitative data of any kind. And finally, there are the inadequacies of norm-referenced tests for the purposes of program evaluation and the negative repercussions of their continued use for these purposes.

But rather than causing us to abandon tests as part of language program evaluation, I believe these concerns must lead us to a reorientation in our thinking about the needs for tests and about the types of tests we need. The need to examine the processes of instruction and learning presents a different set of measurement problems from those encountered in measuring outcomes. For example, the use of communication strategies and the amount of comprehensible input a learner obtains are areas of considerable interest not only for second language acquisition research but, it would seem, for language program evaluation as well. These present a challenge to measurement that is both intriguing and formidable, but not, I believe, insurmountable.

In addition to processes, we need to examine the products of the language program. We must pay attention to a wide range of program

1 This is a revised version of a keynote paper presented at the Conference on Trends in Language Programme Evaluation, Bangkok, Thailand, 9–11 December 1986.

https://doi.org/10.1017/CBO9781139524520.017 Published online by Cambridge University Press

outcomes, including the effects of the program on teachers, the school system, and the community. Learner achievement has historically been a major concern of evaluation, and I believe that this will continue to be of prime interest in the future. It is here, of course, that measurement has played the largest role. And while we must not ignore the vital importance of qualitative information about outcomes, such as classroom observation or teacher reports, I agree with Popham (1978) that 'pupil test performance will always play a pivotal role in any approach to evaluation' (p. 4).

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. This is not to say, however, that these problems must persist. The past decade has seen rapid developments in criterion-referenced testing, so that an emerging measurement technology is now available for application to the needs of language program evaluation (for example, papers in Berk, 1980a, 1984a). This period has also witnessed an expansion in the definition of language proficiency that recognizes both the context – discourse and sociolinguistic – in which language use takes place, and language use as a dynamic negotiation of meaning (for example, Canale and Swain, 1980; Johnson, 1982a; Canale, 1983; Savignon, 1983; Candlin 1986). These developments in criterion-referenced test theory and the broader definition of language ability provide, I believe, the keys to developing language tests that are appropriate to the needs of language program evaluation.

In this paper I will focus on considerations in the development and use of tests of learner outcomes, or achievement, if you will, for purposes of program evaluation. While the measurement of processes is an area that must be addressed, this is ultimately related to our understanding of the developmental process of second language acquisition itself, and consideration of this would take us far beyond the concerns of program evaluation. Similarly, the use of qualitative, that is, non-test information, though essential to evaluation, is beyond the scope of this paper. I will first discuss the needs for measuring outcomes in program evaluation – both formative and summative. I will then present the argument that current language testing practice is inappropriate to these needs. Next, I will outline the fundamental requirements of criterion-referenced tests, and suggest a theoretical framework as a starting point for developing criterion-referenced tests of communicative language ability. Finally, I will repeat a call for a program of research that John Clark and I have outlined elsewhere.

The measurement of learner outcomes for program evaluation

In attempting to delineate the role of tests of learner outcomes in program evaluation, we might begin by considering the range of situations in which evaluation takes place, the types of evaluation decisions that are to be made, and the criteria by which the program will be evaluated. At one end of the range of situations we have contexts in which there is no formal program, but in which there may be some learning and teaching, and in which evaluation might be of interest. If, for example, I were hired as a language tutor by a non-native English speaking student, I might simply arrange to spend some time with him conversing in English. If the student's objective were to attend a college or university in the United States, the evaluation of my effectiveness might come in the form of a TOEFL score. If this were unacceptably low, it would probably terminate my career as a TOEFL preparation coach, irrespective of how interesting I might be as a conversationalist. (It is also possible, of course, that I might succeed despite the fact that I bored my student to tears.) If, on the other hand, the student were more interested in making friends with a native English speaker than in studying in the United States, his TOEFL score might not even be considered relevant to evaluating our interaction.

Formative and summative evaluation

Most of us, however, are concerned with evaluation in the more complex context of a formal instructional program. And here, we find a great deal of variety not only in types of programs, but in the types of decisions to be made. In some situations, we may be interested in identifying ways to improve an on-going program, to upgrade the *status quo*, so to speak, at minimal cost, while in others we may be committed to developing the best 'new and improved' program possible, irrespective of cost. In situations such as these, we will be primarily concerned with what Scriven (1967) has called *formative evaluation*, which is essentially evaluation for the purpose of improving instruction. Formative evaluation takes place during the development of a program, and its major concerns are to determine what the results of the program are, and to diagnose areas of strength and weakness in order to improve instruction. In other situations, we may need to find out whether a new textbook, set of materials, or teaching technique is better than that currently used, or we may need to choose among two or more competing curricula. In contexts such as these, we will be concerned with *summative evaluation*, which typically takes place after the program is complete, and provides information that is relevant to deciding whether or not to adopt a new

program. Formative and summative evaluation can thus be distinguished in terms of the types of decisions to which they are addressed (course improvement vs. adoption) and the point at which they typically take place (during a program vs. after the completion of the program).

Objectives-based and program-free evaluation

The criteria by which the program is to be evaluated are also relevant to the measurement of learner outcomes. The primary question is whether we will limit our evaluation to the stated instructional objectives, or whether we will also look at unstated and unexpected outcomes. Here, we can refer to the distinction that Scriven (1974) has made between objectives-based evaluation and so-called 'goal-free', or program-free evaluation. Traditionally, both formative and summative evaluation have focused on examining the extent to which the program has attained its stated objectives. As Scriven has pointed out, however, the problems with this are twofold: i) it ignores potentially important outcomes, both positive and negative, that program developers have not included, either inadvertently or by choice, and ii) it thus introduces bias into the evaluation, in the sense that it looks only at those outcomes which are valued and for which there is a high probability of success. In advocating program-free evaluation, Scriven argues for considering the broader educational and social context in which programs take place, and looking not only for unanticipated outcomes, but also at the consequences of achieving the stated objectives.

My favorite illustration of the failure to fully consider the consequences of achieving one's objectives comes from a curriculum research and development program in which I was involved several years ago. This was a program to develop an English language arts program for elementary school children in a country where English is a foreign language. (I should also mention that the school system in this country was very traditional and quite authoritarian in its structure and administration.) In addition to promoting the development of skills in English, our program aimed at developing attitudes and skills to promote self-directed study, which we felt would be useful in both the students' English class and in their subject matter classes. After the first year of the program, it was clear that both the language skills and self-direction objectives were succeeding beyond our expectations at one of our demonstration schools. Our flush of success, however, soon turned to the blush of chagrin, when this group of self-directed elementary school children marched *en masse* into the principal's office to protest the assignment of their teacher to another class! We had failed to consider the possibility that self-direction could generalize from study habits to social action.

The implication of program-free evaluation for formative evaluation is that we should be concerned with improving the program so that it not only achieves its stated objectives, but is also consistent with the broader goals of the education system and society. Similarly, in summative evaluation we should compare outcomes of competing programs not only in terms of their stated objectives, but also with reference to educational and societal values.

While there is considerable variation in the types of programs we evaluate and in the types of decisions to be made, the kinds of information about learner outcomes that are needed for all these contexts and decisions are quite similar. If we accept the principle of program-free evaluation, there are two types of information about learner outcomes that are essential to both formative and summative evaluation: i) detailed information related to students' achievement of stated instructional objectives (both cognitive and affective), and ii) information related to possible outcomes that are not included as instructional objectives.

The first problem in measurement, then, is to identify the two types of outcomes to be measured. In formative evaluation, identifying instructional objectives is relatively unproblematic. This is also the case for summative evaluation, even though we may be dealing with differing sets of objectives. What is particularly problematic for summative evaluation, however, is the identification of the broader set of possible outcomes, from which the sets of instructional objectives are presumably drawn. The relationships between instructional objectives and possible outcomes in summative evaluation can be illustrated as in Figure 1:

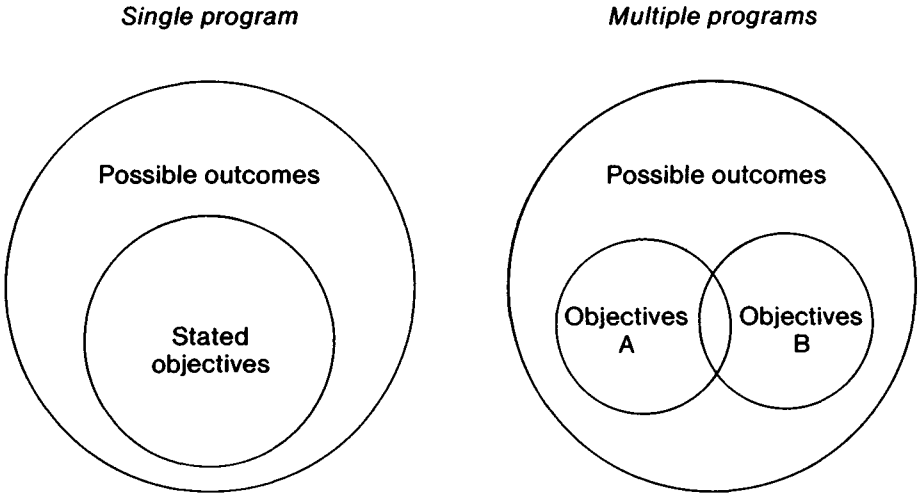

Figure 1 Stated objectives and possible outcomes in summative evaluation

https://doi.org/10.1017/CBO9781139524520.017 Published online by Cambridge University Press

To illustrate these relationships, first consider the evaluation of an English program for individuals whose primary use of English will be to pass an examination that will admit them to a higher level of education. Suppose the instructional objectives of this program were primarily in the areas of grammar and vocabulary. We might find that students in this program perform well on tests of grammar and vocabulary, and that the majority of students who successfully complete this program also obtain high examination scores and subsequent admission to higher education. On the basis of this information, we could say that the program is successful in meeting its objectives. But would we, as program-free evaluators, be satisfied that the program is producing students who are proficient in English? If the developers of the program argue that this is totally irrelevant to the purposes of the program, we might just as easily ask why the program is called 'English', rather than, say, 'test preparation'.

The problem is not different, but is simply multiplied, in the case of two or more competing programs. Suppose, for example, we were asked to compare a course based on a structural syllabus with one based on a functional syllabus, each of which was very successful in meeting its objectives. Here, the impossibility of choosing either as the 'better' program without recourse to a broader set of outcomes is clear. This is because of the distinctiveness of the objectives, and because few of us would accept knowledge of either structures or functions alone as a complete definition of language proficiency. Unfortunately, for evaluators, at least, the distinction between competing programs is seldom this clear, and we are typically faced with choosing between programs with similar objectives. Here, the problem is not so much that of comparing apples with oranges, as comparing *jonathans* with *macintoshes*. The problem, however, is qualitatively the same: unless the two programs have identical sets of objectives (and processes, I might add), they cannot be reasonably compared except with reference to a broader set of outcomes.

To summarize, for virtually every context and purpose of program evaluation, we need to gather detailed information about learners' achievement of instructional objectives. In formative evaluation we need information that is precise enough to permit revision, or fine-tuning, of the program. In summative evaluation I have argued that we must also gather information about outcomes that are not stated as objectives. In evaluating a single program we must examine not only the effectiveness of the program in achieving its objectives, but also the extent to which the outcomes, both stated and unexpected, are worthwhile in the broader context of the education system and society. In evaluating two programs, we cannot use the objective set of either as a criterion, but must use a framework that ideally includes both.

Norm-referenced and criterion-referenced tests

For the past twenty-five years the dominant measurement approach to the development and use of language tests has been that of norm-referencing. This approach characterizes what Spolsky (1978) has referred to as the 'psychometric-structuralist' period of language testing. The distinguishing characteristic of the norm-referenced approach to testing is that test scores are reported and interpreted with reference to the performance of individuals, either in the same group, or in a 'norm' group. That is, a norm-referenced test score provides information about an individual's relative rank with reference to other individuals who have taken the test. The emphasis in developing norm-referenced tests, therefore, is on maximizing differences among individuals. The quintessential norm-referenced test is the 'standardized test' that has two distinguishing characteristics: i) it is administered in a standard way under uniform conditions and ii) it has been tried out with large groups of individuals, whose scores provide standard 'norms' or reference points for interpreting scores.

The other major approach to measurement, that of criterion-referenced testing, has a much longer history and is alive and well in most classroom testing contexts. Nevertheless, with the exception of Cartier's (1968) seminal article and a recent paper by Hudson and Lynch (1984), this approach has been virtually ignored by language testing researchers and writers of texts on language testing. In contrast to scores on norm-referenced tests, criterion-referenced test scores are reported and interpreted with reference to a specific context domain or criterion of performance. They thus provide information about an individual's mastery of a given content domain, or level of performance. One requirement of criterion-referenced test development and interpretation is the specification of a content or ability domain. Because the domain specification is frequently made in terms of instructional objectives, criterion-referenced tests are sometimes also referred to as 'objectives-based' tests.

The inadequacies of standardized norm-referenced tests for purposes of measuring the achievement of instructional objectives have long been recognized (for example, Glaser, 1963; Popham and Husek, 1969; Millman, 1974; Popham, 1978; Cziko, 1983). To mention just one problem, in the classroom setting we are often interested in knowing whether our students have mastered a given set of learning objectives. Knowing that a given student scored in the 90th percentile on the test will not provide this information, since it might well be that the entire group's performance is below a standard that we would be willing to accept as an indication of mastery. Thus, in most classroom uses, criterion-referenced

achievement tests are more appropriate than standardized norm-referenced tests.

While there is no necessary connection between the approach we adopt for developing and using language tests and how we define the abilities measured, over the past twenty-five years we have seen the development of a *de facto* union between the norm-referenced approach and 'language proficiency', as opposed to the criterion-referenced, objectives-based approach and 'language achievement'. And while this distinction may be perfectly serviceable for the varied needs for evaluating individuals, I believe the needs of program evaluation are such that this distinction cannot apply, and that in order to meet these needs we must consider a framework for developing and using criterion-referenced tests of language proficiency.

Inadequacies of norm-referenced tests for formative evaluation

As mentioned above, for the purposes of formative evaluation, we need to gather information about learner outcomes that is detailed enough to guide program revision. A number of evaluation researchers (for example, Weiss, 1972; Millman, 1974; Baker, 1974; Popham, 1978) have argued that standardized tests are unsuitable for this purpose because, in general, they are descriptively inadequate. First, the abilities measured by standardized tests of language proficiency are defined in very general terms, so that they can be interpreted only indirectly with reference to specific instructional objectives. Second, there is the frequent mismatch between instructional objectives and the content of a standardized test. Given the wide variety of contexts in which language is taught around the world and the resultant diversity of instructional objectives, it is unreasonable to expect a single standardized test to measure the objectives of any given program with enough detail to be useful for formative evaluation. A third weakness, mentioned earlier, is that scores on standardized tests provide no information about the degree of mastery of content or skills. And finally, there is the problem of content validity caused by the use of statistical criteria in selecting items for standardized tests. In order to maximize individual differences, only those items that are of medium difficulty and that discriminate well between high and low groups of test takers can be included. What this means is that in the process of developing a standardized test, 'easy' items get weeded out. One of the reasons these items are easy is that they have been included in the instructional objectives, and students have mastered their content. Thus, norm-referenced test development procedures tend

to eliminate those items whose content is of the greatest interest for formative evaluation.

Proponents of the criterion-referenced approach to test development and use argue that this approach provides a means of solving the problems associated with norm-referenced tests (for example, Hively *et al.*, 1973; Popham, 1978). First, since criterion-referenced tests are based on the domain of instructional objectives, they can provide direct and detailed information about students' achievement of those objectives. Basing the content of the test on specific objectives also avoids the mismatch between teaching and test content that is often found with standardized tests. Second, criterion-referenced test scores are reported in terms of the relative degree of mastery of the instructional objectives, frequently as a percentage, thus providing useful information for diagnosing both strengths and deficiencies in learning. Finally, because there is no need to maximize inter-individual differences, test items that are too easy or which do not discriminate need not be eliminated.

Inadequacies of norm-referenced tests for summative evaluation

In the context of summative evaluation, where the focus is on deciding whether a given program is sufficiently effective to implement, the evaluator needs to gather information that is relevant to *both* the stated instructional objectives and unexpected outcomes. Here, standardized test scores are inadequate as indicators of the achievement of instructional objectives for the same reasons as given earlier with respect to formative evaluation. As indicators of broader outcomes, however, standardized tests can provide useful information for the summative evaluation of a single program. The two standardized tests of English that have probably been used the most widely for this purpose are the *Test of English as a Foreign Language* (TOEFL, 1987) and the family of tests that have been developed over the years under the auspices of the British Council (Carroll, 1981; Seaton, 1983; Davies, 1984). The primary limitation to this use of standardized tests is the extent to which the program developer or the evaluator is willing to accept the definition of language abilities that informs the test.

It is in the context of summatively evaluating two competing programs that the delineation and measurement of learner outcomes poses the greatest problems. Since we virtually never compare two programs that are identical in their objectives, we are almost always faced with a dilemma in deciding what objectives to cover in the test, and whether to use a criterion- or a norm-referenced approach. The objectives-based solution to this problem is to base the test on both sets of instructional objectives. This can be done either by preparing a separate objectives-

based test for each program or by developing a single objectives-based test–that covers the instructional objectives of both programs. This approach has generally proven unsatisfactory because test results reveal, to no one's surprise, that students perform best on measures of those objectives they have been taught, and the test scores thus provide little information that is of use in choosing between the two programs. The norm-referenced approach – giving a standardized test to both groups and then comparing their scores – has also proven unsatisfactory. It may be unfair to both programs, since it evaluates them in terms of a third set of objectives – those of the test – that may or may not be related to the objectives of either program. In addition, the norms to which the test is referenced may be inappropriate to the students in the program.

Criterion-referenced tests of language proficiency

In the remainder of this paper, I will outline briefly what I believe to be a viable approach to developing measures of learner outcomes for use in the evaluation of language programs. This approach is not based on a new theory of measurement, and does not involve a radical departure from current views of the nature of language abilities. On the contrary, it simply involves the combination of the criterion-referenced approach to test development with a current specification of the domain of language proficiency.

In order for criterion-referenced tests to provide information about an individual's relative mastery of a given domain of ability, the content of such tests must be sampled from a well-defined domain of ability (Glaser 1963; Nitko, 1984). An additional requirement, if the results of criterion-referenced tests are to be applicable to the needs of both program-free formative, and summative program evaluation, is that the scores must be referenced to an absolute scale of ability. There are thus two issues to be addressed in developing criterion-referenced tests of language proficiency for use in language program evaluation: i) specifying the ability domain, and ii) defining the end points of ability so as to provide an absolute scale.

Communicative language ability

The evidence from recent language testing research is generally consistent with the hypothesis that language proficiency consists of several distinct abilities that are related to each other or which are related to a higher order, general ability (for example, Bachman and Palmer, 1982; Vollmer and Sang, 1983; Carroll, 1983; Upshur and Homburg, 1983; Oller, 1983). Many language testing researchers are thus focusing their efforts on identifying the various abilities involved in using language to commu-

nicate, and are for the most part working within an expanded framework of what I choose to call 'communicative language ability'. I define communicative language ability as consisting of both knowledge, or competence, and skill in implementing, or executing that competence, and my framework of CLA includes three components: language competence, strategic competence and the psychophysiological skills required to implement these abilities in language use. This framework is illustrated in Figure 2.

LANGUAGE COMPETENCE

Language competence can be classified into two types: organizational competence and pragmatic competence. Organizational competence comprises those abilities involved in controlling the formal organization of language for creating or recognizing grammatically correct sentences, comprehending their propositional content, and ordering them to form texts. These abilities are of two types: grammatical and textual. Grammatical competence includes rules of lexis, morphology, and syntax, which govern the choice of words to express specific significations, their forms, and their arrangement in sentences to express propositions. Textual competence includes the knowledge of the conventions of cohesion and rhetorical organization for joining utterances together to form a text.

Pragmatic competence includes those abilities which, *in addition to* organizational competence, are employed in the contextualized performance and interpretation of socially appropriate illocutionary acts in discourse. Pragmatic competence thus includes illocutionary competence, or the knowledge of how to perform illocutionary acts, or language functions, and sociolinguistic competence, or knowledge of the sociolinguistic conventions which govern appropriate language use in a particular culture and in varying situations in that culture.

PSYCHOPHYSIOLOGICAL SKILLS

Language competence may be realized in listening, speaking, reading and writing. These skills can be categorized in terms of mode (receptive, productive) and channel (auditory, visual), and distinguished by the psychophysiological skills that are involved in language use. Thus in the receptive mode (listening, reading) auditory and visual skills are employed, while in the productive mode (speaking, writing), neuromuscular skills (articulatory or digital) are employed.

STRATEGIC COMPETENCE

Communicative language use involves a dynamic interchange between the language user, the discourse, and the context of the situation in which the use occurs. The production of discourse thus requires the ability to

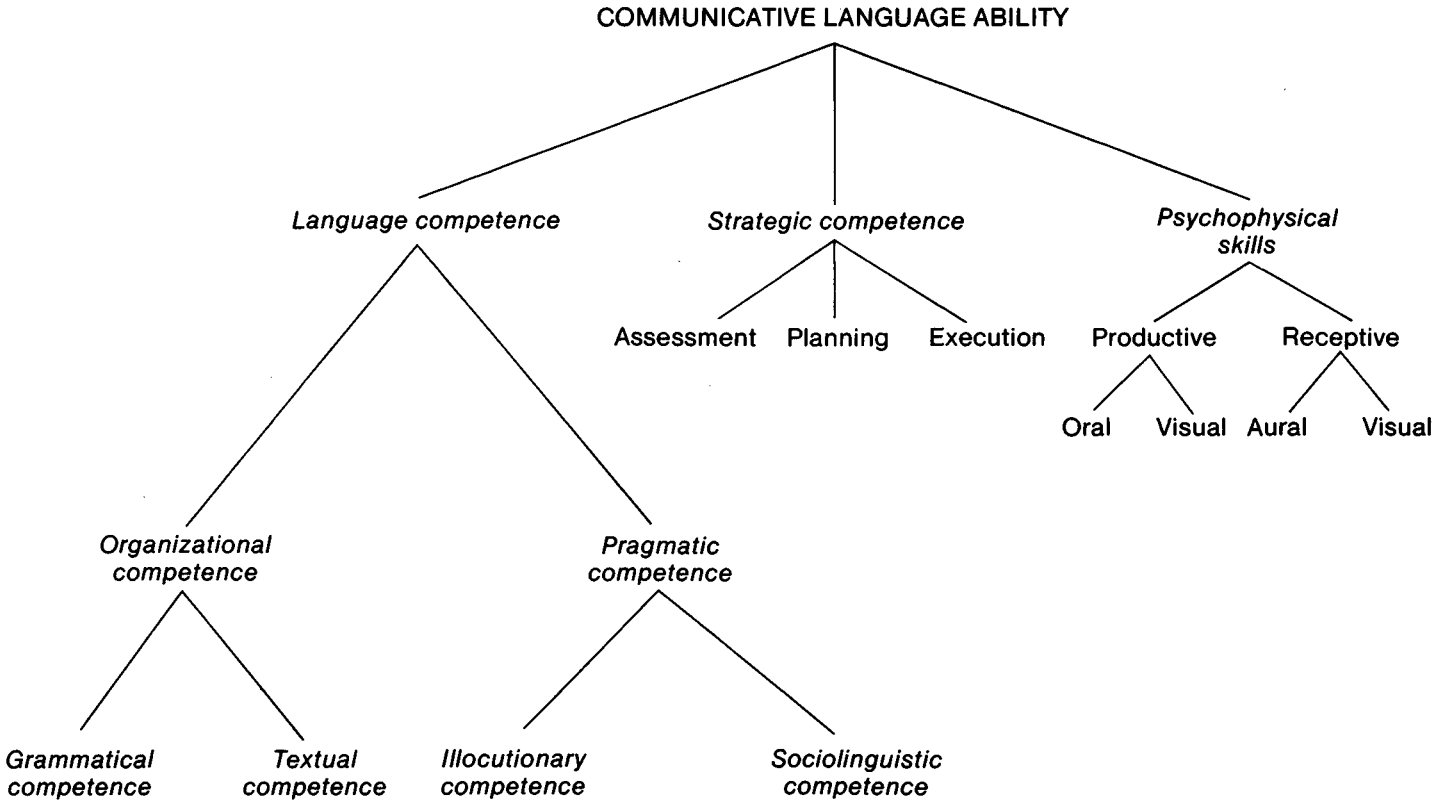

Figure 2 A framework for describing communicative language proficiency

assess the context for information relevant to the communicative goal and to then match information in the discourse to this information. The matching of new information to be encoded with relevant information that is available in the context (including presuppositional and real-world knowledge), and mapping this onto the maximally efficient use of existing language abilites is a function of strategic competence.

Faerch and Kasper (1983), in their discussion of communication strategies, present a model of speech production that includes a communicative goal, a planning process, a plan, and the execution of that plan through neurological and physiological processes. While Faerch and Kasper's model is intended to explain the use of communication strategies in interlanguage communication, I believe it provides a basis for a more general description of strategic competence in communicative language use. I would define strategic competence as the ability that enables us to i) formulate a plan for realizing a particular communicative goal, ii) execute that plan as an utterance, and iii) assess the extent to which the communicative goal has been achieved. These functions can be represented schematically as in Figure 3.

Actual performance versus abstract criteria

For criterion-referenced tests to satisfy the requirement of comparability across programs, they must share a common scale, one that is not defined with reference to the performance of different groups of test takers. Specifically, for test scores to be of use in program-free and comparative program evaluation, they must constitute an absolute scale of measurement, one that has true 'zero' and 'perfect' points. Achieving an absolute scale of foreign language proficiency with true 'zero' and 'perfect' levels is virtually impossible if one attempts to define this scale in terms of actual language use or actual language users. If we consider language proficiency to be similar to other cognitive abilities, such as intelligence, that may not have true zero points, as well as the likely existence of elements of the native language that are either universal to all languages or shared with the foreign language, then true 'zero' second language proficiency does not exist in actual individuals.

At the other end of the spectrum, the individual with absolutely complete language proficiency does not exist. Not only does language proficiency develop diachronically as a function of language change, it also develops in the way that all cognitive abilities constantly develop. And although the language use of native speakers is frequently advocated as a criterion for language proficiency, this is clearly inadequate for several reasons. First, native speakers show considerable variation in proficiency, particularly with regard to abilities such as cohesion, dis-

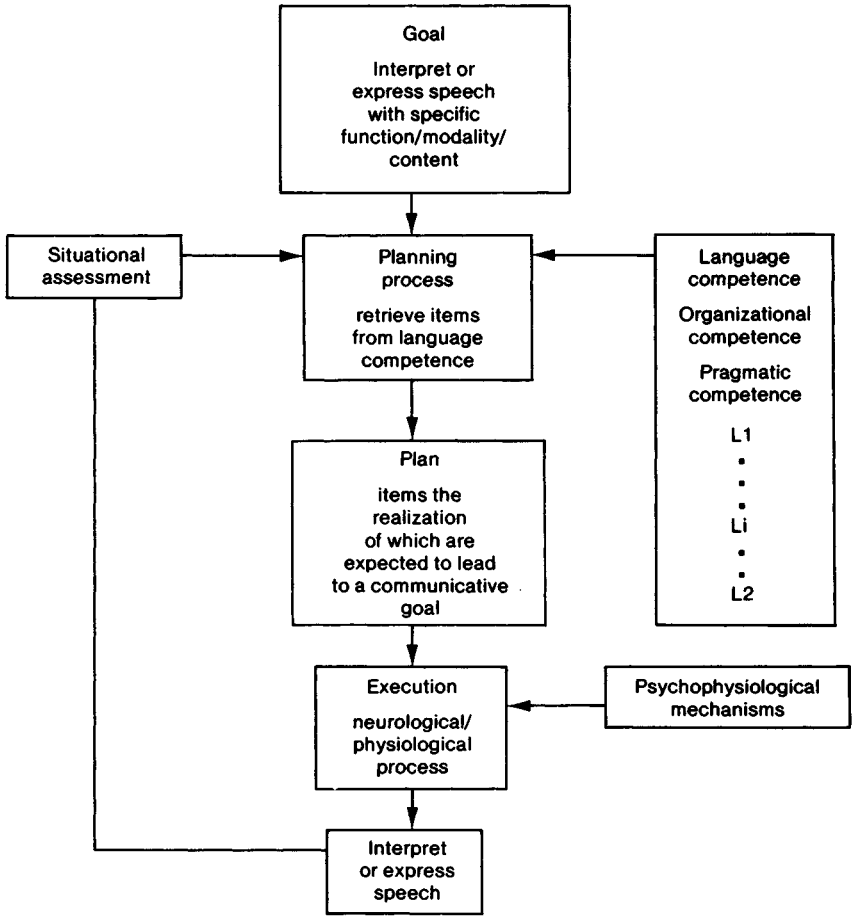

Figure 3 *A model of language use* (adapted from Faerch and Kasper, 1983)

course organization and sociolinguistic appropriateness. Second, there is the problem of identifying which variety or dialect to adopt as the 'native speaker' criterion. This question, which is often political or social rather than linguistic, is further complicated by the fact that boundaries between language varieties are seldom clearcut (Kachru, 1985). We must also consider differences in usage even within varieties or dialects, as well as differences between 'prescriptive' norms and the norms of actual usage. Finally, the whole concept of 'native speaker' has come to be regarded as little more than an abstraction (for example, Coulmas, 1981; Paikeday, 1985).

Because of these problems, it is virtually impossible to define criterion levels of language proficiency in terms of actual individuals or actual

performance. Rather, such levels must be defined abstractly, in terms of the relative presence or absence of the abilities that constitute the domain. To illustrate how levels can be defined abstractly in the context of an oral interview, consider the following scale that Adrian Palmer and I developed a few years ago:

	Vocabulary	*Cohesion*
0	*Extremely limited vocabulary* (A few words and formulaic phrases. Not possible to discuss any topic, due to limited vocabulary.)	*No cohesion* (Utterances completely disjointed, or discourse too short to judge.)
1	*Small vocabulary* (Difficulty in talking with examinee because of vocabulary limitations.)	*Very little cohesion* (Relationships between utterances not adeqately marked; frequent confusing relationship among ideas.)
2	*Vocabulary of moderate size* (Frequently misses or searches for words.)	*Moderate cohesion* (Relationships between utterances generally marked; sometimes confusing relationships among ideas.)
3	*Large vocabulary* (Seldom misses or searches for words.)	*Good cohesion* (Relationships between utterances well-marked.)
4	*Extensive vocabulary* (Rarely, if ever, misses or searches for words. Almost always uses appropriate word.)	*Excellent cohesion* (Uses a variety of appropriate devices; hardly ever confusing relationships among ideas.)

Figure 4 *Scales of ability in vocabulary and cohesion* (Bachman and Palmer, 1983)

While these scale definitions were designed for measuring fairly broad categories of proficiency, and would need to be specified more precisely to meet the needs of any given program evaluation, they do illustrate the

principle of defining scales abstractly, rather than in terms of actual performance or actual speakers.

In summary, I believe that in order to develop tests that are adequate for the uses of language program evaluation, that will yield scores that are comparable across differing sets of instructional objectives, we must begin by i) specifying a domain of communicative language ability that is consistent with current frameworks and ii) defining levels or scales of proficiency abstractly, in terms of relative degrees of ability, and independently of contextual features of language use.

A program of research and development

In a recent paper Bachman and Clark (1987) outline a program of research and development aimed at producing and validating criterion-referenced measures of communicative language ability. This program includes four components, as follows:

1 refining a theoretical model of communicative language ability, with particular attention to defining the specific ability domains in operational terms;
2 developing highly authentic measures of performance, based on the operational definitions of the individual domains of the model, for use as criteria in proficiency testing/development studies;
3 surveying currently available language testing instruments with respect to their degree of congruence with the requirements of a more fully elaborated model, selecting the most promising instruments and validating them against the criterion measures at issue in 2; and
4 developing and validating batteries of new instruments of optimum reliability, validity, and practicality for use in a variety of real-world testing contexts.

Conclusion

In this paper I have discussed the needs and problems of measuring learner outcomes in the evaluation of language programs. I have argued that standardized norm-referenced tests are generally inadequate for these needs, and proposed criterion-referenced tests of language proficiency as a solution. The development of such tests requires a well-specified domain of language proficiency and abstract definitions of proficiency levels. Finally, I have advocated a program of research for developing and validating criterion-referenced tests of language proficiency.

While the solution I have proposed is simple in its conception, its

implementation will require a major commitment on the part of language testers, program evaluators, applied linguists, and language teachers. Part of this commitment, I believe, is the realization that the time for armchair model-building is past, and that in order to move forward we must begin the empirical investigation of current models. This commitment will also involve both the recognition that our current models are probably inadequate, and the patience to continue the cyclical process of hypothesis testing and theory revision that will inform our research and development activities. In this regard, the framework of communicative language ability I have described here is not presented as a complete theory, but is intended as a starting point for research and development. And while I believe this framework is specific enough to guide test development and to generate hypotheses for empirical studies, this framework itself is subject to empirical validation, and will in all likelihood change to reflect our growing knowledge.

Much of the discussion in this paper has been at a theoretical level. However, the concerns to which it is addressed are practical. With the current proliferation of techniques and materials for language teaching, the need for program evaluation has never been greater. In the face of claims and counterclaims, choosing the 'best' program has become increasingly problematic, and I believe program evaluation provides the most effective means for examining the validity of these claims. In addition, there is an increasing need for measures of language proficiency for use in research aimed at better understanding the nature of language acquisition and language attrition. As this research has come to examine these processes in the context of the language classroom, the concerns of the language acquisition researcher and the program evaluator are beginning to merge. It is in the arena of classroom-centered research and program evaluation, therefore, that I believe the most pressing issues of measurement are to be found. It is also in this arena that the researcher and the practitioner are most likely to come together to work toward their solution.